Variable smoothing for convex optimization problems using stochastic gradients

Radu Ioan Boţ[*] Axel Böhm[†]

October 2, 2020

We aim to solve a structured convex optimization problem, where a nonsmooth function is composed with a linear operator. When opting for full splitting schemes, usually, primal-dual type methods are employed as they are effective and also well studied. However, under the additional assumption of Lipschitz continuity of the nonsmooth function which is composed with the linear operator we can derive novel algorithms through regularization via the Moreau envelope. Furthermore, we tackle large scale problems by means of stochastic oracle calls, very similar to stochastic gradient techniques. Applications to total variational denoising and deblurring, and matrix factorization are provided.

Keywords. structured convex optimization problem, variable smoothing algorithm, convergence rate, stochastic gradients
AMS Subject Classification. 90C25, 90C15, 65Y20

1 Introduction

The problem at hand is the following structured convex optimization problem

$$\min_{x\in\mathcal{H}} f(x) + g(Kx), \tag{1}$$

for real Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$, $f : \mathcal{H} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ a proper, convex and lower semicontinuous function, $g : \mathcal{G} \to \mathbb{R}$ a, possibly nonsmooth, convex and Lipschitz continuous function, and $K : \mathcal{H} \to \mathcal{G}$ a linear continuous operator.

[*] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: radu.bot@univie.ac.at. Research partially supported by FWF (Austrian Science Fund), project I 2419-N32.

[†] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: axel.boehm@univie.ac.at. Research supported by the doctoral programme *Vienna Graduate School on Computational Optimization (VGSCO)*, FWF (Austrian Science Fund), project W 1260.

Our aim will be to devise an algorithm for solving (1) following the *full splitting* paradigm (see [5, 6, 8, 9, 15, 17, 29]). In other words, we allow only proximal evaluations for simple nonsmooth functions, but no proximal evaluations for compositions with linear continuous operators, like, for instance, for $g \circ K$.

We will accomplish this feat by the means of a *smoothing strategy*, which, for the purpose of this paper, means, making use of the Moreau-Yosida approximation. The approach can be described as follows: we "smooth" g, i.e. we replace it by its Moreau envelope, and solve the resulting optimization problem by an *accelerated proximal-gradient algorithm* (see [3, 13, 21]). This approach is similar to the those in [7, 10, 11, 20, 22], where a convergence rate of $\mathcal{O}\left(\frac{\log(k)}{k}\right)$ is proved. However, our techniques (for the deterministic case) resemble more the ones in [28], where an improved rate of $\mathcal{O}(\frac{1}{k})$ is shown, with the most notable difference to our work is that we use a simpler stepsize and treat the stochastic case.

The only other family of methods able to solve problems of type (1) are the so called primal-dual algorithms, first and foremost the *primal-dual hybrid gradient (PDHG)* introduced in [15]. In comparison, this method does not need the Lipschitz continuity of g in order to proof convergence. However, in this very general case, convergence rates can only be shown for the so-called *restricted primal-dual gap* function. In order to derive from here convergence rates for the primal objective function, either Lipschitz continuity of g or finite dimensionality of the problem plus the condition that g must have full domain are necessary (see, for instance, [5, Theorem 9]). This means, that for infinite dimensional problems the assumptions required by both, PDHG and our method, for deriving convergence rates for the primal objective function are in fact equal, but for finite dimensional problems the assumption of PDHG are weaker. In either case, however, we are able to proof these rates for the sequence of iterates $(x_k)_{k\geq 1}$ itself whereas PDHG only has them for the sequence of so-called *ergodic iterates*, i.e. $(\frac{1}{k}\sum_{i=1}^k x_i)_{k\geq 1}$, which is naturally undesirable as the averaging slows the convergence down. Furthermore, we do not show any convergence for the iterates as these are notoriously hard to obtain for accelerated method whereas PDHG gets these in the strongly convex setting via standard fixed point arguments (see e.g. [29]).

Furthermore, we will also consider the case where only a stochastic oracle of the proximal operator of g is available to us. This setup corresponds e.g. to the case where the objective function is given as

$$\min_{x\in\mathcal{H}} f(x) + \sum_{i=1}^m g_i(K_i x), \tag{2}$$

where, for $i = 1, \dots, m$, $\mathcal{G}_i$ are real Hilbert spaces, $g_i : \mathcal{G}_i \to \mathbb{R}$ are convex and Lipschitz continuous functions and $K_i : \mathcal{H} \to \mathcal{G}_i$ are linear continuous operators, but the number of summands being large we wish to not compute all proximal operators of all $g_i, i = 1, \dots, m$, for purpose of making iterations cheaper to compute.

For the finite sum case (2), there exist algorithms of similar spirit such as those in [14, 24]. Some algorithms do in fact deal with a similar setup of stochastic gradient like evaluations, see [26], but only for smooth terms in the objective function.

In Section 2 we will cover the preliminaries about the Moreau-Yosida envelope as well as useful identities and estimates connected to it. In Section 3 we will deal with the deterministic case and prove a convergence rate of $\mathcal{O}(\frac{1}{k})$ for the function values at the iterates. Next up, in Section 4, we will consider the stochastic case as described above and prove a convergence rate of $\mathcal{O}\left(\frac{\log(k)}{\sqrt{k}}\right)$. Last but not least, we will look at some numerical examples in image processing in Section 5.

It is important to note that the proof for the deterministic setting differs surprisingly from the one for the stochastic setting. The technique for the stochastic setting is less refined in the sense that there is no coupling between the smoothing parameter and the extrapolation parameter. Where as this technique works also works for the deterministic setting it gives a worse convergence rate of $\mathcal{O}\left(\frac{\log k}{k}\right)$. The tight coupling of the two sequences of parameters, however does not work in the proof of the stochastic algorithm as it does not allow for the particular choice of the smoothing parameters needed there.

2 Preliminaries

In the main problem (1), the nonsmooth function regularizer g is supposed to be Lipschitz continuous. This assumption is necessary to ensure our main convergence results, however, many of the preliminary lemmata of this section hold true similarly if the function is only assumed to be lower semicontinuous. We will point this out in every statement of this section individually.

Definition 2.1. *For a proper, convex and lower semicontinuous function $g : \mathcal{H} \to \overline{\mathbb{R}}$, its convex conjugate is denoted by g^* defined as a function from $\mathcal{H}$ to $\overline{\mathbb{R}}$, given by*

$$g^*(x) := \sup_{p\in\mathcal{H}} \{\langle x, p\rangle - g(p)\} \quad \forall x \in \mathcal{H}.$$

As mentioned in the introduction, we want to *smooth* a nonsmooth function by considering its Moreau envelope. The next definition will clarify exactly what object we are talking about.

Definition 2.2. *For a proper, convex and lower semicontinuous function $g : \mathcal{H} \to \overline{\mathbb{R}}$, its Moreau envelope with the parameter $\mu \geq 0$ is defined as a function from $\mathcal{H}$ to $\mathbb{R}$, given by*

$${}^{\mu}g(\cdot) := \left(g^* + \frac{\mu}{2}\|\cdot\|^2\right)^*(\cdot) = \sup_{p\in\mathcal{H}} \left\{\langle \cdot, p\rangle - g^*(p) - \frac{\mu}{2}\|p\|^2\right\}.$$

From this definition, however, it is not completely evident that the Moreau envelope indeed fulfills its purpose in being a smooth representation of the original function. The next lemma will remedy this fact.

Lemma 2.1 (see [2, Proposition 12.29])**.** *Let $g : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function and $\mu > 0$. Then its Moreau envelope is Fréchet differentiable on $\mathcal{H}$. In particular, the gradient itself is given by*

$$\nabla({}^{\mu}g)(x) = \frac{1}{\mu}\left(x - \operatorname{prox}_{\mu g}(x)\right) = \operatorname{prox}_{\frac{1}{\mu}g^*}\left(\frac{x}{\mu}\right) \quad \forall x \in \mathcal{H}$$

and is μ^{-1}-Lipschitz continuous.

In particular, for all $\mu > 0$, a gradient step with respect to the Moreau envelope corresponds to a proximal step

$$x - \mu\nabla({}^{\mu}g)(x) = \operatorname{prox}_{\mu g}(x) \quad \forall x \in \mathcal{H}.$$

The previous lemma establishes two things. Not only does it clarify the smoothness of the Moreau envelope, but it also gives a way of computing its gradient. Obviously, a smooth representation whose gradient we would not be able to compute would not be any good.

As mentioned in the introduction, we want to smooth the nonsmooth summand of the objective function which is composed with the linear operator as this can be considered the crux of problem (1). The function $g \circ K$ will be *smoothed* via considering instead ${}^{\mu}g \circ K : \mathcal{H} \to \mathbb{R}$. Clearly, by the chain rule, this function is continuously differentiable with gradient given for every $x \in \mathcal{H}$ by

$$\nabla\left({}^{\mu}g \circ K\right)(x) = K^{*}\nabla\left({}^{\mu}g\right)(Kx) = \frac{1}{\mu}K^{*}\left(Kx - \operatorname{prox}_{\mu g}(Kx)\right) = K^{*}\operatorname{prox}_{\frac{1}{\mu}g^{*}}\left(\frac{Kx}{\mu}\right),$$

and is thus Lipschitz continuous with Lipschitz constant $\frac{\|K\|^2}{\mu}$, where $\|K\|$ denotes the operator norm of K.

Lipschitz continuity will play an integral role in our investigations, as can be seen by the following lemmas.

Lemma 2.2 (see [4, Proposition 4.4.6])**.** *Let $g : \mathcal{H} \to \mathbb{R}$ be a convex and L_g-Lipschitz continuous function. Then, the domain of its Fenchel conjugate is bounded, i.e.*

$$\operatorname{dom} g^{*} \subseteq B(0, L_g),$$

where $B(0, L_g)$ denotes the open ball with radius L_g around the origin.

The Moreau envelope even preserves the Lipschitzness of the original function.

Lemma 2.3 (see [18, Lemma 2.1])**.** *Let $g : \mathcal{H} \to \mathbb{R}$ be a convex and L_g-Lipschitz continuous function. Then its Moreau envelope ${}^{\mu}g$ is L_g-Lipschitz as well, i.e.*

$$|{}^{\mu}g(x) - {}^{\mu}g(y)| \leq L_g\|x - y\| \quad \forall x, y \in \mathcal{H}.$$

Proof. We observe that for all $x \in \mathcal{H}$

$$\nabla^{\mu}g(x) \in \partial g(\operatorname{prox}_{\mu g}(x)).$$

Therefore we can bound the gradient norm

$$\|\nabla^{\mu}g(x)\| \leq \sup\{\|v\| : y \in \mathcal{H}, v \in \partial g(y)\} \leq L_g \quad \forall x \in \mathcal{H}, \tag{3}$$

where we used in the last step that the Lipschitz continuity of g. The statement follows from the mean-value theorem. □

The following lemmata are not new, but we provide proofs anyways in order to remain self-contained and to shed insight on how to use the Moreau envelope for the interested reader.

Lemma 2.4 (see [28, Lemma 10 (a)])**.** *Let $g : \mathcal{H} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. The maximizing argument in the definition of the Moreau-Yosida envelope is given by its gradient, i.e. for $\mu > 0$ it holds that*

$$\operatorname*{arg\,max}_{p\in\mathcal{H}} \left\{ \langle \cdot, p\rangle - g^*(p) - \frac{\mu}{2}\|p\|^2 \right\} = \nabla{}^{\mu}g(\cdot).$$

Proof. Let $x \in \mathcal{H}$ be fixed. It holds

$$\begin{aligned}\operatorname*{arg\,max}_{p\in\mathcal{H}} \left\{ \langle x, p\rangle - g^*(p) - \frac{\mu}{2}\|p\|^2 \right\} =& \operatorname*{arg\,max}_{p\in\mathcal{H}} \left\{ -\frac{1}{2\mu}\|x\|^2 + \langle x, p\rangle - \frac{\mu}{2}\|p\|^2 - g^*(p) \right\} \\ =& \operatorname*{arg\,max}_{p\in\mathcal{H}} \left\{ -\frac{\mu}{2}\left\|\frac{x}{\mu} - p\right\|^2 - g^*(p) \right\} \\ =& \operatorname*{arg\,min}_{p\in\mathcal{H}} \left\{ g^*(p) + \frac{\mu}{2}\left\|\frac{x}{\mu} - p\right\|^2 \right\} \\ =& \operatorname{prox}_{\frac{1}{\mu}g^*}\left(\frac{x}{\mu}\right)\end{aligned}$$

and the conclusion follows by using Lemma 2.1. □

Lemma 2.5 (see [28, Lemma 10 (a)])**.** *For a proper, convex and lower semicontinuous function $g : \mathcal{H} \to \overline{\mathbb{R}}$ and every $x \in \mathcal{H}$ we can consider the mapping from $(0, +\infty)$ to $\mathbb{R}$ given by*

$$\mu \mapsto {}^{\mu}g(x). \tag{4}$$

This mapping is convex and differentiable and its derivative is given by

$$\frac{\partial}{\partial\mu}{}^{\mu}g(x) = -\frac{1}{2}\|\nabla{}^{\mu}g(x)\|^2 \qquad \forall x \in \mathcal{H}\ \forall\mu \in (0, +\infty).$$

Proof. Let $x \in \mathcal{H}$ be fixed. From the definition of the Moreau-Yosida envelope we can see that the mapping given in (4) is a pointwise supremum of functions which are linear in μ. It is therefore convex. Furthermore, since the objective function is strongly concave, this supremum is uniquely attained at $\nabla{}^{\mu}g(x) = \operatorname{arg\,max}_{p\in\mathcal{H}} \left\{ \langle x, p\rangle - g^*(p) - \frac{\mu}{2}\|p\|^2 \right\}$. According to the Danskin Theorem, the function $\mu \mapsto {}^{\mu}g(x)$ is differentiable and its gradient is given by

$$\begin{aligned}\frac{\partial}{\partial\mu}{}^{\mu}g(x) =& \frac{\partial}{\partial\mu}\sup_{p\in\mathcal{H}} \left\{ \langle x, p\rangle - g^*(p) - \frac{\mu}{2}\|p\|^2 \right\} \\ =& -\frac{1}{2}\|\nabla{}^{\mu}g(x)\|^2 \quad \forall\mu \in (0, +\infty).\end{aligned}$$

□

Lemma 2.6 ([28, Lemma 10 (b)]). *Let $g : \mathcal{H} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. For $\mu_1, \mu_2 > 0$ and every $x \in \mathcal{H}$ it holds*

$$ {}^{\mu_1}g(x) \leq {}^{\mu_2}g(x) + (\mu_2 - \mu_1)\frac{1}{2}\|\nabla^{\mu_1}g(x)\|^2. \tag{5} $$

If g is additionally L_g-Lipschitz and if $\mu_2 \geq \mu_1 > 0$, then

$$ {}^{\mu_2}g(x) \leq {}^{\mu_1}g(x) \leq {}^{\mu_2}g(x) + (\mu_2 - \mu_1)\frac{L_g^2}{2}. \tag{6} $$

Proof. Let $x \in \mathcal{H}$ be fixed. Via Lemma 2.5 we know that the map $\mu \mapsto {}^{\mu}g(x)$ is convex and differentiable. We can therefore use the gradient inequality to deduce that

$$ \begin{aligned} {}^{\mu_2}g(x) &\geq {}^{\mu_1}g(x) + (\mu_2 - \mu_1)\left(\frac{\partial}{\partial \mu}{}^{\mu}g(x)\Big|_{\mu=\mu_1}\right) \\ &= {}^{\mu_1}g(x) - (\mu_2 - \mu_1)\frac{1}{2}\|\nabla^{\mu_1}g(x)\|^2, \end{aligned} $$

which is exactly the first statement of the lemma. The first inequality of (6) follows directly from the definition of the Moreau envelope and the second one from (5) and (3). □

By applying a limiting argument it is easy to see that (6) implies that for any $\mu > 0$

$$ {}^{\mu}g(x) \leq g(x) \leq {}^{\mu}g(x) + \mu\frac{L_g^2}{2}, \tag{7} $$

which shows that the Moreau envelope is always a lower approximation the original function.

Lemma 2.7 (see [28, Lemma 10 (c)]). *Let $g : \mathcal{H} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. Then, for $\mu > 0$ and every $x, y \in \mathcal{H}$ we have that*

$$ {}^{\mu}g(x) + \langle \nabla^{\mu}g(x), y - x\rangle \leq g(y) - \frac{\mu}{2}\|\nabla^{\mu}g(x)\|^2. $$

Proof. Using Lemma 2.4 and the definition of the Moreau-Yosida envelope we get that

$$ \begin{aligned} {}^{\mu}g(x) + \langle \nabla^{\mu}g(x), y - x\rangle &= \langle x, \nabla^{\mu}g(x)\rangle - g^*(\nabla^{\mu}g(x)) - \frac{\mu}{2}\|\nabla^{\mu}g(x)\|^2 + \langle \nabla^{\mu}g(x), y - x\rangle \\ &= \langle \nabla^{\mu}g(x), y\rangle - g^*(\nabla^{\mu}g(x)) - \frac{\mu}{2}\|\nabla^{\mu}g(x)\|^2 \\ &\leq \sup_{p\in\mathcal{H}}\{\langle p, y\rangle - g^*(p)\} - \frac{\mu}{2}\|\nabla^{\mu}g(x)\|^2 \\ &= g(y) - \frac{\mu}{2}\|\nabla^{\mu}g(x)\|^2. \end{aligned} $$

□

In the convergence proof of Lemma 3.3 we will need the inequality in the above lemma at the points Kx and Ky, namely

$$\begin{aligned} g(Ky) - \frac{\mu}{2}\|\nabla^{\mu} g(Kx)\|^2 &\geq {}^{\mu}g(Kx) + \langle \nabla^{\mu} g(Kx), Ky - Kx\rangle \\ &= {}^{\mu}g(Kx) + \langle K^* \nabla^{\mu} g(Kx), y - x\rangle \\ &= {}^{\mu}g(Kx) + \langle \nabla({}^{\mu}g \circ K)(x), y - x\rangle \quad \forall x, y \in \mathcal{H}. \end{aligned} \tag{8}$$

The following lemma is a standard result for convex and Fréchet differentiable functions.

Lemma 2.8 (see [23]). *For a convex and Fréchet differentiable function $h : \mathcal{H} \to \mathbb{R}$ with L_h-Lipschitz continuous gradient we have that*

$$h(x) + \langle \nabla h(x), y - x\rangle \leq h(y) - \frac{1}{2L_h}\|\nabla h(x) - \nabla h(y)\|^2 \quad \forall x, y \in \mathcal{H}.$$

By applying Lemma 2.8 with ${}^{\mu}g$, Kx and Ky instead of h, x and y respectively, we obtain

$${}^{\mu}g(Kx) + \langle \nabla({}^{\mu}g \circ K)(x), y - x\rangle \leq {}^{\mu}g(Ky) - \frac{\mu}{2}\|\nabla^{\mu} g(Kx) - \nabla^{\mu} g(Ky)\|^2 \quad \forall x, y \in \mathcal{H}. \tag{9}$$

The following technical result will be used in the proof of the convergence statement.

Lemma 2.9. *For $\alpha \in (0, 1)$ and every $x, y \in \mathcal{H}$ we have that*

$$(1 - \alpha)\|x - y\|^2 + \alpha\|y\|^2 \geq \alpha(1 - \alpha)\|x\|^2.$$

3 Deterministic Method

Problem 3.1. *The problem at hand reads*

$$\min_{x \in \mathcal{H}} F(x) := f(x) + g(Kx),$$

for a proper, convex and lower semicontinuous function $f : \mathcal{H} \to \overline{\mathbb{R}}$, a convex and L_g-Lipschitz continuous ($L_g > 0$) function $g : \mathcal{G} \to \mathbb{R}$, and a nonzero linear continuous operator $K : \mathcal{H} \to \mathcal{G}$.

The idea of the algorithm which we propose to solve (1) is to smooth g and then to solve the resulting problem by means of an accelerated proximal-gradient.

Algorithm 3.1 (Variable Accelerated SmooThing (VAST)). *Let $y_0 = x_0 \in \mathcal{H}$, $(\mu_k)_{k\geq 0} \subseteq (0, +\infty)$, and $(t_k)_{k\geq 1}$ a sequence of real numbers with $t_1 = 1$ and $t_k \geq 1$ for every $k \geq 2$. Consider the following iterative scheme*

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} L_k = \frac{\|K\|^2}{\mu_k} \\ \gamma_k = \frac{1}{L_k} \\ x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k K^* \operatorname{prox}_{\frac{1}{\mu_k} g^*}\left(\frac{K y_{k-1}}{\mu_k}\right)\right) \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}). \end{array} \right.$$

Remark 3.1. The assumption $t_1 = 1$ can be removed but guarantees easier computation and is also in line with classical choices of $(t_k)_{k\geq 1}$ in [13, 21].

Remark 3.2. The sequence $(u_k)_{k\geq 1}$ given by

$$u_k := x_{k-1} + t_k(x_k - x_{k-1}) \quad \forall k \geq 1,$$

despite not appearing in the algorithm, will feature a prominent role in the convergence proof. Due to the convention $t_1 = 1$ we have that

$$u_1 := x_0 + t_1(x_1 - x_0) = x_1.$$

We also denote

$$F^k = f + {}^{\mu_k}g \circ K \quad \forall k \geq 0.$$

The next theorem is the main result of this section and it will play a fundamental role when proving a convergence rate of $\mathcal{O}(\frac{1}{k})$ for the sequence $(F(x_k))_{k\geq 0}$.

Theorem 3.1. *Consider the setup of Problem 3.1 and let $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ be the sequences generated by Algorithm 3.1. Assume that for every $k \geq 1$*

$$\mu_k - \mu_{k+1} - \frac{\mu_{k+1}}{t_{k+1}} \leq 0$$

and

$$\left(1 - \frac{1}{t_{k+1}}\right)\gamma_{k+1}t_{k+1}^2 = \gamma_k t_k^2.$$

Then, for every optimal solution x^* of Problem 3.1, it holds

$$F(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \mu_N \frac{L_g^2}{2} \quad \forall N \geq 1.$$

The proof of this result relies on several partial results which we will prove as follows.

Lemma 3.1. *The following statement holds for every $z \in \mathcal{H}$ and every $k \geq 0$*

$$F^{k+1}(x_{k+1}) + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq$$
$$f(z) + {}^{\mu_{k+1}}g(Ky_k) + \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), z - y_k\rangle + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

Proof. Let $k \geq 0$ be fixed. Since, by the definition of the proximal map, x_{k+1} is the minimizer of a $\frac{1}{\gamma_{k+1}}$-strongly convex function we know that for every $z \in \mathcal{H}$

$$f(x_{k+1}) + {}^{\mu_{k+1}}g(Ky_k) + \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x_{k+1} - y_k\rangle + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - y_k\|^2 +$$
$$\frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq$$
$$f(z) + {}^{\mu_{k+1}}g(Ky_k) + \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), z - y_k\rangle + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

Next we use the L_{k+1}-smoothness of ${}^{\mu_{k+1}}g \circ K$ and the fact that $\frac{1}{\gamma_{k+1}} = L_{k+1}$ to deduce

$$f(x_{k+1}) + {}^{\mu_{k+1}}g(Kx_{k+1}) + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq \\ f(z) + {}^{\mu_{k+1}}g(Ky_k) + \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), z - y_k\rangle + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

□

Lemma 3.2. *Let x^* be an optimal solution of Problem 3.1. Then it holds*

$$\gamma_1(F^1(x_1) - F(x^*)) + \frac{1}{2}\|u_1 - x^*\|^2 \leq \frac{1}{2}\|x^* - x_0\|^2.$$

Proof. We use the gradient inequality to deduce that for every $z \in \mathcal{H}$ and every $k \geq 0$

$${}^{\mu_{k+1}}g(Ky_k) + \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), z - y_k\rangle \leq {}^{\mu_{k+1}}g(Kz) \leq g(Kz)$$

and plug this into the statement of Lemma 3.1 to conclude that

$$F^{k+1}(x_{k+1}) + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq F(z) + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

For $k = 0$ we get that

$$F^1(x_1) + \frac{1}{2\gamma_1}\|x_1 - x^*\|^2 \leq F(x^*) + \frac{1}{2\gamma_1}\|x^* - y_0\|^2.$$

Now we us the fact that $u_1 = x_1$ and $y_0 = x_0$ to obtain the conclusion. □

Lemma 3.3. *Let x^* be an optimal solution of Problem 3.1. The following descent-type inequality holds for every $k \geq 0$*

$$F^{k+1}(x_{k+1}) - F(x^*) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \leq \left(1 - \frac{1}{t_{k+1}}\right)\left(F^k(x_k) - F(x^*)\right) + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \\ + \left(1 - \frac{1}{t_{k+1}}\right)\left(\mu_k - \mu_{k+1} - \frac{\mu_{k+1}}{t_{k+1}}\right)\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2.$$

Proof. Let $k \geq 0$ be fixed. We apply Lemma 3.1 with $z := \left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*$ to deduce that

$$F^{k+1}(x_{k+1}) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \leq f\left(\left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*\right) + {}^{\mu_{k+1}}g(Ky_k) \\ + \left(1 - \frac{1}{t_{k+1}}\right)\langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x_k - y_k\rangle \\ + \frac{1}{t_{k+1}}\langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x^* - y_k\rangle + \frac{1}{2\gamma_{k+1}t_{k+1}^2}\|u_k - x^*\|^2.$$

Using the convexity of f gives

$$\begin{aligned}F^{k+1}(x_{k+1}) + \frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} &\leq \left(1-\frac{1}{t_{k+1}}\right) f(x_k) + \frac{1}{t_{k+1}} f(x^*) \\ &+ \left(1-\frac{1}{t_{k+1}}\right) {}^{\mu_{k+1}}g(Ky_k) + \left(1-\frac{1}{t_{k+1}}\right) \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x_k - y_k\rangle \\ &+ \frac{1}{t_{k+1}} {}^{\mu_{k+1}}g(Ky_k) + \frac{1}{t_{k+1}} \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x^* - y_k\rangle + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.\end{aligned} \tag{10}$$

Now, we use (8) to deduce that

$$\begin{aligned}\frac{1}{t_{k+1}} {}^{\mu_{k+1}}g(Ky_k) + \frac{1}{t_{k+1}} \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x^* - y_k\rangle \leq \\ \frac{1}{t_{k+1}} g(Kx^*) - \frac{1}{t_{k+1}} \frac{\mu_{k+1}}{2} \|\nabla^{\mu_{k+1}}g(Ky_k)\|^2\end{aligned} \tag{11}$$

and (9) to conclude that

$$\begin{aligned}\left(1-\frac{1}{t_{k+1}}\right) {}^{\mu_{k+1}}g(Ky_k) + \left(1-\frac{1}{t_{k+1}}\right) \langle \nabla({}^{\mu_{k+1}}g \circ K)(y_k), x_k - y_k\rangle \leq \\ \left(1-\frac{1}{t_{k+1}}\right) {}^{\mu_{k+1}}g(Kx_k) - \left(1-\frac{1}{t_{k+1}}\right) \frac{\mu_{k+1}}{2} \|\nabla^{\mu_{k+1}}g(Ky_k) - \nabla^{\mu_{k+1}}g(Kx_k)\|^2.\end{aligned} \tag{12}$$

Combining (10), (11) and (12) gives

$$\begin{aligned}F^{k+1}(x_{k+1}) + \frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} &\leq \left(1-\frac{1}{t_{k+1}}\right) {}^{\mu_{k+1}}g(Kx_k) + \left(1-\frac{1}{t_{k+1}}\right) f(x_k) \\ &+ \frac{1}{t_{k+1}} g(Kx^*) + \frac{1}{t_{k+1}} f(x^*) \\ &- \left(1-\frac{1}{t_{k+1}}\right) \frac{\mu_{k+1}}{2} \|\nabla^{\mu_{k+1}}g(Ky_k) - \nabla^{\mu_{k+1}}g(Kx_k)\|^2 \\ &- \frac{1}{t_{k+1}} \frac{\mu_{k+1}}{2} \|\nabla^{\mu_{k+1}}g(Ky_k)\|^2 + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.\end{aligned}$$

The first term on the right hand side is ${}^{\mu_{k+1}}g(Kx_k)$ but we would like it to be ${}^{\mu_k}g(Kx_k)$. Therefore we use Lemma 2.6 to deduce that

$$\begin{aligned}F^{k+1}(x_{k+1}) + \frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} &\leq \left(1-\frac{1}{t_{k+1}}\right) {}^{\mu_k}g(Kx_k) + \left(1-\frac{1}{t_{k+1}}\right) f(x_k) \\ &+ \frac{1}{t_{k+1}} g(Kx^*) + \frac{1}{t_{k+1}} f(x^*) + \left(1-\frac{1}{t_{k+1}}\right) (\mu_k - \mu_{k+1}) \frac{1}{2} \|\nabla^{\mu_{k+1}}g(Kx_k)\|^2 \\ &- \left(1-\frac{1}{t_{k+1}}\right) \frac{\mu_{k+1}}{2} \|\nabla^{\mu_{k+1}}g(Ky_k) - \nabla^{\mu_{k+1}}g(Kx_k)\|^2 \\ &- \frac{1}{t_{k+1}} \frac{\mu_{k+1}}{2} \|\nabla^{\mu_{k+1}}g(Ky_k)\|^2 + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.\end{aligned} \tag{13}$$

Next up we want to estimate all the norms of gradients by using Lemma 2.9 which says that

$$\begin{aligned}\left(1-\frac{1}{t_{k+1}}\right)\|\nabla^{\mu_{k+1}}g(Ky_k)-\nabla^{\mu_{k+1}}g(Kx_k)\|^2+\frac{1}{t_{k+1}}\|\nabla^{\mu_{k+1}}g(Ky_k)\|^2 \geq \\ \left(1-\frac{1}{t_{k+1}}\right)\frac{1}{t_{k+1}}\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2.\end{aligned} \tag{14}$$

Combining (13) and (14) gives

$$\begin{aligned}F^{k+1}(x_{k+1})+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} &\leq \left(1-\frac{1}{t_{k+1}}\right){}^{\mu_k}g(Kx_k)+\left(1-\frac{1}{t_{k+1}}\right)f(x_k) \\ &+\frac{1}{t_{k+1}}g(Kx^*)+\frac{1}{t_{k+1}}f(x^*)+\left(1-\frac{1}{t_{k+1}}\right)(\mu_k-\mu_{k+1})\frac{1}{2}\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2 \\ &-\frac{\mu_{k+1}}{2}\left(1-\frac{1}{t_{k+1}}\right)\frac{1}{t_{k+1}}\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.\end{aligned}$$

Now we combine the two terms containing $\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2$ and get that

$$\begin{aligned}F^{k+1}(x_{k+1})+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} &\leq \left(1-\frac{1}{t_{k+1}}\right){}^{\mu_k}g(Kx_k)+\left(1-\frac{1}{t_{k+1}}\right)f(x_k) \\ &+\frac{1}{t_{k+1}}g(Kx^*)+\frac{1}{t_{k+1}}f(x^*)+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \\ &+\left(1-\frac{1}{t_{k+1}}\right)\left(\mu_k-\mu_{k+1}-\frac{\mu_{k+1}}{t_{k+1}}\right)\frac{1}{2}\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2.\end{aligned}$$

By subtracting $F(x^*)=f(x^*)+g(Kx^*)$ on both sides we finally obtain

$$\begin{aligned}F^{k+1}(x_{k+1})-F(x^*)+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} &\leq \left(1-\frac{1}{t_{k+1}}\right)\Big(F^k(x_k)-F(x^*)\Big)+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \\ &+\left(1-\frac{1}{t_{k+1}}\right)\left(\mu_k-\mu_{k+1}-\frac{\mu_{k+1}}{t_{k+1}}\right)\frac{1}{2}\|\nabla^{\mu_{k+1}}g(Kx_k)\|^2.\end{aligned}$$

□

Now we are in the position to prove Theorem 3.1.

Proof of Theorem 3.1. We start with the statement of Lemma 3.3 and use the assumption that

$$\mu_k-\mu_{k+1}-\frac{\mu_{k+1}}{t_{k+1}}\leq 0$$

to make the last term in the inequality disappear for every $k\geq 0$

$$F^{k+1}(x_{k+1})-F(x^*)+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\leq\left(1-\frac{1}{t_{k+1}}\right)\Big(F^k(x_k)-F(x^*)\Big)+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.$$

Now we use the assumption that

$$\left(1 - \frac{1}{t_{k+1}}\right) \gamma_{k+1} t_{k+1}^2 = \gamma_k t_k^2$$

to get that for every $k \geq 0$

$$\gamma_{k+1} t_{k+1}^2 (F^{k+1}(x_{k+1}) - F(x^*)) + \frac{\|u_{k+1} - x^*\|^2}{2} \leq \gamma_k t_k^2 (F^k(x_k) - F(x^*)) + \frac{\|u_k - x^*\|^2}{2}. \tag{15}$$

Let $N \geq 2$. Summing (15) from $k = 1$ to $N - 1$ and getting rid of the nonnegative term $\|u_N - x^*\|^2$ gives

$$\gamma_N t_N^2 (F^N(x_N) - F(x^*)) \leq \gamma_1 (F^1(x_1) - F(x^*)) + \frac{\|u_1 - x^*\|^2}{2} \quad \forall N \geq 2.$$

Since $t_1 = 1$, the above inequality is fulfilled also for $N = 1$. Using Lemma 3.2 shows that

$$F^N(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{\gamma_N t_N^2} \quad \forall N \geq 1.$$

The above inequality, however, is still in terms of the smoothed objective function. In order to go to the actual objective function we apply (7) and deduce that

$$F(x_N) - F(x^*) \leq F^N(x_N) - F(x^*) + \mu_N \frac{L_g^2}{2} \leq \frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \mu_N \frac{L_g^2}{2} \quad \forall N \geq 1.$$

□

Corollary 3.1. *By choosing the parameters $(\mu_k)_{k \geq 1}, (t_k)_{k \geq 1}, (\gamma_k)_{k \geq 1}$ in the following way,*

$$t_1 = 1, \quad \mu_1 = b\|K\|^2, \text{ for } b > 0,$$

and for every $k \geq 1$

$$t_{k+1} := \sqrt{t_k^2 + 2t_k}, \quad \mu_{k+1} := \mu_k \frac{t_k^2}{t_{k+1}^2 - t_{k+1}}, \quad \gamma_k := \frac{\mu_k}{\|K\|^2}, \tag{16}$$

they fulfill

$$\mu_k - \mu_{k+1} - \frac{\mu_{k+1}}{t_{k+1}} \leq 0 \tag{17}$$

and

$$\left(1 - \frac{1}{t_{k+1}}\right) \gamma_{k+1} t_{k+1}^2 = \gamma_k t_k^2 \tag{18}$$

For this choice of the parameters we have that

$$F(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{b(N+1)} + \frac{bL_g^2 \|K\|^2}{(N+1)} \exp\left(\frac{4\pi^2}{6}\right) \quad \forall N \geq 1.$$

Proof. Since γ_k and μ_k are a scalar multiple of each other (18) is equivalent to

$$\left(1 - \frac{1}{t_{k+1}}\right) \mu_{k+1} t_{k+1}^2 = \mu_k t_k^2 \quad \forall k \geq 1$$

and further to (by taking into account that $t_{k+1} > 1$ for every $k \geq 1$)

$$\mu_{k+1} = \mu_k \frac{t_k^2}{t_{k+1}^2} \frac{t_{k+1}}{t_{k+1} - 1} = \mu_k \frac{t_k^2}{t_{k+1}^2 - t_{k+1}} \quad \forall k \geq 1. \tag{19}$$

Our update choice in (16) for the sequence $(\mu_k)_{k \geq 1}$ is exactly chosen in such a way that it satisfies this. Plugging (19) into (17) gives for every $k \geq 1$ the condition

$$1 \leq \left(1 + \frac{1}{t_{k+1}}\right) \frac{t_k^2}{t_{k+1}^2} \frac{t_{k+1}}{t_{k+1} - 1} = \frac{t_k^2}{t_{k+1}^2} \frac{t_{k+1} + 1}{t_{k+1} - 1},$$

which is equivalent to

$$0 \geq t_{k+1}^3 - t_{k+1}^2 - t_k^2 t_{k+1} - t_k^2$$

and further to

$$t_{k+1}^2 + t_k^2 \geq t_{k+1} \left(t_{k+1}^2 - t_k^2\right).$$

Plugging in $t_{k+1} = \sqrt{t_k^2 + 2t_k}$ we get that this equivalent to

$$t_{k+1}^2 + t_k^2 \geq t_{k+1} 2t_k \quad \forall k \geq 1,$$

which is evidently fulfilled. Thus, the choices in (16) are indeed feasible for our algorithm.

Now we want to prove the claimed convergence rates. Via induction we show that

$$\frac{k+1}{2} \leq t_k \leq k \quad \forall k \geq 1. \tag{20}$$

Evidently, this holds for $t_1 = 1$. Assuming that (20) holds for $k \geq 1$, we easily see that

$$t_{k+1} = \sqrt{t_k^2 + 2t_k} \leq \sqrt{k^2 + 2k} \leq \sqrt{k^2 + 2k + 1} = k + 1$$

and, on the other hand,

$$t_{k+1} = \sqrt{t_k^2 + 2t_k} \geq \sqrt{\frac{(k+1)^2}{4} + k + 1} = \frac{1}{2}\sqrt{k^2 + 6k + 5} \geq \frac{1}{2}\sqrt{k^2 + 4k + 4} = \frac{k+2}{2}.$$

In the following we prove a similar estimate for the sequence $(\mu_k)_{k \geq 1}$. To this end we show, again by induction, the following recursion for every $k \geq 2$

$$\mu_k = \mu_1 \frac{\prod_{j=1}^{k-1} t_j}{\prod_{j=2}^{k} (t_j - 1)} \frac{1}{t_k}. \tag{21}$$

For $k = 2$ this follows from the definition (19). Assume now that (21) holds for $k \geq 2$. From here we have that

$$\mu_{k+1} = \mu_k \frac{t_k^2}{t_{k+1}(t_{k+1}-1)} = \mu_1 \frac{\prod_{j=1}^{k-1} t_j}{\prod_{j=2}^{k}(t_j - 1)} \frac{1}{t_k} \frac{t_k^2}{t_{k+1}(t_{k+1}-1)} = \mu_1 \frac{\prod_{j=1}^{k} t_j}{\prod_{j=2}^{k+1}(t_j - 1)} \frac{1}{t_{k+1}}.$$

Using (21) together with (20) we can check that for every $k \geq 1$

$$\mu_{k+1} = \mu_1 \frac{\prod_{j=1}^{k} t_j}{\prod_{j=2}^{k+1}(t_j - 1)} \frac{1}{t_{k+1}} = \frac{\mu_1}{t_{k+1}} \prod_{j=1}^{k} \frac{t_j}{(t_{j+1}-1)} \geq \frac{\mu_1}{t_{k+1}} = b\|K\|^2 \frac{1}{t_{k+1}}, \tag{22}$$

where we used in the last step the fact that $t_{k+1} \leq t_k + 1$.

The last thing to check is the fact that μ_k goes to zero like $\frac{1}{k}$. First we check that for every $k \geq 1$

$$\frac{t_k}{t_{k+1}-1} \leq 1 + \frac{1}{t_{k+1}(t_{k+1}-1)}. \tag{23}$$

This can be seen via

$$(t_k + 1)t_{k+1} \leq (t_k + 1)^2 = t_{k+1}^2 + 1 \quad \forall k \geq 1.$$

By bringing t_{k+1} to the other side we get that

$$t_{k+1}t_k \leq t_{k+1}^2 - t_{k+1} + 1,$$

from which we can deduce (23) by dividing by $t_{k+1}^2 - t_{k+1}$.

We plug in the estimate (23) in (21) and get for every $k \geq 2$

$$\begin{aligned}
\mu_k &= \mu_1 \frac{\prod_{j=1}^{k-1} t_j}{\prod_{j=1}^{k-1}(t_{j+1}-1)} \frac{1}{t_k} \\
&\leq \mu_1 \prod_{j=1}^{k-1} \left(1 + \frac{1}{t_{j+1}(t_{j+1}-1)}\right) \frac{1}{t_k} \leq \mu_1 \prod_{j=1}^{k-1} \left(1 + \frac{4}{(j+2)j}\right) \frac{1}{t_k} \\
&\leq \mu_1 \prod_{j=1}^{k-1} \left(1 + \frac{4}{j^2}\right) \frac{1}{t_k} \leq \mu_1 \exp\left(\frac{\pi^2 4}{6}\right) \frac{1}{t_k} = b\|K\|^2 \exp\left(\frac{\pi^2 4}{6}\right) \frac{1}{t_k}.
\end{aligned}$$

With the above inequalities we can to deduce the claimed convergence rates. First note that from Theorem 3.1 we have

$$F(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \mu_N \frac{L_g^2}{2} \quad \forall N \geq 1.$$

Now, in order to obtain the desired conclusion, we used the above estimates and deduce for every $N \geq 1$

$$\begin{aligned}
\frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \mu_N \frac{L_g^2}{2} &\leq \frac{\|x_0 - x^*\|^2}{2bt_N} + \frac{bL_g^2\|K\|^2}{2t_N} \exp\left(\frac{4\pi^2}{6}\right) \\
&\leq \frac{\|x_0 - x^*\|^2}{b(N+1)} + \frac{bL_g^2\|K\|^2}{(N+1)} \exp\left(\frac{4\pi^2}{6}\right),
\end{aligned}$$

where we used that

$$\gamma_N t_N = \frac{\mu_N t_N}{\|K\|^2} \geq b,$$

as shown in (22). □

Remark 3.3. Consider the choice (see [21])

$$t_1 = 1, \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \quad \forall k \geq 1$$

and

$$\mu_1 = b\|K\|^2, \text{ for } b > 0.$$

Since

$$t_k^2 = t_{k+1}^2 - t_{k+1} \quad \forall k \geq 1,$$

we see that in this setting we have to choose

$$\mu_k = b\|K\|^2 \text{ and } \gamma_k = b \quad \forall k \geq 1.$$

Thus, the sequence of optimal function values $(F(x_N))_{N\geq 1}$ approaches a $b\|K\|^2 \frac{L_g}{2}$-approximation of the optimal objective value $F(x^*)$ with a convergence rate of $\mathcal{O}(\frac{1}{N^2})$, i.e.

$$F(x_N) - F(x^*) \leq 2\frac{\|x_0 - x^*\|^2}{b(N+1)^2} + b\frac{\|K\|^2 L_g^2}{2} \quad \forall N \geq 1.$$

4 Stochastic Method

Problem 4.1. *The problem is the same as in the deterministic case*

$$\min_{x \in \mathcal{H}} f(x) + g(Kx)$$

other than the fact that at each iteration we are only given a stochastic estimator of the quantity

$$\nabla({}^{\mu_k} g \circ K)(\cdot) = K^* \operatorname{prox}_{\frac{1}{\mu_k} g^*}\left(\frac{1}{\mu_k} K \cdot\right) \quad \forall k \geq 1.$$

Remark 4.1. See Algorithm 4.3 for a setting where such an estimator is easily computed.

For the stochastic quantities arising in this section we will use the following notation. For every $k \geq 0$, we denote by $\sigma(x_0, \dots, x_k)$ the smallest σ-algebra generated by the family of random variables $\{x_0, \dots, x_k\}$ and by $\mathbb{E}_k(\cdot) := \mathbb{E}(\cdot | \sigma(x_0, \dots, x_k))$ the conditional expectation with respect to this σ-algebra.

Algorithm 4.1 (stochastic Variable Accelerated SmooThing (sVAST))**.** *Let* $y_0 = x_0 \in \mathcal{H}$, $(\mu_k)_{k\geq 1}$ *a sequence of positive and nonincreasing real numbers, and* $(t_k)_{k\geq 1}$ *a sequence* *of real numbers with $t_1 = 1$ and $t_k \geq 1$ for every $k \geq 2$. Consider the following iterative scheme*

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} L_k = \frac{\|K\|^2}{\mu_k} \\ \gamma_k = \frac{1}{L_k} \\ x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k \xi_{k-1}\right) \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}), \end{array} \right.$$

where we make the standard assumptions about our gradient estimator of being unbiased, i.e.

$$\mathbb{E}_k(\xi_k) = \nabla({}^{\mu_{k+1}}g \circ K)(y_k),$$

and having bounded variance

$$\mathbb{E}_k\left(\|\xi_k - \nabla({}^{\mu_{k+1}}g \circ K)(y_k)\|^2\right) \leq \sigma^2$$

for every $k \geq 0$.

Note that we use the same notations as in the deterministic case

$$u_k := x_{k-1} + t_k(x_k - x_{k-1}) \text{ and } F^k(\cdot) := f + {}^{\mu_k}g \circ K \quad \forall k \geq 1.$$

Lemma 4.1. *The following statement holds for every (deterministic) $z \in \mathcal{H}$ and every $k \geq 0$*

$$\mathbb{E}_k\left(F^{k+1}(x_{k+1}) + \frac{\|x_{k+1} - z\|^2}{2\gamma_{k+1}}\right) \leq F^{k+1}(z) + \frac{\|z - y_k\|^2}{2\gamma_{k+1}} + \gamma_{k+1}\left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2}\right)$$

Proof. Here we have to proceed a little bit different from Lemma 3.1. Namely, we have to treat the gradient step and the proximal step differently. For this purpose we define the auxiliary variable

$$z_k := y_{k-1} - \gamma_k \xi_{k-1} \quad \forall k \geq 1.$$

Let $k \geq 1$ be fixed. From the gradient step we get

$$\begin{aligned} \|z - z_k\|^2 &= \|y_{k-1} - \gamma_k \xi_{k-1} - z\|^2 \\ &= \|y_{k-1} - z\|^2 - 2\gamma_k \langle \xi_{k-1}, y_{k-1} - z \rangle + \gamma_k^2 \|\xi_{k-1}\|^2. \end{aligned}$$

Taking the conditional expectation gives

$$\mathbb{E}_{k-1}\left(\|z - z_k\|^2\right) = \|y_{k-1} - z\|^2 - 2\gamma_k \langle \nabla({}^{\mu_k}g \circ K)(y_{k-1}), y_{k-1} - z \rangle + \gamma_k^2 \mathbb{E}_{k-1}\left(\|\xi_{k-1}\|^2\right).$$

Using the gradient inequality we deduce

$$\begin{aligned} \mathbb{E}_{k-1}\left(\|z - z_k\|^2\right) \leq\ & \|y_{k-1} - z\|^2 - 2\gamma_k(({}^{\mu_k}g \circ K)(y_{k-1}) - ({}^{\mu_k}g \circ K)(z)) \\ & + \gamma_k^2 \mathbb{E}_{k-1}\left(\|\xi_{k-1}\|^2\right) \end{aligned}$$

and therefore

$$\begin{aligned}\gamma_k({}^{\mu_k}g \circ K)(y_{k-1}) + \frac{1}{2}\mathbb{E}_{k-1}\left(\|z - z_k\|^2\right) \leq \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k({}^{\mu_k}g \circ K)(z) \\ + \frac{\gamma_k^2}{2}\mathbb{E}_{k-1}\left(\|\xi_{k-1}\|^2\right).\end{aligned} \tag{24}$$

Also from the smoothness of $({}^{\mu_k}g \circ K)$ we deduce via the Descent Lemma that

$${}^{\mu_k}g(Kz_k) \leq {}^{\mu_k}g(Ky_{k-1}) + \langle \nabla({}^{\mu_k}g \circ K)(y_{k-1}), z_k - y_{k-1}\rangle + \frac{L_k}{2}\|z_k - y_{k-1}\|^2.$$

Plugging in the definition of z_k and using the fact that $L_k = \frac{1}{\gamma_k}$ we get

$${}^{\mu_k}g(Kz_k) \leq {}^{\mu_k}g(Ky_{k-1}) - \gamma_k \langle \nabla({}^{\mu_k}g \circ K)(y_{k-1}), \xi_{k-1}\rangle + \frac{\gamma_k}{2}\|\xi_{k-1}\|^2.$$

Now we take the conditional expectation to deduce that

$$\mathbb{E}_{k-1}({}^{\mu_k}g(Kz_k)) \leq {}^{\mu_k}g(Ky_{k-1}) - \gamma_k \|\nabla({}^{\mu_k}g \circ K)(y_{k-1})\|^2 + \frac{\gamma_k}{2}\mathbb{E}_{k-1}\left(\|\xi_{k-1}\|^2\right). \tag{25}$$

Multiplying (25) by γ_k and adding it to (24) gives

$$\begin{aligned}&\gamma_k\mathbb{E}_{k-1}\left({}^{\mu_k}g(Kz_k)\right) + \frac{1}{2}\mathbb{E}_{k-1}\left(\|z - z_k\|^2\right) \leq \\ &\qquad \gamma_k{}^{\mu_k}g(Kz) + \frac{1}{2}\|y_{k-1} - z\|^2 - \gamma_k^2 \|\nabla({}^{\mu_k}g \circ K)(y_{k-1})\|^2 + \gamma_k^2\mathbb{E}_{k-1}\left(\|\xi_{k-1}\|^2\right).\end{aligned}$$

Now we use the assumption about the bounded variance to deduce that

$$\gamma_k\mathbb{E}_{k-1}\left({}^{\mu_k}g(Kz_k)\right) + \frac{1}{2}\mathbb{E}_{k-1}\left(\|z - z_k\|^2\right) \leq \gamma_k{}^{\mu_k}g(Kz) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2\sigma^2. \tag{26}$$

Next up for the proximal step we deduce

$$f(x_k) + \frac{1}{2\gamma_k}\|x_k - z_k\|^2 + \frac{1}{2\gamma_k}\|x_k - z\|^2 \leq f(z) + \frac{1}{2\gamma_k}\|z - z_k\|^2. \tag{27}$$

Taking the conditional expectation and combining (26) and (27) we get

$$\begin{aligned}\mathbb{E}_{k-1}\left(\gamma_k({}^{\mu_k}g(Kz_k) + f(x_k)) + \frac{1}{2}\|x_k - z_k\|^2 + \frac{1}{2}\|x_k - z\|^2\right) \leq \\ \gamma_k F^k(z) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2\sigma^2.\end{aligned}$$

From here, using now Lemma 2.3, we get that

$$\begin{aligned}\mathbb{E}_{k-1}\left(\gamma_k F^k(x_k) - \gamma_k L_g\|K\|\|x_k - z_k\| + \frac{1}{2}\|x_k - z_k\|^2 + \frac{1}{2}\|x_k - z\|^2\right) \leq \\ \gamma_k F^k(z) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2\sigma^2.\end{aligned}$$

Now we use

$$-\frac{1}{2}\gamma_k^2 L_g^2\|K\|^2 \leq \frac{1}{2}\|x_k - z_k\|^2 - \gamma_k L_g\|K\|\|x_k - z_k\|$$

to obtain that

$$\mathbb{E}_{k-1}\left(\gamma_k F^k(x_k) + \frac{1}{2}\|x_k - z\|^2\right) \leq$$
$$\gamma_k F^k(z) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2\sigma^2 + \frac{1}{2}\gamma_k^2 L_g^2\|K\|^2.$$

□

Lemma 4.2. *Let x^* be an optimal solution of Problem 4.1. Then it holds*

$$\mathbb{E}\left(\gamma_1(F^1(x_1) - F^1(x^*))\right) + \frac{1}{2}\|u_1 - x^*\|^2 \leq \frac{1}{2}\|x_0 - x^*\|^2 + \gamma_1^2\sigma^2 + \frac{1}{2}\gamma_1^2 L_g^2\|K\|^2.$$

Proof. Applying the previous lemma with $k = 0$ and $z = x^*$, we get that

$$\mathbb{E}\left(\gamma_1 F^1(x_1) + \frac{1}{2}\|x_1 - x^*\|^2\right) \leq \gamma_1 F^1(x^*) + \frac{1}{2}\|y_0 - x^*\|^2 + \gamma_1^2\sigma^2 + \frac{1}{2}\gamma_1^2 L_g^2\|K\|^2.$$

Therefore, using the fact that $y_0 = x_0$ and $u_1 = x_1$,

$$\mathbb{E}\left(\gamma_1(F^1(x_1) - F^1(x^*)) + \frac{1}{2}\|u_1 - x^*\|^2\right) \leq \frac{1}{2}\|x_0 - x^*\|^2 + \gamma_1^2\sigma^2 + \frac{1}{2}\gamma_1^2 L_g^2\|K\|^2,$$

which finishes the proof. □

Theorem 4.1. *Consider the setup of Problem 4.1 and let $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ denote the sequences generated by Algorithm 4.1. Assume that for all $k \geq 1$*

$$\rho_{k+1} := t_k^2 - t_{k+1}^2 + t_{k+1} \geq 0.$$

Then, for every optimal solution x^* of Problem 4.1, it holds

$$\begin{aligned}\mathbb{E}\left(F(x_N) - F(x^*)\right) \leq{}& \frac{1}{\gamma_N t_N^2}\frac{1}{2}\|x_0 - x^*\|^2 + \frac{1}{\gamma_N t_N^2}\frac{\|K\|^2 L_g^2}{2}\sum_{k=1}^{N}\gamma_k^2(t_k + \rho_k) \\ &+ \frac{1}{\gamma_N t_N^2}\left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2}\right)\sum_{k=1}^{N} t_k^2\gamma_k^2 \quad \forall N \geq 1.\end{aligned}$$

Proof of Theorem 4.1. Let $k \geq 0$ be fixed. Lemma 4.1 for $z := \left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*$ gives

$$\mathbb{E}_k\left(F^{k+1}(x_{k+1}) + \frac{1}{2\gamma_{k+1}}\left\|\frac{1}{t_{k+1}}u_{k+1} - \frac{1}{t_{k+1}}x^*\right\|^2\right) \leq$$
$$F^{k+1}\left(\left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*\right) + \frac{1}{2\gamma_{k+1}}\left\|\frac{1}{t_{k+1}}x^* - \frac{1}{t_{k+1}}u_k\right\|^2$$
$$+\gamma_{k+1}\left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2}\right).$$

From here and from the convexity of F^{k+1} follows

$$\mathbb{E}_k\left(F^{k+1}(x_{k+1})-F^{k+1}(x^*)\right)-\left(1-\frac{1}{t_{k+1}}\right)(F^{k+1}(x_k)-F^{k+1}(x^*))\leq \\ \frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}-\mathbb{E}_k\left(\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\right)+\gamma_{k+1}\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).$$

Now, by multiplying both sides with by t_{k+1}^2, we deduce

$$\mathbb{E}_k\left(t_{k+1}^2(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\right)+(t_{k+1}-t_{k+1}^2)(F^{k+1}(x_k)-F^{k+1}(x^*))\leq \\ \frac{1}{2\gamma_{k+1}}\left(\|u_k-x^*\|^2-\mathbb{E}_k\left(\|u_{k+1}-x^*\|^2\right)\right)+t_{k+1}^2\gamma_{k+1}\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right). \tag{28}$$

Next, by adding $t_k^2(F^{k+1}(x_k)-F^{k+1}(x^*))$ on both sides of (28), gives

$$\mathbb{E}_k\left(t_{k+1}^2(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\right)+\rho_{k+1}(F^{k+1}(x_k)-F^{k+1}(x^*))\leq \\ t_k^2(F^{k+1}(x_k)-F^{k+1}(x^*))+\frac{1}{2\gamma_{k+1}}\left(\|u_k-x^*\|^2-\mathbb{E}_k\left(\|u_{k+1}-x^*\|^2\right)\right) \\ +t_{k+1}^2\gamma_{k+1}\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).$$

Utilizing (6) together with the assumption that $(\mu_k)_{k\geq 1}$ is nonincreasing leads to

$$\mathbb{E}_k\left(t_{k+1}^2(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\right)+\rho_{k+1}(F^{k+1}(x_k)-F^{k+1}(x^*))\leq \\ t_k^2(F^k(x_k)-F^k(x^*))+\frac{1}{2\gamma_{k+1}}\left(\|u_k-x^*\|^2-\mathbb{E}_k\left(\|u_{k+1}-x^*\|^2\right)\right)+t_k^2(\mu_k-\mu_{k+1})\frac{L_g^2}{2} \\ +t_{k+1}^2\gamma_{k+1}\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).$$

Now, using that $t_k^2\geq t_{k+1}^2-t_{k+1}$, we get

$$\mathbb{E}_k\left(t_{k+1}^2(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\right)+\rho_{k+1}(F^{k+1}(x_k)-F^{k+1}(x^*))\leq \\ t_k^2(F^k(x_k)-F^k(x^*))+\frac{1}{2\gamma_{k+1}}(\|u_k-x^*\|^2-\mathbb{E}_k\left(\|u_{k+1}-x^*\|^2\right)) \\ +t_k^2\mu_k\frac{L_g^2}{2}-t_{k+1}^2\mu_{k+1}\frac{L_g^2}{2}+t_{k+1}\mu_{k+1}\frac{L_g^2}{2} \\ +t_{k+1}^2\gamma_{k+1}\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).$$

Multiplying both sides with γ_{k+1} and putting all terms on the correct sides yields

$$\begin{aligned}\mathbb{E}_k\Bigg(\gamma_{k+1}t_{k+1}^2\left(F^{k+1}(x_{k+1})-F^{k+1}(x^*)+\mu_{k+1}\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_{k+1}-x^*\|^2\Bigg)+&\\ \gamma_{k+1}\rho_{k+1}(F^{k+1}(x_k)-F^{k+1}(x^*))\leq&\\ \gamma_{k+1}t_k^2\left(F^k(x_k)-F^k(x^*)+\mu_k\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_k-x^*\|^2+&\\ +\gamma_{k+1}t_{k+1}\mu_{k+1}\frac{L_g^2}{2}+t_{k+1}^2\gamma_{k+1}^2\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).&\end{aligned}\tag{29}$$

At this point we would like to discard the term $\gamma_{k+1}\rho_{k+1}(F^{k+1}(x_k)-F^{k+1}(x^*))$ which we currently cannot as the positivity of $F^{k+1}(x_k)-F^{k+1}(x^*)$ is not ensured. So we add $\gamma_{k+1}\rho_{k+1}\mu_{k+1}\frac{L_g^2}{2}$ on both sides of (29) and get

$$\begin{aligned}\mathbb{E}_k\Bigg(\gamma_{k+1}t_{k+1}^2\left(F^{k+1}(x_{k+1})-F^{k+1}(x^*)+\mu_{k+1}\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_{k+1}-x^*\|^2\Bigg)+&\\ \gamma_{k+1}\rho_{k+1}\left(F^{k+1}(x_k)-F^{k+1}(x^*)+\mu_{k+1}\frac{L_g^2}{2}\right)\leq&\\ \gamma_{k+1}t_k^2\left(F^k(x_k)-F^k(x^*)+\mu_k\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_k-x^*\|^2+&\\ +\gamma_{k+1}\mu_{k+1}\frac{L_g^2}{2}(t_{k+1}+\rho_{k+1})+t_{k+1}^2\gamma_{k+1}^2\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).&\end{aligned}\tag{30}$$

Using again (6) to deduce that

$$\gamma_{k+1}\rho_{k+1}\left(F^{k+1}(x_k)-F^{k+1}(x^*)+\mu_{k+1}\frac{L_g^2}{2}\right)\geq\gamma_{k+1}\rho_{k+1}(F(x_k)-F(x^*))\geq 0$$

we can now discard said term from (30), giving

$$\begin{aligned}\mathbb{E}_k\left(\gamma_{k+1}t_{k+1}^2\left(F^{k+1}(x_{k+1})-F^{k+1}(x^*)+\mu_{k+1}\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_{k+1}-x^*\|^2\right)\leq&\\ \gamma_{k+1}t_k^2\left(F^k(x_k)-F^k(x^*)+\mu_k\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_k-x^*\|^2&\\ +\gamma_{k+1}\mu_{k+1}\frac{L_g^2}{2}(t_{k+1}+\rho_{k+1})+t_{k+1}^2\gamma_{k+1}^2\left(\sigma^2+\frac{\|K\|^2L_g^2}{2}\right).&\end{aligned}\tag{31}$$

Last but not least we use the that $F^k(x_k)-F^k(x^*)+\mu_k\frac{L_g^2}{2}\geq F(x_k)-F(x^*)\geq 0$ and

$\gamma_{k+1} \leq \gamma_k$ to follow that

$$\gamma_{k+1} t_k^2 \left(F^k(x_k) - F^k(x^*) + \mu_k \frac{L_g^2}{2} \right) \leq \gamma_k t_k^2 \left(F^k(x_k) - F^k(x^*) + \mu_k \frac{L_g^2}{2} \right). \tag{32}$$

Combining (31) and (32) yields

$$\begin{aligned} \mathbb{E}_k \left(\gamma_{k+1} t_{k+1}^2 \left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \mu_{k+1} \frac{L_g^2}{2} \right) + \frac{1}{2}\|u_{k+1} - x^*\|^2 \right) \leq \\ \gamma_k t_k^2 \left(F^k(x_k) - F^k(x^*) + \mu_k \frac{L_g^2}{2} \right) + \frac{1}{2}\|u_k - x^*\|^2 \\ + \gamma_{k+1}\mu_{k+1}\frac{L_g^2}{2}(t_{k+1} + \rho_{k+1}) + t_{k+1}^2 \gamma_{k+1}^2 \left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2} \right). \end{aligned} \tag{33}$$

Let $N \geq 2$. We take the expected value on both sides (33) and sum from $k = 1$ to $N-1$. Getting rid of the non-negative terms $\|u_N - x^*\|^2$ gives

$$\begin{aligned} \mathbb{E}\left(\gamma_N t_N^2 \left(F^N(x_N) - F^N(x^*) + \mu_N \frac{L_g^2}{2} \right) \right) \leq \\ \mathbb{E}\left(\gamma_1 \left(F^1(x_1) - F^1(x^*) + \mu_1 \frac{L_g^2}{2} \right) \right) + \frac{1}{2}\|u_1 - x^*\|^2 + \sum_{k=2}^{N} \gamma_k \mu_k \frac{L_g}{2}(t_k + \rho_k) \\ + \sum_{k=2}^{N} t_k^2 \gamma_k^2 \left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2} \right). \end{aligned}$$

Since $t_1 = 1$, the above inequality holds also for $N = 1$. Now, using Lemma 4.2 we get that for every $N \geq 1$

$$\begin{aligned} \mathbb{E}\left(\gamma_N t_N^2 \left(F^N(x_N) - F^N(x^*) + \mu_N \frac{L_g^2}{2} \right) \right) \leq \frac{1}{2}\|x_0 - x^*\|^2 + \sum_{k=1}^{N} \gamma_k \mu_k \frac{L_g^2}{2}(t_k + \rho_k) \\ + \sum_{k=1}^{N} t_k^2 \gamma_k^2 \left(\sigma^2 + \frac{\|K\|^2}{2} \right). \end{aligned}$$

From (7) we follow that

$$\gamma_N t_N^2 \left(F(x_N) - F(x^*) \right) \leq \gamma_N t_N^2 \left(F^N(x_N) - F^N(x^*) + \mu_N \frac{L_g^2}{2} \right),$$

therefore, for every $N \geq 1$

$$\begin{aligned} \mathbb{E}\left(\gamma_N t_N^2 \left(F^N(x_N) - F^N(x^*) \right) \right) \leq \frac{1}{2}\|x_0 - x^*\|^2 + \sum_{k=1}^{N} \gamma_k \mu_k \frac{L_g^2}{2}(t_k + \rho_k) \\ + \sum_{k=1}^{N} t_k^2 \gamma_k^2 \left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2} \right). \end{aligned}$$

By using the fact that $\mu_k = \gamma_k \|K\|^2$ for every $k \geq 1$ gives

$$\mathbb{E}\left(\gamma_N t_N^2 (F(x_N) - F(x^*))\right) \leq \frac{1}{2}\|x_0 - x^*\|^2 + \frac{\|K\|^2 L_g^2}{2} \sum_{k=1}^{N} \gamma_k^2 (t_k + \rho_k) \\ + \left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2}\right) \sum_{k=1}^{N} t_k^2 \gamma_k^2 \quad \forall N \geq 1.$$

Thus,

$$\mathbb{E}\left(F(x_N) - F(x^*)\right) \leq \frac{1}{\gamma_N t_N^2} \frac{1}{2}\|x_0 - x^*\|^2 + \frac{1}{\gamma_N t_N^2} \frac{\|K\|^2 L_g^2}{2} \sum_{k=1}^{N} \gamma_k^2 (t_k + \rho_k) \\ + \frac{1}{\gamma_N t_N^2} \left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2}\right) \sum_{k=1}^{N} t_k^2 \gamma_k^2 \quad \forall N \geq 1.$$

□

Corollary 4.1. *Let*

$$t_1 = 1, \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \quad \forall k \geq 1,$$

and, for $b > 0$,

$$\mu_k = \frac{b}{k^{\frac{3}{2}}}\|K\|^2, \text{ and } \gamma_k = \frac{b}{k^{\frac{3}{2}}} \quad \forall k \geq 1.$$

Then,

$$\mathbb{E}\left(F(x_N) - F(x^*)\right) \leq 2\frac{\|x_0 - x^*\|^2}{b\sqrt{N}} + b\|K\|^2 L_g^2 \frac{\pi^2}{3} \frac{1}{\sqrt{N}} \\ + 2b\left(2\sigma^2 + \|K\|^2 L_g^2\right) \frac{1 + \log(N)}{\sqrt{N}} \quad \forall N \geq 1.$$

Furthermore, we have that $F(x_N)$ *converges almost surely to* $F(x^*)$ *as* $N \to +\infty$.

Proof. First we notice that the choice of $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ fulfills that

$$\rho_{k+1} = t_k^2 - t_{k+1}^2 + t_{k+1} = 0 \quad \forall k \geq 1.$$

Now we derive the stated convergence result by first showing via induction that

$$\frac{1}{k} \leq \frac{1}{t_k} \leq \frac{2}{k} \quad \forall k \geq 1.$$

Assuming that this holds for $k \geq 1$, we have that

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \leq \frac{1 + \sqrt{1 + 4k^2}}{2} \leq \frac{1 + \sqrt{1 + 4k + 4k^2}}{2} = k + 1$$

and

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \geq \frac{1+\sqrt{1+4(\frac{k}{2})^2}}{2} \geq \frac{1+\sqrt{k^2}}{2} \geq \frac{k+1}{2}.$$

Furthermore, for every $N \geq 1$ we have that

$$\begin{aligned} \frac{1}{\gamma_N t_N^2} \frac{\|K\|^2 L_g^2}{2} \sum_{k=1}^{N} \gamma_k^2 (t_k + \rho_k) &\leq \frac{4}{b\sqrt{N}} \frac{\|K\|^2 L_g^2}{2} \sum_{k=1}^{N} \frac{b^2}{k^3} k = \frac{2b\|K\|^2 L_g^2}{\sqrt{N}} \sum_{k=1}^{N} k^{-2} \\ &\leq \frac{2b\|K\|^2 L_g^2}{\sqrt{N}} \sum_{k=1}^{\infty} k^{-2} = b\|K\|^2 L_g^2 \frac{\pi^2}{3} \frac{1}{\sqrt{N}}. \end{aligned} \tag{34}$$

The statement of the convergence rate in expectation follows now by plugging in our parameter choices into the statement of Theorem 4.1, using the estimate (34) and checking that

$$\sum_{k=1}^{N} t_k^2 \gamma_k^2 \leq b^2 \sum_{k=1}^{N} \frac{1}{k} \leq b^2(1+\log(N)) \quad \forall N \geq 1.$$

The almost sure convergence of $(F(x_N))_{N\geq 1}$ can be deduced by looking at (33) and dividing by $\gamma_{k+1}t_{k+1}^2$ and using that $\gamma_{k+1}t_{k+1}^2 \geq \gamma_k t_k^2$ as well as $\rho_k = 0$, which gives for every $k \geq 0$

$$\mathbb{E}_k \left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \mu_{k+1}\frac{L_g^2}{2} + \frac{1}{2\gamma_{k+1}t_{k+1}^2}\|u_{k+1} - x^*\|^2 \right) \leq$$
$$F^k(x_k) - F^k(x^*) + \mu_k \frac{L_g^2}{2} + \frac{1}{2\gamma_k t_k^2}\|u_k - x^*\|^2 + \frac{\mu_{k+1}}{t_{k+1}} \frac{L_g^2}{2} + \gamma_{k+1}\left(\sigma^2 + \frac{\|K\|^2 L_g^2}{2}\right).$$

Plugging in our choice of parameters gives for every $k \geq 0$

$$\mathbb{E}_k \left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \mu_{k+1}\frac{L_g^2}{2} + \frac{1}{2\gamma_{k+1}t_{k+1}^2}\|u_{k+1} - x^*\|^2 \right) \leq$$
$$F^k(x_k) - F^k(x^*) + \mu_k \frac{L_g^2}{2} + \frac{1}{2\gamma_k t_k^2}\|u_k - x^*\|^2 + \frac{C}{k^{\frac{3}{2}}},$$

where $C > 0$.

Thus, by the famous Robbins-Siegmund Theorem (see [25, Theorem 1]) we get that $(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \mu_{k+1}\frac{L_g^2}{2})_{k\geq 0}$ converges almost surely. In particular, from the convergence to 0 in expectation we know that the almost sure limit must also be the constant zero. □

Finite Sum. The formulation of the previous section can be used to deal e.g. with problems of the form

$$\min_{x\in\mathcal{H}} f(x) + \sum_{i=1}^{m} g_i(K_i x) \tag{35}$$

for $f : \mathcal{H} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function, $g_i : \mathcal{G}_i \to \mathbb{R}$ convex and L_{g_i}-Lipschitz continuous functions and $K_i : \mathcal{H} \to \mathcal{G}_i$ linear continuous operators for $i = 1, \dots, m$.

Clearly one could consider

$$\boldsymbol{K} := \begin{cases} \mathcal{H} \to \times_{i=1}^m \mathcal{G}_i \\ x \mapsto \times_{i=1}^m K_i x \end{cases}$$

with $\|\boldsymbol{K}\|^2 = \sum_{i=1}^m \|K_i\|^2$ and

$$\boldsymbol{g} := \begin{cases} \times_{i=1}^m \mathcal{G}_i \to \overline{\mathbb{R}} \\ \times_{i=1}^m y_i \mapsto \sum_{i=1}^m g_i(y_i). \end{cases}$$

in order to reformulate the problem as

$$\min_{x \in \mathcal{H}} f(x) + \boldsymbol{g}(\boldsymbol{K}x)$$

and use Algorithm 3.1 together with the parameter choices described in Corollary 3.1 on this. This results in the following algorithm.

Algorithm 4.2. *Let $y_0 = x_0 \in \mathcal{H}, \mu_1 = b\|\boldsymbol{K}\|$, for $b > 0$, and $t_1 = 1$. Consider the following iterative scheme*

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} \gamma_k = \frac{\sum_{i=1}^m \|K_i\|^2}{\mu_k} \\ x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k \sum_{i=1}^m K_i^* \operatorname{prox}_{\frac{1}{\mu_k} g_i^*}\left(\frac{K_i y_{k-1}}{\mu_k} \right) \right) \\ t_{k+1} = \sqrt{t_k^2 + 2t_k} \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \\ \mu_{k+1} = \mu_k \frac{t_k^2}{t_{k+1}^2 - t_{k+1}}. \end{array} \right.$$

However, problem (35) also lends itself to be tackled via the stochastic version of our method, Algorithm 4.1, by randomly choosing a subset of the summands. Together with the parameter choices described in Corollary 4.1 which results in the following scheme.

Algorithm 4.3. *Let $y_0 = x_0 \in \mathcal{H}, b > 0$, and $t_1 = 1$. Consider the following iterative scheme*

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} \mu_k = b \sum_{i=1}^m \|K_i\|^2 k^{-\frac{3}{2}} \\ \gamma_k = b k^{-\frac{3}{2}} \\ x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k \frac{\epsilon_{i,k}}{p_i} \sum_{i=1}^m K_i^* \operatorname{prox}_{\frac{1}{\mu_k} g_i^*}\left(\frac{K_i y_{k-1}}{\mu_k} \right) \right) \\ t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}), \end{array} \right.$$

with $\epsilon_k := (\epsilon_{1,k}, \epsilon_{2,k}, \dots, \epsilon_{m,k})$ a sequence of i.i.d., $\{0,1\}^m$ random variables and $p_i = \mathbb{P}[\epsilon_{i,1} = 1]$.

Since the above two methods were not explicitly developed for this separable case and can therefore not make use of more refined estimation of the constant $\|\boldsymbol{K}\|$, as it is done in e.g. [14]. However, in the stochastic case, this fact is remedied due to the scaling of the stepsize with respect to the i-th component by p_i^{-1}.

Remark 4.2. In theory Algorithm 4.1 could be used to treat more general stochastic problems than finite sums like (35), but in the former case it is not clear anymore how a gradient estimator can be found, so we do not discuss it here.

5 Numerical Examples

We will focus our numerical experiments on image processing problems. The examples are implemented in python using the operator discretization library (ODL) [1]. We define the discrete gradient operators D_1 and D_2 representing the discretized derivative in the first and second coordinate respectively, which we will need for the numerical examples. Both map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times n}$ and are defined by

$$(D_1 u)_{i,j} := \begin{cases} u_{i+1,j} - u_{i,j} & 1 \leq i < m, \\ 0 & \text{else,} \end{cases}$$

and

$$(D_2 u)_{i,j} := \begin{cases} u_{i,j+1} - u_{i,j} & 1 \leq j < m, \\ 0 & \text{else.} \end{cases}$$

The operator norm of D_1 and D_2, respectively, is 2 (where we equipped $\mathbb{R}^{m\times n}$ with the Frobenius norm). This yields an operator norm of $\sqrt{8}$ for the total gradient $D := D_1 \times D_2$ as a map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$, see also [12].

We will compare our methods, i.e. the Variable Accelerated SmooThing (VAST) and its stochastic counterpart (sVAST) to the Primal Dual Hybrid Gradient (PDHG) of [15] as well as its stochastic version (sPDHG) from [14]. Furthermore, we will illustrate another competitor, the method by Pesquet and Repetti, see [24], which is another a stochastic version of PDHG (see also [29]).

In all examples we choose the parameters in accordance with [14]:

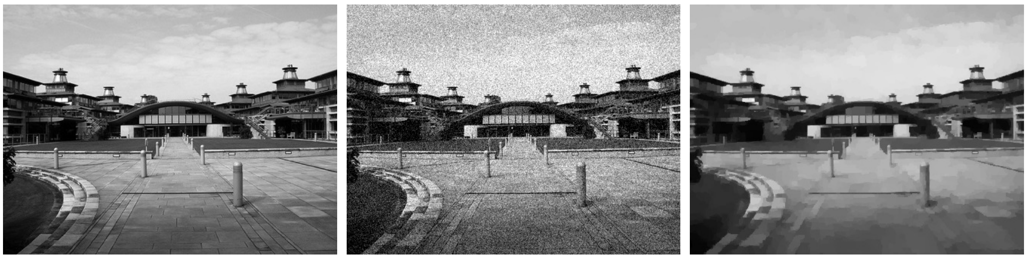
- for PDHG and Pesquet&Repetti: $\tau = \sigma_i = \frac{\gamma}{\|K\|}$
- for sPDHG: $\sigma_i = \frac{\gamma}{\|K\|}$ and $\tau = \frac{\gamma}{n \max_i \|K_i\|}$,

where $\gamma = 0.99$.

5.1 Total Variation Denoising

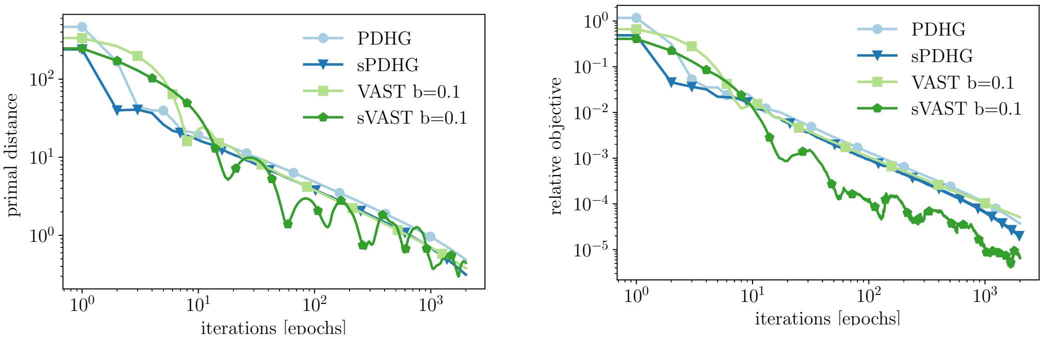
The task at hand is to reconstruct an image from its noisy observation. We do this by solving

$$\min_{x\in\mathbb{R}^{m\times n}} \alpha\|x-b\|_2 + \|D_1 x\|_1 + \|D_2 x\|_1,$$



(a) Groundtruth (b) Data (c) Approximate solution

Figure 1: TV denoising. Images used. The approximate solution is computed by running PDHG for 7000 iterations.



(a) Distance to the solution. (b) Relative objective $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$.

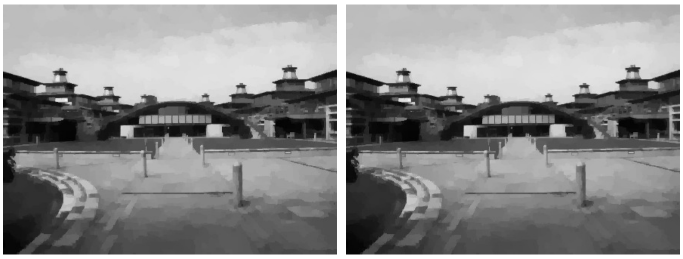
Figure 2: TV denoising. Plots.

with $\alpha > 0$ as regularization parameter, in the following setting: $f = \alpha\|\cdot - b\|_2, g_1 = g_2 = \|\cdot\|_1, K_1 = D_1, K_2 = D_2$.

Figure 1 illustrates the images (of dimension $m = 442$ and $n = 331$) used in for this example. These include the groundtruth, i.e. the uncorrupted image, as well as the data for the optimization problem b, which visualizes the level of noise. In Figure 2 we can see that for the deterministic setting our method is as good as PDHG. For the objective function values, Subfigure 2b, this is not too surprising as both algorithms share the same convergence rate. For the distance to a solution however we completely lack a convergence result. Nevertheless in Subfigure 2a we can see that our method performs also well with respect to this measure.

In the stochastic setting we can see in Figure 2 that, while sPDHG provides some benefit over its deterministic counterpart, the stochastic version of our method, although significantly increasing the variance, provides great benefit, at least for the objective function values.

Furthermore, Figure 3, shows the reconstructions of sPDHG and our method which are, despite the different objective function values, quite comparable.



(a) sVAST (b) sPDHG

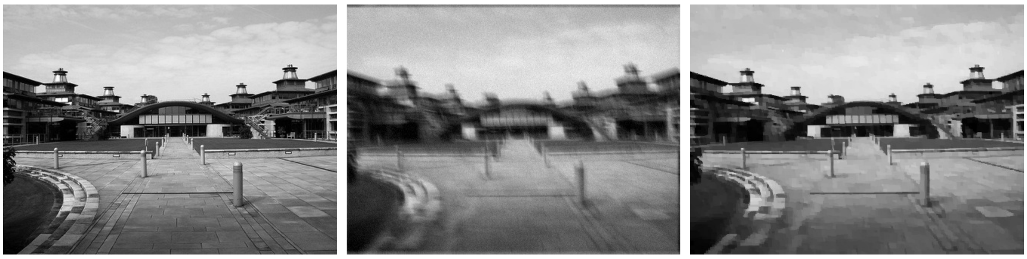
Figure 3: TV Denoising. A comparison of the reconstruction for the stochastic variable smoothing method and the stochastic PDHG.

5.2 Total Variation Deblurring

For this example we want to reconstruct an image from a blurred and noisy image. We assume to know the blurring operator $C : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$. This is done by solving

$$\min_{x\in\mathbb{R}^{m\times n}} \alpha\|Cx - b\|_2 + \|D_1 x\|_1 + \|D_2 x\|_1, \tag{36}$$

for $\alpha > 0$ as regularization parameter, in the following setting: $f = 0, g_1 = \alpha\|\cdot - b\|_2, g_2 = g_3 = \|\cdot\|_1, K_1 = C, K_2 = D_1, K_2 = D_2$.

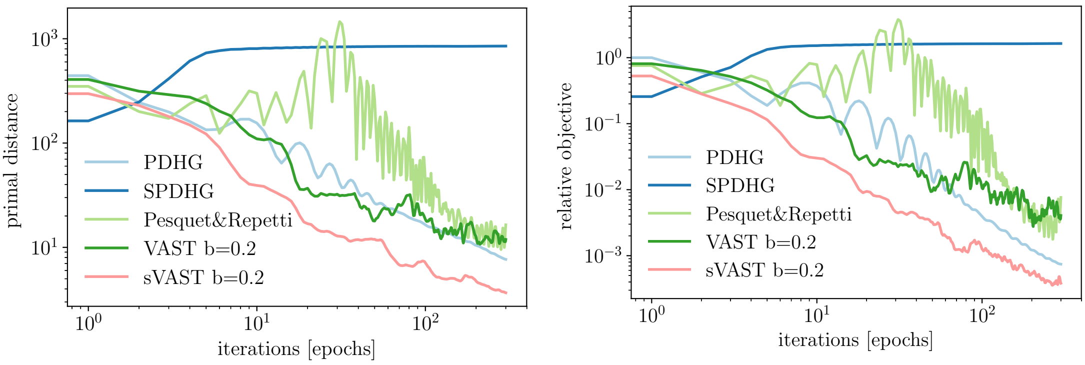


(a) Groundtruth (b) Data (c) Approximate solution

Figure 4: TV Deblurring. The approximate solution is computed by running PDHG for 3000 iterations.

Figure 4 shows the images used to set up the optimization problem (36), in particular Subfigure 4b which corresponds to b in said problem.

In Figure 5 we see that while PDGH performs better in the deterministic setting, in particular in the later iteration, the stochastic variable smoothing method provides a significant improvement where sPDHG method seems not to converge. It is interesting to note that in this setting even the deterministic version of our algorithm exhibits a slightly chaotic behaviour. Although neither of the two methods is monotone in the primal objective function PDHG seems here much more stable.



(a) Distance to the solution.

(b) Relative objective $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$.

Figure 5: TV deblurring. Plots.

5.3 Matrix Factorization

In this section we want to solve a *nonconvex* and nonsmooth optimization problem of completely positive matrix factorization, see [16, 19, 27]. For an observed matrix $A \in \mathbb{R}^{d\times d}$ we want to find a completely positive low rank factorization, meaning we are looking for $x \in \mathbb{R}^{r\times d}_{\geq 0}$ with $r \ll d$ such that $x^T x = A$. This can be formulated as the following optimization problem

$$\min_{x\in\mathbb{R}^{r\times d}_{\geq 0}} \|x^T x - A\|_1, \tag{37}$$

where x^T denotes the transpose of the matrix x. The more natural approach might be to use a smooth formulation where $\|\cdot\|_2^2$ is used instead of the 1-Norm we are suggesting. However, the former choice of distance measure, albeit smooth, comes with its own set of problems (mainly a non-Lipschitz gradient).

The so called *Prox-Linear method* presented in [18] solves the above problem (37), by linearizing the smooth ($\mathbb{R}^{d\times d}$-valued) function $x \mapsto x^T x$ inside the nonsmooth distance function. In particular for the problem

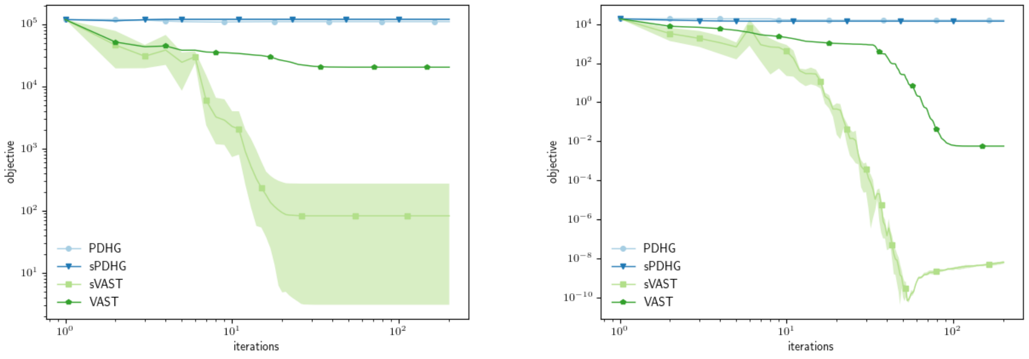
$$\min_x g(c(x)),$$

for a smooth vector valued function c and a convex and Lipschitz function g, [18] proposes to iteratively solve the subproblem

$$x_{k+1} = \arg\min_x \left\{ g(c(x_k) + \nabla c(x_k)(x - x_k)) + \frac{1}{2t}\|x - x_k\|_2^2 \right\} \quad \forall k \geq 0, \tag{38}$$

for a stepsize $t \leq (L_g L_{D\nabla c})^{-1}$. For our particular problem described in (37) the subproblem looks as follows

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^{r\times d}_{\geq 0}} \left\{ \|x_k^T x - A\|_1 + \frac{1}{2}\|x - x_k\|_2^2 \right\}, \tag{39}$$



(a) Random starting point.

(b) Starting point close to the solution.

Figure 6: Comparison of the evolutions of the objective function values for different starting points. We run 40 epochs with 5 iterations each. For each epoch we choose the last iterate of the previous epoch as the linearization. For the stochastic methods we fix the number of rows (batchsize) which are randomly chosen in each update a priori and count d divided by this number as one iteration. For the randomly chosen initial point we use a batchsize of 3 (to allow for more exploration) and for the one close to the solution we use 5 in order to give a more accuracy. The parameter b in the variable smoothing method was chosen with minimal tuning to be 0.1 for both the deterministic and the stochastic version.

and therefore fits our general setup described in (1) with the identification $f = \|\cdot - x_k\|_2^2 + \delta_{\mathbb{R}^{r\times d}_{\geq 0}}(x)$, $g = \|\cdot\|_1$ and $K = x_k^T$. Moreover, due to its separable structure, the subproblem (39) fits the special case described in (35) and can therefore be tackled by the stochastic version of our algorithm presented in Algorithm 4.3. In particular reformulating (38) for the stochastic finite sum setting we interpret the subproblem as

$$x_{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^{r\times d}_{\geq 0}} \left\{ \sum_{i=1}^{d} \left\| x_k^T[i,:]x - A[i,:] \right\|_1 + \frac{1}{2}\|x - x_k\|_2^2 \right\},$$

where $A[i,:]$ denotes the i-th row of the matrix A.

In comparison to Section 5.1 and Section 5.2 a new aspect becomes important when evaluating methods for solving (38). Now, it is not only relevant how well subproblem (39) is solved, but also the trajectory taken in doing so as different paths might lead to different local minima. This can be seen in Figure 6 where PDHG gets stuck early on in bad local minima. The variable smoothing method (especially the stochastic version) is able to move further from the starting point and find better local minima. Note that in general the methods have a difficulty in finding the global minimum $x_{true} \in \mathbb{R}^{3\times 60}$ (with optimal objective function value zero, as constructed $A := x_{true}^T x_{true} \in \mathbb{R}^{60\times 60}$ in all examples).

Acknowledgements. The authors are thankful to two anonymous reviewers for comments and remarks which improved the quality of the presentation and led to the numerical experiment on matrix factorization.

References

[1] Jonas Adler, Holger Kohr, and Ozan Öktem. Operator Discretization Library, https://odlgroup.github.io/odl/, 2017.

[2] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, Springer, New York, 2011.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] Jonathan M Borwein and Jon D Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*, Cambridge University Press, Cambridge, 2010.

[5] Radu I Boţ and Ernö R Csetnek. On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems. *Optimization*, 64(1):5–23, 2015.

[6] Radu I Boţ, Ernö R Csetnek, André Heinrich, and Christopher Hendrich. On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. *Mathematical Programming*, 150(2):251–279, 2015.

[7] Radu I Boţ and Christopher Hendrich. A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. *Computational Optimization and Applications*, 54(2):239–262, 2013.

[8] Radu I Boţ and Christopher Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4):2541–2565, 2013.

[9] Radu I Boţ and Christopher Hendrich. Convergence analysis for a primal-dual monotone+ skew splitting algorithm with applications to total variation minimization. *Journal of Mathematical Imaging and Vision*, 49(3):551–568, 2014.

[10] Radu I Boţ and Christopher Hendrich. On the acceleration of the double smoothing technique for unconstrained convex optimization problems. *Optimization*, 64(2):265–288, 2015.

[11] Radu I Boţ and Christopher Hendrich. A variable smoothing algorithm for solving convex optimization problems. *TOP*, 23(1):124–150, 2015.

[12] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.

[13] Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". *Journal of Optimization theory and Applications*, 166(3):968–982, 2015.

[14] Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[15] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[16] Chen Chen, Ting Kei Pong, Lulin Tan, and Liaoyuan Zeng. A difference-of-convex approach for split feasibility with applications to matrix factorizations and outlier detection. *Journal of Global Optimization*, DOI: 10.1007/s10898-020-00899-8, 2020.

[17] Laurent Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[18] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:1–56, 2019.

[19] Patrick Groetzner and Mirjam Dür. A factorization method for completely positive matrices. *Linear Algebra and its Applications*, 591:1–24, 2020.

[20] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[21] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady Akademija Nauk USSR*, 269:543–547, 1983.

[22] Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2007.

[23] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, Springer Science & Business Media, New York, 2013.

[24] Jean-Christophe Pesquet and Audrey Repetti. A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis*, 16(12):2453–2490, 2015.

[25] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In: *Optimizing Methods in Statistics*, Proceedings of a Symposium Held at the Center for Tomorrow, Ohio State University, June 14–16, 1971, pages 233–257, Elsevier, 1971.

[26] Lorenzo Rosasco, Silvia Villa, and Băng C Vũ. A first-order stochastic primal-dual algorithm with correction step. *Numerical Functional Analysis and Optimization*, 38(5):602–626, 2017.

[27] Qingjiang Shi, Haoran Sun, Songtao Lu, Mingyi Hong, and Meisam Razaviyayn. Inexact block coordinate descent methods for symmetric nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 65(22): 5995–6008, 2017.

[28] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018.

[29] Băng C Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.