Variance Reduced Coordinate Descent with Acceleration: New Method With a Surprising Application to Finite-Sum Problems

Filip Hanzely [1] **Dmitry Kovalev** [1] **Peter Richtárik** [1]

Abstract

We propose an accelerated version of stochastic variance reduced coordinate descent – ASVRCD. As other variance reduced coordinate descent methods such as SEGA or SVRCD, our method can deal with problems that include a non-separable and non-smooth regularizer, while accessing a random block of partial derivatives in each iteration only. However, ASVRCD incorporates Nesterov's momentum, which offers favorable iteration complexity guarantees over both SEGA and SVRCD. As a by-product of our theory, we show that a variant of Katyusha (Allen-Zhu, 2017) is a specific case of ASVRCD, recovering the optimal oracle complexity for the finite sum objective.

1. Introduction

In this paper, we aim to solve the regularized optimization problem

$$\min_{x\in\mathbb{R}^d} P(x) := f(x) + \psi(x), \tag{1}$$

where function f is convex and differentiable, while the regularizer ψ is convex and non-smooth. Furthermore, we assume that the dimensionality d is large.

The most standard approach to deal with the huge d is to decompose the space, i.e., use coordinate descent, or, more generally, subspace descent methods (Nesterov, 2012; Wright, 2015; Kozak et al., 2019). Those methods are especially popular as they achieve a linear convergence rate on strongly convex problems while enjoying a relatively cheap cost of performing each iteration.

However, coordinate descent methods are only feasible if the regularizer ψ is separable (Richtárik & Takáč, 2014). In contrast, if ψ is not separable, the corresponding stochastic gradient estimator has an inherent (non-zero) variance at the optimum, and thus the linear convergence rate is not achievable.

This phenomenon is, to some extent, similar when applying Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951; Nemirovski et al., 2009) on finite sum objective – the corresponding stochastic gradient estimator has a (non-zero) variance at the optimum, which prevents SGD from converging linearly. Recently, the issue of sublinear convergence of SGD has been resolved using the idea of control variates (Hickernell et al., 2005), resulting in famous variance reduced methods such as SVRG (Johnson & Zhang, 2013) and SAGA (Defazio et al., 2014a).

Motivated by the massive success of variance reduced methods for finite sums, control variates have been proposed to "fix" coordinate descent methods to minimize problem (1) with non-separable ψ. To best of our knowledge, there are two such algorithms in the literature – SEGA (Hanzely et al., 2018) and SVRCD (Hanzely & Richtárik, 2019), which we now quickly describe.[1]

Let x^k be the current iterate of SEGA (or SVRCD) and suppose that the oracle reveals the i-th partial derivative $\nabla_i f(x^k)$ (for i chosen uniformly at random). The simplest unbiased gradient estimator of $\nabla f(x^k)$ can be constructed as $\tilde{g}^k = d\nabla_i f(x^k)e_i$ (where $e_i \in \mathbb{R}^d$ is the i-th standard basis vector). The idea behind these methods is to enrich $\tilde{g}^k$ using a control variate $h^k \in \mathbb{R}^d$, resulting in a new (still unbiased) gradient estimator g^k:

$$g^k = d\nabla_i f(x^k)e_i - dh_i^k e_i + h^k,$$

where $h_i^k \in \mathbb{R}$ stands for the i-th element of vector h^k. *How to choose the sequence of control variates* $\{h^k\}$*?* Intuitively, we wish for both sequences $\{h^k\}$ and $\{\nabla f(x^k)\}$ to have an identical limit point. In such case, we have $\lim_{k\to\infty} \mathrm{Var}(g^k) = 0$, and thus one shall expect faster convergence. There is no unique way of setting $\{h^k\}$ to have the mentioned property satisfied – this is where SEGA and SVRCD differ. See Algorithm 1 for details.

[1]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Correspondence to: Filip Hanzely <filip.hanzely@kaust.edu.sa>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]VRSSD (Kozak et al., 2019) is yet another stochastic subspace descent algorithm aided by control variates; however, it was

Algorithm 1 SEGA and SVRCD

Require: Stepsize $\alpha > 0$, starting point $x^0 \in \mathbb{R}^d$, probability vector p: $p_i := \mathbb{P}(i \in S)$
Set $h^0 = 0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
Sample random $S \subseteq \{1, 2, \ldots, d\}$
$g^k = \sum_{i \in S} \frac{1}{p_i}(\nabla_i f(x^k) - h_i^k)e_i + h^k$
$x^{k+1} = \text{prox}_{\alpha\psi}(x^k - \alpha g^k)$
$h^{k+1} = \begin{cases} h^k + \sum_{i \in S}(\nabla_i f(x^k) - h_i^k)e_i & \text{for SEGA} \\ \begin{cases} \nabla f(x^k) & \text{w.p. } \rho \\ h^k & \text{w.p. } 1-\rho \end{cases} & \text{for SVRCD} \end{cases}$
end for

In this work, we continue the above research along the lines of variance reduced coordinate descent algorithms, with surprising consequences.

1.1. Contributions

Here we list the main contributions of this paper.

▷ **Exploiting prox in SEGA/SVRCD.** Assume that the regularizer ψ includes an indicator function of some affine subspace of $\mathbb{R}^d$. We show that both SEGA and SVRCD might exploit this fact, resulting in a faster convergence rate. As a byproduct, we establish the same result in the more general framework from (Hanzely & Richtárik, 2019) (presented in the appendix).

▷ **Accelerated SVRCD.** We propose an accelerated version of SVRCD - ASVRCD. ASVRCD is the first accelerated variance reduced coordinate descent to minimize objectives with non-separable, proximable regularizer.[2]

▷ **SEGA/SVRCD/ASCRVD generalizes SAGA/L-SVRG/L-Katyusha.** We show a surprising link between SEGA and SAGA. In particular, SAGA is a special case of SEGA; and the new rate we obtain for SEGA recovers the tight complexity of SAGA (Qian et al., 2019b; Gazagnadou et al., 2019). Similarly, we recover loopless SVRG (L-SVRG) (Hofmann et al., 2015; Kovalev et al., 2020) along with its best-known rate (Hanzely & Richtárik, 2019; Qian et al., 2019a) using a result for SVRCD. Lastly, as a particular case of ASVRCD, we recover an algorithm which is marginally preferable to loopless Katyusha (L-Katyusha) (Qian et al., 2019a): while we recover their iteration complexity result, our proof is more straightforward, and at the same time, the stepsize for the proximal operator is smaller.[3]

1.2. Preliminaries

As mentioned in Section 1.1, the new results we provide are particularly interesting if the regularizer ψ contains an indicator function of some affine subspace of $\mathbb{R}^d$. This step is crucial in order to recover the tight convergence rate of the finite-sum variance reduced methods, as we shall see later.

Assumption 1.1 *Assume that* $\mathbf{W}$ *is an orthogonal projection matrix such that*

$$\psi(x) = \begin{cases} \psi'(x) & if \quad x \in \{x^0 + \text{Range}(\mathbf{W})\} \\ \infty & if \quad x \notin \{x^0 + \text{Range}(\mathbf{W})\} \end{cases} \tag{2}$$

for some convex function $\psi'(x)$*. Furthermore, suppose that the proximal operator of* ψ *is cheap to compute.*

Example 1.1 *Given that* $\psi' = 0$*, the regularizer* ψ *becomes the indicator function of the affine subspace given by* $\{x^0 + \text{Range}(\mathbf{W})\}$.

Remark 1.2 *If* ψ *is convex, there is always some* $\mathbf{W}$ *such that* (2) *holds as one might choose* $\mathbf{W} = \mathbf{I}$.

Next, we require smoothness of the objective, as well as the strong convexity over the affine subspace given by $\text{Range}(\mathbf{W})$.

Assumption 1.2 *Function* f *is* $\mathbf{M}$*-smooth, i.e., for all* $x, y \in \mathbb{R}^d$*:*[4]

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\|x - y\|_{\mathbf{M}}^2.$$

Function f *is* μ*-strongly convex over* $\{x^0 + \text{Range}(\mathbf{W})\}$*, i.e., for all* $x, y \in \{x^0 + \text{Range}(\mathbf{W})\}$*:*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2. \tag{3}$$

Remark 1.3 *Smoothness with respect to matrix* $\mathbf{M}$ *arises naturally in various applications. For example, if* $f(x) = f'(\mathbf{A}x)$*, where* f' *is* L'*-smooth (for scalar* $L' > 0$*), we can derive that* f *is* $\mathbf{M} = L'\mathbf{A}^\top\mathbf{A}$*-smooth.*

In order to stress the distinction between the finite sum setup and the setup from the rest of the paper, we are denoting the finite-sum variables that differ from the non-finite sum case in red and with an extra tilde on top of them. We thus, *recommend printing this paper in color*.

proposed to minimize f only (i.e., considers $\psi = 0$).

[2] We shall note that an accelerated version of SEGA was already proposed in (Hanzely et al., 2018) for $\psi = 0$ – this was rather an impractical result demonstrating that SEGA can match state-of-the art convergence rate of accelerated coordinate descent from (Allen-Zhu et al., 2016; Nesterov & Stich, 2017; Hanzely & Richtárik, 2018). In contrast, our results cover any convex ψ.

[3] This is preferable especially if the proximal operator has to be estimated numerically.

[4] We define $\|x\|^2 := \langle x, x \rangle$ and $\|x\|_{\mathbf{M}}^2 := \langle \mathbf{M}x, x \rangle$.

2. Better Rates for SEGA and SVRCD

In this section, we show that a specific structure of non-smooth function ψ might lead to faster convergence of SEGA and SVRCD.

The next lemma is a direct consequence of Assumption 1.1 – it shows that proximal operator of ψ is contractive under $\mathbf{W}$-norm.

Lemma 2.1 *Let $\{x^k\}_{k\geq 0}$ be a sequence of iterates of Algorithm 1 and let x^* be optimal solution of (1). Then*

$$x^k \in \{x^0 + \text{Range}(\mathbf{W})\},\ x^* \in \{x^0 + \text{Range}(\mathbf{W}))\}. \quad (4)$$

for all k. Furthermore, for any $x, y \in \mathbb{R}^d$ we have

$$\|\operatorname{prox}_{\alpha\psi}(x) - \operatorname{prox}_{\alpha\psi}(y)\|^2 \leq \|x - y\|^2_{\mathbf{W}}. \quad (5)$$

Next, we state the convergence rate of both SEGA and SVRCD under Assumption 1.1 as Theorem 2.2. We also generalize the main theorem from (Hanzely & Richtárik, 2019) (fairly general algorithm which covers SAGA, SVRG, SEGA, SVRCD, and more as a special case; see Section D of the appendix); from which the convergence rate of SEGA/SVRCD follows as a special case.

Theorem 2.2 *Let Assumptions 1.1, 1.2 hold and denote $p_i \coloneqq \mathbb{P}(i \in S)$. Consider vector $v = \sum_{i=1}^d e_i v_i, v_i \geq 0$ such that*

$$\mathbf{M}^{\frac{1}{2}}\mathbb{E}\left[\sum_{i\in S}\frac{1}{p_i}e_i e_i^\top \mathbf{W} \sum_{i\in S}\frac{1}{p_i}e_i e_i^\top\right]\mathbf{M}^{\frac{1}{2}} \preceq \mathbf{D}(p^{-1}\circ v), \quad (6)$$

where $\mathbf{D}(\cdot)$ is a diagonal operator which returns a matrix with the input on the diagonal, and zeros everywhere else. Then, iteration complexity of SEGA with $\alpha = \min_i \frac{p_i}{4v_i+\mu}$ is $\max_i\left(\frac{4v_i+\mu}{p_i\mu}\right)\log\frac{1}{\epsilon}$. At the same time, iteration complexity of SVRCD with $\alpha = \min_i \frac{1}{4v_i {p_i}^{-1}+\mu\rho^{-1}}$ is $\left(\frac{4\max_i(v_i {p_i}^{-1})+\mu\rho^{-1}}{\mu}\right)\log\frac{1}{\epsilon}$.

Let us look closer to convergence rate of SVRCD from Theorem 2.2. The optimal vector v is a solution to the following optimization problem

$$\min_{v\in\mathbb{R}^d}\ \left(\frac{4\max_i\{v_i {p_i}^{-1}\}+\mu\rho^{-1}}{\mu}\right)\log\frac{1}{\epsilon}\ \text{ s. t. (6) holds.}$$

Clearly, there exists a solution of the form $v \propto p$; let us thus choose $v \coloneqq \mathcal{L}p$ with $\mathcal{L} > 0$. In this case, to satisfy (6) we must have

$$\mathcal{L} = \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbb{E}\left[\sum_{i\in S}\frac{1}{p_i}e_i e_i^\top \mathbf{W} \sum_{i\in S}\frac{1}{p_i}e_i e_i^\top\right]\mathbf{M}^{\frac{1}{2}}\right) \quad (7)$$

and the iteration complexity of SVRCD becomes[5]

$$\left(\frac{4\mathcal{L}+\mu\rho^{-1}}{\mu}\right)\log\frac{1}{\epsilon}.$$

How does $\mathbf{W}$ influence the rate? As mentioned, one can always consider $\mathbf{W} = \mathbf{I}$. In such a case, we recover the convergence rate of SEGA and SVRCD from (Hanzely & Richtárik, 2019). However, the smaller rank of $\mathbf{W}$ is, the faster rate is Theorem 2.2 providing. To see this, it suffices to realize that if $\mathcal{L}$ is increasing in $\mathbf{W}$ (in terms of Loewner ordering).

Example 2.3 *Let $\mathbf{M} = \mathbf{I}$ and $S = \{i\}$ with probability d^{-1} for all $1 \leq i \leq d$. Given that $\mathbf{W} = \mathbf{I}$, it is easy to see that $\mathcal{L} = d$. In such case, the iteration complexity of SVRCD is $\left(\frac{4d+\mu\rho^{-1}}{\mu}\right)\log\frac{1}{\epsilon}$. In the other extreme, if $\mathbf{W} = \frac{1}{d}ee^\top$, we have $\mathcal{L} = 1$, which yields complexity (of SVRCD) $\left(\frac{4+\mu\rho^{-1}}{\mu}\right)\log\frac{1}{\epsilon}$. Therefore, given that $\mu = \mathcal{O}(\rho)$, the low rank of $\mathbf{W}$ caused the speedup of order $\Theta(d)$.*

We shall also note that the tight rate of SAGA and L-SVRG might be recovered from Theorem 2.2 only using a non-trivial $\mathbf{W}$ (see Section 3), while the original theory of SEGA and SVRCD only yield a suboptimal rate for both SAGA and L-SVRG.

Connection with Subspace SEGA (Hanzely et al., 2018). Assume that function f is of structure $f(x) = h(\mathbf{A}x)$. As a consequence, we have $\nabla f(x) = \mathbf{A}^\top \nabla h(\mathbf{A}x)$ and thus $\nabla f(x) \in \text{Range}\left(\mathbf{A}^\top\right)$. This fact was exploited by Subspace SEGA in order to achieve a faster convergence rate. Our results can mimic Subspace SEGA by setting ψ to be an indicator function of $x^0 + \text{Range}\left(\mathbf{A}^\top\right)$, given that there is no extra non-smooth term in the objective.

Remark 2.4 *Throughout all proofs of this section, we have used a weaker conditions than Assumption 1.2. In particular, instead of-$\mathbf{M}$-smoothness, it is sufficient to have[6] $D_f(x, x^*) \geq \frac{1}{2}\|\nabla f(x) - \nabla f(x^*)\|^2_{\mathbf{M}^{-1}}$ for all $x \in \mathbb{R}^d$ (Lemma D.3 shows that it is indeed a consequence of $\mathbf{M}$ smoothness and convexity). At the same time, instead of μ-strong convexity, it is sufficient to have μ-quasi strong convexity, i.e., for all $x \in \{x^0 + \text{Range}(\mathbf{W})\}$: $f(x^*) \geq f(x) + \langle\nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2$. However, the accelerated method (presented in Section 4) requires the fully general version of Assumption 1.2.*

[5] We decided to not present this, simplified rate in Theorem 2.2 for the following two reasons: 1) it would yields a slightly subpotimal rate of SEGA and 2) the connection of to the convergence rate of SAGA from (Qian et al., 2019b) is more direct via (6).

[6] By $D_f(x, y)$ we denote Bregman distance between x, y, i.e., $D_f(x, y) \coloneqq f(x) - f(y) - \langle\nabla f(x)$

3. Connection between SEGA (SVRCD) and SAGA (L-SVRG)

In this section, we show that SAGA and L-SVRG are special cases of SEGA and SVRCD, respectively. At the same time, the previously tightest convergence rate of SAGA (Gazagnadou et al., 2019; Qian et al., 2019b) and L-SVRG (Hanzely & Richtárik, 2019; Qian et al., 2019a) follow from Theorem 2.2 (convergence rate of SEGA and SVRCD).

3.1. Convergence rate of SAGA and L-SVRG

We quickly state the best-known convergence rate for both SAGA and L-SVRG to minimize the following objective:

$$\min_{\tilde{x}\in\mathbb{R}^{\tilde{d}}} \tilde{P}(\tilde{x}) := \underbrace{\frac{1}{n}\sum_{j=1}^{n} \tilde{f}_j(\tilde{x})}_{:=\tilde{f}(\tilde{x})} + \tilde{\psi}(\tilde{x}). \tag{8}$$

Assumption 3.1 *Each $\tilde{f}_j$ is convex, $\tilde{\mathbf{M}}_j$-smooth and $\tilde{f}$ is $\tilde{\mu}$-strongly convex.*

Assuming the oracle access to $\nabla \tilde{f}_i(\tilde{x}^k)$ for $i \in \tilde{S}$ (where $\tilde{S}$ is a random subset of $\{1,\dots,n\}$), the minibatch SGD (Gower et al., 2019) uses moves in the direction of the "plain" unbiased stochastic gradient $\frac{1}{n}\sum_{i\in\tilde{S}} \frac{1}{\tilde{p}_i}\nabla \tilde{f}_i(\tilde{x}^k)$, where $\tilde{p}_i := \mathbb{P}\left(i \in \tilde{S}\right)$.

In contrast, variance reduced methods such as SAGA and L-SVRG enrich the "plain" unbiased stochastic gradient with control variates:

$$\tilde{g}^k = \frac{1}{n}\sum_{i\in\tilde{S}} \frac{1}{\tilde{p}_i}\left(\nabla \tilde{f}_i(\tilde{x}^k) - \mathbf{J}^k_{:,i}\right) + \frac{1}{n}\mathbf{J}^k\tilde{e}. \tag{9}$$

where $\mathbf{J}^k \in \mathbb{R}^{\tilde{d}\times n}$ is the control matrix and $\tilde{e} \in \mathbb{R}^n$ is vector of ones. The difference between SAGA and L-SVRG lies in the procedure to update $\mathbf{J}^k$; SAGA uses the freshest gradient information to replace corresponding columns in $\mathbf{J}^k$; i.e.

$$\mathbf{J}^{k+1}_{:,i} = \begin{cases} \nabla \tilde{f}_i(\tilde{x}^k) & \text{if } i \in \tilde{S} \\ \mathbf{J}^k_{:,i} & \text{if } i \notin \tilde{S}. \end{cases} \tag{10}$$

On the other hand, L-SVRG sets $\mathbf{J}^k$ to the true Jacobian of f upon a successful, unfair coin toss:

$$\mathbf{J}^{k+1} = \begin{cases} \left[\nabla \tilde{f}_1(\tilde{x}^k), \dots, \nabla \tilde{f}_n(\tilde{x}^k)\right] & \text{w. p. } \rho \\ \mathbf{J}^k & \text{w. p. } 1-\rho. \end{cases} \tag{11}$$

The formal statement of SAGA and L-SVRG is provided as Algorithm 2, while Proposition 3.1 states their convergence rate.

Algorithm 2 SAGA/L-SVRG

Require: $\alpha > 0$, $\rho \in (0,1)$
$\tilde{x}^0 \in \mathbb{R}^{\tilde{d}}$, $\mathbf{J}^0 = 0 \in \mathbb{R}^{d\times n}$
for $k = 0, 1, 2, \dots$ **do**
  Sample random $\tilde{S} \subseteq \{1,\dots,n\}$
  $\tilde{g}^k = \frac{1}{n}\mathbf{J}^k\tilde{e} + \frac{1}{n}\sum_{i\in\tilde{S}} \frac{1}{\tilde{p}_i}(\nabla \tilde{f}_i(\tilde{x}^k) - \mathbf{J}^k_{:,i})$
  $\tilde{x}^{k+1} = \text{prox}_{\tilde{\alpha}\tilde{\psi}}(\tilde{x}^k - \tilde{\alpha}\tilde{g}^k)$
  Update $\mathbf{J}^{k+1}$ according to (10) or (11)
end for

Proposition 3.1 *Suppose that Assumption 3.1 holds and let $\tilde{v}$ be a nonegative vector such that for all $h_1,\dots,h_n \in \mathbb{R}^{\tilde{d}}$ we have*

$$\mathbb{E}\left[\left\|\sum_{j\in\tilde{S}} \tilde{\mathbf{M}}_j^{\frac{1}{2}} h_j\right\|^2\right] \le \sum_{j=1}^{n} \tilde{p}_j\tilde{v}_j \left\|h_j\right\|^2. \tag{12}$$

Then the iteration complexity of SAGA with $\tilde{\alpha} = \min_j \frac{n\tilde{p}_j}{4\tilde{v}_j + n\tilde{\mu}}$ is $\max_j \left(\frac{4\tilde{v}_j + n\tilde{\mu}}{n\tilde{\mu}\tilde{p}_j}\right)\log\frac{1}{\epsilon}$. At the same time, iteration complexity of L-SVRG with $\tilde{\alpha} = \min_j \frac{n}{4\frac{\tilde{v}_j}{\tilde{p}_j} + \frac{\tilde{\mu}n}{\rho}}$ is $\max_j \left(4\frac{\tilde{v}_j}{n\tilde{\mu}\tilde{p}_j} + \frac{1}{\rho}\right)\log\frac{1}{\epsilon}$.

3.2. SAGA is a special case of SEGA

Consider setup from Section 3.1; i.e., problem (8) along with Assumption 3.1 and $\tilde{v}$ defined according to (12). We will construct an instance of (1) (i.e., specific f, ψ), which is equivalent to (8), such that applying SEGA on (1) is equivalent applying SAGA on (8).

Let $d := \tilde{d}n$. For convenience, define $R_j := \{\tilde{d}(j-1)+1, \tilde{d}(j-1)+1, \dots, \tilde{d}j\}$ (i.e., $|R_j| = \tilde{d}$) and lifting operator $Q(\cdot) : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^d$ defined as $Q(\tilde{x}) := \left[\underbrace{\tilde{x}^\top, \dots, \tilde{x}^\top}_{n \text{ times}}\right]^\top$.

Construction of f, ψ. Let I be indicator function of the set[7] $x_{R_1} = \cdots = x_{R_n}$ and choose

$$f(x) := \frac{1}{n}\sum_{j=1}^{n} \tilde{f}_j(x_{R_j}), \quad \psi(x) := I(x) + \tilde{\psi}(x_{R_1}) \tag{13}$$

Now, it is easy to see that problem (8) and problem (1) with the choice (13) are equivalent; each $x \in \mathbb{R}^d$ such that $P(x) < \infty$ must be of the form $x = Q(\tilde{x})$ for some $\tilde{x} \in \mathbb{R}^{\tilde{d}}$. In such case, we have $P(x) = \tilde{P}(\tilde{x})$. The next lemma goes further, and derives the values $\mathbf{M}$, μ, $\mathbf{W}$ and v based on $\tilde{\mathbf{M}}_i$ $(1 \le i \le n)$, $\tilde{\mu}$, $\tilde{v}$.

[7] Indicator function of a set returns 0 for each point inside of the set and ∞ for each point outside of the set.

Lemma 3.2 *Consider f, ψ defined by (13). Function f satisfies Assumption 1.2 with $\mu := \frac{\tilde{\mu}}{n}$ and $\mathbf{M} := \frac{1}{n}$BlockDiag$(\tilde{\mathbf{M}}_1, \ldots, \tilde{\mathbf{M}}_n)$. Function ψ and $x^0 = Q\left(\tilde{x}^0\right)$ satisfy Assumption with $\mathbf{W} := \frac{1}{n}\tilde{e}\tilde{e}^\top \otimes \mathbf{I}$. At the same time, given that $\tilde{v}$ satisfies (12), inequaility (6) holds with $v = \tilde{v}n^{-1}$.*

Next, we show that running Algorithm 1 in this particular setup is equivalent to running Algorithm 2 for the finite sum objective.

Lemma 3.3 *Consider f, ψ from (13), S as described in the last paragraph and $x^0 = Q\left(\tilde{x}^0\right)$. Running SEGA (SVRCD) on (1) with $S := \cup_{j \in \tilde{S}} R_j$ and $\alpha := n\tilde{\alpha}$ is equivalent to running SAGA (L-SVRG) on (8).; i.e., we have for all k*

$$x^k = Q\left(\tilde{x}^k\right). \tag{14}$$

As a consequence of Lemmas 3.2 and 3.3, we get the next result.

Corollary 3.4 *Let f, ψ, S be as described above. Convergence rate of SAGA (L-SVRG) given by Proposition 3.1 to solve (1) is identical to convergence rate of SEGA (SVRCD) given by Theorem 2.2.*

4. Accelerated SVRCD

In this section we present SVRCD with Nesterov's momentum (Nesterov, 1983) – ASVRCD. The development of ASVRCD along with the theory (Theorem 4.1) was motivated by Katyusha (Allen-Zhu, 2017), ASVRG (Shang et al., 2018) and their loopless variants (Kovalev et al., 2020; Qian et al., 2019a). In Section 5.2, we show that a variant of L-Katyusha (Algorithm 4) is a special case of ASVRCD, and argue that it is slightly superior to the methods mentioned above.

The main component of ASVRCD is the gradient estimator g^k constructed analogously to SVRCD. In particular, g^k is an unbiased estimator of $\nabla f(x^k)$ controlled by $\nabla f(w^k)$:[8]

$$g^k = \nabla f(w^k) + \sum_{i \in S} \frac{1}{p_i}(\nabla_i f(x^k) - \nabla_i f(w^k))e_i. \tag{15}$$

Next, ASVRCD requires two more sequences of iterates $\{y^k\}_{k\geq 0}, \{z^k\}_{k \geq 0}$ in order to incorporate Nesterov's momentum. The update rules of those sequences consist of subtracting g^k alongside with convex combinations or interpolations of the iterates. See Algorithm 3 for specific formulas.

We are now ready to present ASVRCD along with its convergence guarantees.

[8] This is efficient to implement as sequence of iterates $\{w^k\}$ is updated rarely.

Algorithm 3 Accelerated SVRCD (ASVRCD)

Require: $0 < \theta_1, \theta_2 < 1$, $\eta, \beta, \gamma > 0$, $\rho \in (0,1)$, $y^0 = z^0 = x^0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

$x^k = \theta_1 z^k + \theta_2 w^k + (1 - \theta_1 - \theta_2) y^k$

Sample random $S \subseteq \{1, 2, \ldots, d\}$

$g^k = \nabla f(w^k) + \sum_{i \in S} \frac{1}{p_i}(\nabla_i f(x^k) - \nabla_i f(w^k))e_i$

$y^{k+1} = \text{prox}_{\eta\psi}(x^k - \eta g^k)$

$z^{k+1} = \beta z^k + (1-\beta)x^k + \frac{\gamma}{\eta}(y^{k+1} - x^k)$

$w^{k+1} = \begin{cases} y^k, & \text{with probability } \rho \\ w^k, & \text{with probability } 1-\rho \end{cases}$

end for

Theorem 4.1 *Let Assumption 1.1, 1.2 hold and denote $L := \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{W}\mathbf{M}^{\frac{1}{2}}\right)$. Further, let $\mathcal{L}'$ be such that for all k we have*

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \leq 2\mathcal{L}' D_f(w^k, x^k). \tag{16}$$

Define the following Lyapunov function:

$$\Psi^k := \left\|z^k - x^*\right\|^2 + \frac{2\gamma\beta}{\theta_1}\left[P(y^k) - P(x^*)\right] + \frac{(2\theta_2 + \theta_1)\gamma\beta}{\theta_1 \rho}\left[P(w^k) - P(x^*)\right],$$

and let

$$\begin{aligned} \eta &= \frac{1}{4}\max\{\mathcal{L}', L\}^{-1}, \\ \theta_2 &= \frac{\mathcal{L}'}{2\max\{L, \mathcal{L}'\}}, \\ \gamma &= \frac{1}{\max\{2\mu, 4\theta_1/\eta\}}, \\ \beta &= 1 - \gamma\mu \text{ and} \\ \theta_1 &= \min\left\{\frac{1}{2}, \sqrt{\eta\mu\max\left\{\frac{1}{2}, \frac{\theta_2}{\rho}\right\}}\right\}. \end{aligned}$$

Then the following inequality holds:

$$\mathbb{E}\left[\Psi^{k+1}\right] \leq \left[1 - \frac{1}{4}\min\left\{\rho, \sqrt{\frac{\mu}{2\max\left\{L, \frac{\mathcal{L}'}{\rho}\right\}}}\right\}\right]\Psi^0.$$

As a consequence, iteration complexity of Algorithm 3 is

$$\mathcal{O}\left(\left(\frac{1}{\rho} + \sqrt{\frac{L}{\mu}} + \sqrt{\frac{\mathcal{L}'}{\rho\mu}}\right)\log\frac{1}{\epsilon}\right).$$

Convergence rate of ASVRCD depends on constant $\mathcal{L}'$ such that (16) holds. The next lemma shows that $\mathcal{L}'$ can be obtained indirectly from $\mathbf{M}$-smoothness (via $\mathcal{L}$), in which case

the convergence rate provided by Theorem 4.1 significantly simplifies.

Lemma 4.2 *Inequality 16 holds for $\mathcal{L}' = \mathcal{L}$ (defined in (7)). Further, we have $L \leq \mathcal{L}$. Therefore, setting $\rho \geq \sqrt{\frac{\mu}{\mathcal{L}}}$ yields the following complexity of ASVRCD:*

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}}{\rho\mu}}\log\frac{1}{\epsilon}\right). \tag{17}$$

Setting $\mathcal{L}' = \mathcal{L}$ might be, however, loose in some cases. In particular, inequality (16) is slightly weaker than (6) and consequently, the bound bound from Theorem 4.1 is slightly better than (17). To see this, notice that the proof of Lemma 4.2 bounds variance of $g^k + \nabla f(w^k)$ by its second moment. Admittedly, this bound might not worsen the rate by more than a constant factor when $\frac{\mathbb{E}[|S|]}{d}$ is not close to 1. Therefore, bound (17) is good in essentially all practical cases. The next reason why we keep inequality (16) is that an analogous assumption was required for the analysis of L-Katyusha in (Qian et al., 2019a) (see Section 5.1) – and so we can now recover L-Katyusha results directly.

Let us give a quick taste how the rate of ASVRCD behaves depending on $\mathbf{W}$. In particular, Lemma 4.3 shows that nontrivial $\mathbf{W}$ might lead to speedup of order $\Theta(\sqrt{d})$ for ASVRCD.

Lemma 4.3 *Let $S = i$ for each $1 \leq i \leq d$ with probability $\frac{1}{d}$ and $\rho = \frac{1}{d}$. Then, if $\mathbf{W} = \mathbf{I}$, iteration complexity of ASVRCD is $\mathcal{O}\left(d\sqrt{\frac{\lambda_{\max}\mathbf{M}}{\mu}}\log\frac{1}{\epsilon}\right)$. If, however, $\mathbf{W} = \frac{1}{d}ee^\top$, iteration complexity of ASVRCD is $\mathcal{O}\left(\sqrt{\frac{d\lambda_{\max}\mathbf{M}}{\mu}}\log\frac{1}{\epsilon}\right)$.*

5. Connection between ASVRCD and L-Katyusha

Next, we show that L-Katyusha can be seen as a particular case of ASVRCD.

5.1. Convergence rate of L-Katyusha

In this section, we quickly introduce the loopless Katyusha (L-Katyusha) from (Qian et al., 2019a) along with its convergence guarantees. In the next section, we show that an improved version of L-Katyusha can be seen as a special case of ASVRCD, and at the same time, the tight convergence guarantees from (Qian et al., 2019a) can be obtained as a special case of Theorem 4.1.

Consider problem (8) and suppose that $\tilde{f}$ is $\tilde{L}$-smooth and $\tilde{\mu}$-strongly convex. Let $\tilde{S}$ be a random subset of $\{1,\dots,n\}$ (sampled from arbitrary fixed distribution) such that $\tilde{p}_i := \mathbb{P}\left(i \in \tilde{S}\right)$. For each k let $\tilde{g}^k$ be the following unbiased, variance reduced estimator of $\nabla\tilde{f}(x^k)$:

$$\tilde{g}^k = \frac{1}{n}\left(\sum_{i\in\tilde{S}}\tilde{p}_i^{-1}\left(\nabla\tilde{f}_i(\tilde{x}^k) - \nabla\tilde{f}_i(\tilde{w}^k)\right)\right) + \nabla\tilde{f}(\tilde{w}^k).$$

Next, L-Katyusha requires the variance of $\tilde{g}^k$ to be bounded by Bregman distance between $\tilde{w}^k$ and $\tilde{x}^k$ with constant $\tilde{\mathcal{L}}$, as the next assumption states.

Assumption 5.1 *For all k we have*

$$\mathbb{E}\left[\|\tilde{g}^k - \nabla\tilde{f}(\tilde{x}^k)\|^2\right] \leq 2\tilde{\mathcal{L}}D_f(\tilde{w}^k,\tilde{x}^k). \tag{18}$$

Proposition 5.1 provides a convergence rate of L-Katyusha.

Proposition 5.1 *(Qian et al., 2019a) Let $\tilde{f}$ be $\tilde{L}$-smooth and $\tilde{\mu}$-strongly convex while Assumption 5.1 holds. Iteration complexity of L-Katyusha is*

$$\mathcal{O}\left(\left(\frac{1}{\tilde{p}} + \sqrt{\frac{\tilde{L}}{\tilde{\mu}}} + \sqrt{\frac{\tilde{\mathcal{L}}}{\tilde{\mu}\tilde{p}}}\right)\log\frac{1}{\epsilon}\right).$$

5.2. L-Katyusha is a special case of ASVRCD

In this section, we show that a modified version of L-Katyusha (Algorithm 4) is a special case of ASVRCD. Furthermore, we show that the tight convergence rate of L-Katyusha (Qian et al., 2019a) follows from Theorem 4.1 (convergence rate of ASVRCD).

Consider again f, ψ chosen according to (13). With this choice, problem (1) and (8) are equivalent. At the same time, Lemma 3.3 establishes that f satisfies Assumption 1.2 with $\mu = \frac{\tilde{\mu}}{n}$ and $\mathbf{M} = \frac{1}{n}\text{BlockDiag}(\tilde{\mathbf{M}}_1,\dots,\tilde{\mathbf{M}}_n)$ while ψ and x^0 satisfy Assumption with $\mathbf{W} := \frac{1}{n}\tilde{e}\tilde{e}^\top \otimes \mathbf{I}$.

Note that the update rule of sequences x^k, z^k, w^k are identical for both algorithms; we shall thus verify that the update rule on y^k is identical as well. The last remaining thing is to relate $\mathcal{L}'$ and $\tilde{\mathcal{L}}$. The next lemma establishes both results.

Lemma 5.2 *Running ASVRCD on (1) with $S := \cup_{j\in\tilde{S}}R_j$ and $\eta := n\tilde{\eta}$, $\gamma := n\tilde{\gamma}$ is equivalent to running Algorithm 4 on (8). At the same time, inequality 16 holds with $\mathcal{L}' = n^{-1}\tilde{\mathcal{L}}$, while we have $L = n^{-1}\tilde{L}$.*

As a direct consequence of Lemma 5.2 and Theorem 4.1, we obtain the next corollary.

Corollary 5.3 *Let f, ψ, S be as described above. Iteration complexity of Algorithm 4 is*

$$\mathcal{O}\left(\left(\frac{1}{\tilde{p}}+\sqrt{\frac{\tilde{L}}{\tilde{\mu}}}+\sqrt{\frac{\tilde{\mathcal{L}}}{\tilde{\mu}\tilde{p}}}\right)\log\frac{1}{\epsilon}\right).$$

As promised, the convergence rate of Algorithm 4 matches the convergence rate of L-Katyusha from Proposition 5.1 and thus matches the lower bound for finite sum minimization by Lan & Zhou (2018); Woodworth & Srebro (2016). Let us now argue that Algorithm 4 is slightly superior to other accelerated SVRG variants.

First, Algortihm 4 is loopless; thus has a simpler analysis and slightly better properties (as shown by Kovalev et al. (2020)) over Katyusha (Allen-Zhu, 2017) and ASVRG (Shang et al., 2018). Next, the analysis is simpler than (Qian et al., 2019a) (i.e., we do not require one page of going through special cases). At the same time, Algorithm 4 uses a smaller stepsize for the proximal operator than L-Katyusha, which is useful if the proximal operator does is estimated numerically. However, Algorithm 4 is almost indistinguishable from L-Katyusha if $\tilde{\psi}=0$.

Remark 5.4 *The convergence rate of L-Katyusha from (Qian et al., 2019a) allows exploiting the strong convexity of regularizer ψ (given that it is strongly convex). While such a result is possible to obtain in our case, we have omitted it for simplicity.*

Algorithm 4 Variant of L-Katyusha (special case of Algorithm 3)

Require: $0<\theta_1,\theta_2<1$, $\tilde{\eta},\beta,\tilde{\gamma}>0$, $\rho\in(0,1)$

$\tilde{y}^0=\tilde{z}^0=\tilde{x}^0\in\mathbb{R}^{\tilde{d}}$

for $k=0,1,2,\ldots$ **do**

$\tilde{x}^k=\theta_1\tilde{z}^k+\theta_2\tilde{w}^k+(1-\theta_1-\theta_2)\tilde{y}^k$

Sample random $\tilde{S}\subseteq\{1,2,\ldots,n\}$

$g^k=\nabla\tilde{f}(\tilde{w}^k)+\sum_{i\in\tilde{S}}\frac{1}{\tilde{p}_i}(\nabla\tilde{f}_i(\tilde{x}^k)-\nabla\tilde{f}_i(\tilde{w}^k))$

$\tilde{y}^{k+1}=\mathrm{prox}_{\tilde{\eta}\psi}(\tilde{x}^k-\tilde{\eta}\tilde{g}^k)$

$\tilde{z}^{k+1}=\beta\tilde{z}^k+(1-\beta)\tilde{x}^k+\frac{\tilde{\gamma}}{\tilde{\eta}}(\tilde{y}^{k+1}-\tilde{x}^k)$

$\tilde{w}^{k+1}=\begin{cases}\tilde{y}^k, & \text{with probability } \rho\\ \tilde{w}^k, & \text{with probability } 1-\rho\end{cases}$

end for

6. Experiments

In this section, we numerically verify the performance of ASVRCD, as well as the improved performance of SVRCD under Assumption 1.1. In order to better understand and control the experimental setup, we consider a quadratic minimization (four different types) over the unit ball intersected with a linear subspace.[9] The specific choice of the objective is presented in Section 6.1.

In the first experiment we demonstrate the superiority of ASVRCD to SVRCD for problems with $\mathbf{W}=\mathbf{I}$. We consider four different methods – ASVRCD and SVRCD, both with uniform and importance sampling such that $|S|=1$ with probability 1. The importance sampling is the same as one from (Hanzely & Richtárik, 2019). In short, the goal is to have $\mathcal{L}$ from (7) as small as possible. Using $\mathbf{W}=\mathbf{I}$, it is easy to see that $\mathcal{L}=\lambda_{\max}\left(\mathbf{D}(p)^{-\frac{1}{2}}\mathbf{M}\mathbf{D}(p)^{-\frac{1}{2}}\right)$. While the optimal p is still hard to find, we set $p_i\propto\mathbf{M}_{i,i}$ (i.e., the effect of importance sampling is the same as the effect of Jacobi preconditioner). Figure 1 shows the result. As expected, accelerated SVRCD always outperforms non-accelerated variant, while at the same time, the importance sampling improves the performance too.

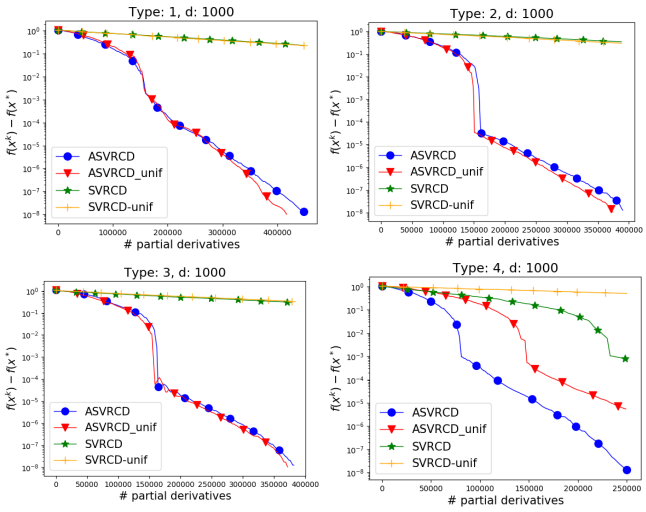

Figure 1. Comparison of both ASVRCD and SVRCD with importance and uniform sampling.

The second experiment compares the performance of both ASVRCD and SVRCD for various $\mathbf{W}$. We only consider methods with the importance sampling ($p_i\propto\mathbf{M}_{i,i}\mathbf{W}_{i,i}$) and theory supported stepsize. Figure 2 presents the result. We see that the smaller $\mathrm{Range}\,(\mathbf{W})$ is, the faster the convergence is. This observation is well-aligned with our theory: $\mathcal{L}$ is increasing as a function of $\mathbf{W}$ (in terms of Loewner ordering).

[9] Note that the practicality of ASVRCD immediately follows as it recovers Algorithm 4 as a special case, which is (especially for $\psi=0$) almost indistinguishable to L-Katyusha – state-of-the-art method for smooth finite sum minimization. For this reason, we decided to focus on less practical, but better-understood experiments.

Table 1. Choice of $\mathbf{M}$. O_{dd} is set of all odd positive integers smaller than $d+1$, while matrix $\mathbf{U}$ was set as random orthonormal matrix (generated by QR decomposition from a matrix with independent standard normal entries).

Type	$\mathbf{M}$	Figure 1: L	Figure 2: L
1	$\mathbf{U}\left(\mathbf{I}+\mathbf{I}_{:,O_{dd}}\mathbf{D}\left(((L-1)\frac{1}{500})^{(1:500)}\right)\mathbf{I}_{O_{dd},:}\right)\mathbf{U}^\top$	100	1000
2	$\mathbf{U}\left(\mathbf{I}+\sum_{i=1}^{100}(L-1)e_ie_i^\top\right)\mathbf{U}^\top$	100	1000
3	$\mathbf{U}\left(\kappa\mathbf{I}-\sum_{i=1}^{100}(L-1)e_ie_i^\top\right)\mathbf{U}^\top$	100	1000
4	$\left(\mathbf{I}+\frac{L}{500}\mathbf{I}_{:,O_{dd}}\mathbf{D}\,(1:500)\,\mathbf{I}_{O_{dd},:}\right)$	100	1000

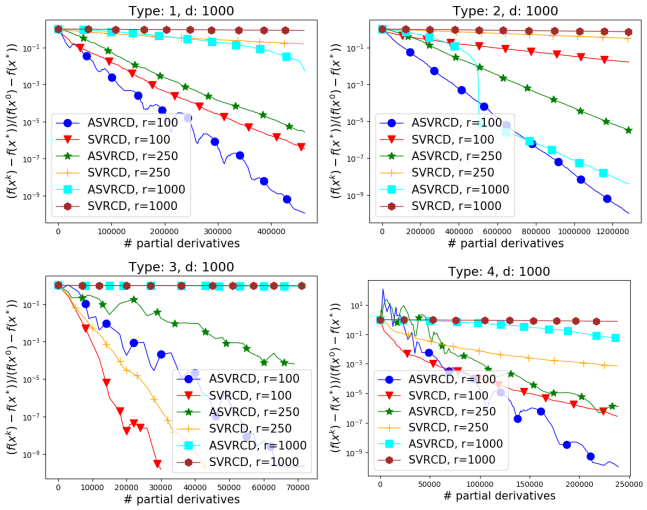

Figure 2. Comparison of ASVRCD and SVRCD for various $\mathbf{W}$. Label 'r' indicates the dimension of $\mathrm{Range}\,(\mathbf{W})$.

6.1. Experiments: The choice of the objective

In all experiments from this paper, we have chosen $f(x)=\frac{1}{2}x^\top\mathbf{M}x-b^\top x$, where $x\in\mathbb{R}^{1000}$, while ψ is an indicator function of the unit ball intersected with $\mathrm{Range}\,(\mathbf{W})$. First, matrix $\mathbf{M}$ was chosen according to Table 1. Next, vector b was chosen as follows: first we generate $\tilde{x}\in\mathbb{R}^d$ with independent normal entries, then compute $\tilde{b}=\mathbf{M}^{-1}\tilde{x}$ and set $b=\frac{3}{2\|\tilde{b}\|}\tilde{b}$. Lastly, for Figure 2, the projection matrix $\mathbf{W}$ of rank r was chosen as a block diagonal matrix with r blocks, each of them being the matrix of ones multiplied by $\frac{r}{d}$.

7. Implications

Finite sum algorithms are a special case of methods with partial derivative oracle. Using the trick described in Sections 3 and 5.2, it is possible to show that essentially any finite-sum stochastic algorithm is a special case of analogous method with partial derivative oracle (those are yet to be discovered/analyzed) in a given setting (i.e., strongly convex, convex, non-convex). Those include, but are not limited to SGD (Robbins & Monro, 1951; Nemirovski et al., 2009), over-parametrized SGD (Vaswani et al., 2018), SAG (Roux et al., 2012), SVRG (Johnson & Zhang, 2013), S2GD (Konečný & Richtárik, 2017), SARAH (Nguyen et al., 2017), incremental methods such as Finito (Defazio et al., 2014b), MISO (Mairal, 2015) or accelerated algorithms such as point-SAGA (Defazio, 2016), Katyusha (Allen-Zhu, 2017), MiG (Zhou et al., 2018b), SAGA-SSNM (Zhou, 2018), Catalyst (Lin et al., 2015; Kulunchakov & Mairal, 2019), non-convex variance reduced algorithms (Reddi et al., 2016; Allen-Zhu & Hazan, 2016; Fang et al., 2018; Zhou et al., 2018a) and others. In particular, SGD can be seen as a special case of block coordinate descent, while SAG is a special case of bias-SEGA from (Hanzely et al., 2018) (neither of CD with non-separable prox, nor bias-SEGA were analyzed yet).

Zero order optimization with non-separable non-smooth regularizer. We believe it would be interesting to develop an inexact version of ASVRCD, as this would immediately enable the application in zero-order optimization, where the partial derivatives are (inexactly) estimated using finite differences.

Acknowledgments

The authors would like to express their gratitude to Konstantin Mishchenko. In particular, Konstantin has introduced us to the product space objective (13) and at the same time, the idea that SAGA is a special case of SEGA was born during the discussion with him.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pp. 699–707, 2016.

Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pp. 1110–1119, 2016.

Defazio, A. A simple practical accelerated method for finite sums. In *Advances in neural information processing systems*, pp. 676–684, 2016.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pp. 1646–1654, 2014a.

Defazio, A., Domke, J., et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pp. 1125–1133, 2014b.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018.

Gazagnadou, N., Gower, R. M., and Salmon, J. Optimal mini-batch and step sizes for SAGA. In *International conference on machine learning*, 2019.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5200–5209, 09–15 Jun 2019.

Gutman, D. H. and Pena, J. F. The condition number of a function relative to a set. *arXiv preprint arXiv:1901.08359*, 2019.

Hanzely, F. and Richtárik, P. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Hanzely, F. and Richtárik, P. One method to rule them all: Variance reduction for data, parameters and many new methods. *arXiv preprint arXiv:1905.11266*, 2019.

Hanzely, F., Mishchenko, K., and Richtárik, P. Sega: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pp. 2082–2093, 2018.

Hickernell, F. J., Lemieux, C., Owen, A. B., et al. Control variates for quasi-monte carlo. *Statistical Science*, 20(1): 1–31, 2005.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pp. 2305–2313, 2015.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Konečný, J. and Richtárik, P. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3:9, 2017.

Kovalev, D., Horváth, S., and Richtárik, P. Dont jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

Kozak, D., Becker, S., Doostan, A., and Tenorio, L. Stochastic subspace descent. *arXiv preprint arXiv:1904.01145*, 2019.

Kulunchakov, A. and Mairal, J. A generic acceleration framework for stochastic composite optimization. In *Advances in Neural Information Processing Systems*, pp. 12556–12567, 2019.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1-2): 167–215, 2018.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances in neural information processing systems*, pp. 3384–3392, 2015.

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Nesterov, Y. and Stich, S. U. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2613–2621. JMLR. org, 2017.

Qian, X., Qu, Z., and Richtárik, P. L-svrg and L-Katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019a.

Qian, X., Qu, Z., and Richtárik, P. SAGA with arbitrary sampling. *arXiv preprint arXiv:1901.08669*, 2019b.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pp. 314–323, 2016.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2):1–38, 2014.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400407, 1951.

Roux, N. L., Schmidt, M., and Bach, F. R. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in neural information processing systems*, pp. 2663–2671, 2012.

Shang, F., Jiao, L., Zhou, K., Cheng, J., Ren, Y., and Jin, Y. Asvrg: Accelerated proximal svrg. In *Proceedings of The 10th Asian Conference on Machine Learning*, 2018.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Woodworth, B. E. and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pp. 3639–3647, 2016.

Wright, S. J. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems 31*, pp. 3921–3932. Curran Associates, Inc., 2018a.

Zhou, K. Direct acceleration of saga using sampled negative momentum. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *The 35th International Conference on Machine Learning*, 2018b.

A. Missing lemmas and proofs: SAGA/L-SVRG is a special case of SEGA/SVRCD

A.1. Proof of Lemma 3.2

Let $\mathbf{W}' := \frac{1}{n}\tilde{e}\tilde{e}^\top \otimes \mathbf{I}$ and denote $\mathbf{D_B}(\tilde{\mathbf{M}}) := \text{BlockDiag}(\tilde{\mathbf{M}}_1, \ldots, \tilde{\mathbf{M}}_n)$ for simplicity. Now clearly $x^0 \in \text{Range}\,(\mathbf{W}')$, while $\mathbf{W}'$ is a projection matrix such that $I(x) < \infty$ if and only if $\mathbf{W}'x = x$. Consequently, $\mathbf{W} = \mathbf{W}'$. Next, if $x, y \in \text{Range}\,(\mathbf{W})$, there is $\tilde{x}, \tilde{y} \in \mathbb{R}^{\tilde{d}}$ such that $x = Q\,(\tilde{x})\,, y = Q\,(\tilde{y})$. Therefore we can write

$$\begin{aligned} f(x) = f(\mathbf{W}(x)) = \frac{1}{n}\sum_{j=1}^n \tilde{f}_j(\tilde{x}) &\geq \frac{1}{n}\sum_{j=1}^n \tilde{f}_j(\tilde{y}) + \left\langle \nabla\left(\frac{1}{n}\sum_{j=1}^n \tilde{f}_j(\tilde{y})\right), \tilde{x} - \tilde{y}\right\rangle + \frac{\tilde{\mu}}{2}\|\tilde{x} - \tilde{y}\|^2 \\ &= f(y) + \langle \nabla f(y), x - y\rangle + \frac{\tilde{\mu}}{2n}\|x - y\|^2. \end{aligned}$$

Similarly,

$$\begin{aligned} f(x) = \frac{1}{n}\sum_{j=1}^n \tilde{f}_j(\tilde{x}) &\leq \frac{1}{n}\sum_{j=1}^n \tilde{f}_j(\tilde{y}) + \left\langle \nabla\left(\frac{1}{n}\sum_{j=1}^n \tilde{f}_j(\tilde{y})\right), \tilde{x} - \tilde{y}\right\rangle + \sum_{j=1}^n \frac{1}{2n}\|\tilde{x} - \tilde{y}\|^2_{\tilde{\mathbf{M}}_j} \\ &= f(y) + \langle \nabla f(y), x - y\rangle + \frac{1}{2n}\|x - y\|^2_{\mathbf{D_B}(\tilde{\mathbf{M}})}. \end{aligned}$$

Thus we conclude $\mu = \frac{\tilde{\mu}}{n}$ and $\mathbf{M} = \frac{1}{n}\mathbf{D_B}(\tilde{\mathbf{M}})$. Further, for any $h \in \mathbb{R}^d$, we have:

$$\begin{aligned} &h^\top \mathbf{M}^{\frac{1}{2}} \mathbb{E}\left[\sum_{i\in S} p_i{}^{-1} e_i e_i{}^\top \mathbf{W} \sum_{i\in S} p_i{}^{-1} e_i e_i{}^\top\right] \mathbf{M}^{\frac{1}{2}} h \\ &= \frac{1}{n} h^\top \mathbf{D_B}(\tilde{\mathbf{M}})^{\frac{1}{2}} \mathbb{E}\left[\left(\sum_{i\in \tilde{S}} \tilde{p}_i^{-1}\left(\sum_{j\in R_i} e_j e_j^\top\right)\right) \mathbf{W} \left(\sum_{i\in \tilde{S}} \tilde{p}_i^{-1}\left(\sum_{j\in R_i} e_j e_j^\top\right)\right)\right] \mathbf{D_B}(\tilde{\mathbf{M}})^{\frac{1}{2}} h \\ &= \frac{1}{n}\mathbb{E}\left[\left\|\sum_{i\in \tilde{S}} \tilde{\mathbf{M}}_i^{\frac{1}{2}} \tilde{p}_i^{-1} h_{R_i}\right\|^2\right] \\ &\overset{(12)}{\leq} \frac{1}{n}\sum_{i=1}^n \tilde{p}_i \tilde{v}_i \left\|h_{R_i}\right\|^2 \end{aligned}$$

and thus (6) holds with $v = \frac{1}{n}\tilde{v}$ as desired.

A.2. Proof of Lemma 3.3

Denote $\text{Vec}\,(\cdot)$ to be the vectorization operator, i.e., operator which takes a matrix as an input, and returns a vector constructed by a column-wise stacking of the matrix columns. We will show both

$$h^k = \frac{1}{n}\text{Vec}\left(\mathbf{J}^k\right) \tag{19}$$

and (14) using mathematical induction. Clearly, if $k = 0$ both (19) and (14) hold. Now, let us proceed with the second induction step.

$$
\begin{aligned}
x^{k+1} &= \operatorname{prox}_{\alpha\psi}(x^k - \alpha g^k) = \operatorname*{argmin}_{x\in\mathbb{R}^d} \alpha I(x) + \alpha\tilde{\psi}(x_{R_1}) + \|x - (x^k - \alpha g^k)\|^2 \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^d} \alpha I(x) + \alpha\tilde{\psi}(x_{R_1}) + \left\|x - x^k + \alpha\left(h^k + \sum_{i\in S}\frac{1}{p_i}(\nabla_i f(x^k) - h_i^k)e_i\right)\right\|^2 \\
&= \operatorname*{argmin}_{x=\mathbf{W}x} \alpha\tilde{\psi}(x_{R_1}) + \left\|x - x^k + \alpha\left(h^k + \sum_{i\in S}\frac{1}{p_i}(\nabla_i f(x^k) - h_i^k)e_i\right)\right\|^2 \\
&= \operatorname*{argmin}_{x=\mathbf{W}x} \alpha\tilde{\psi}(x_{R_1}) + \left\|x - x^k + \alpha\left(h^k + \sum_{i\in S}\frac{1}{p_i}(\nabla_i f(x^k) - h_i^k)e_i\right)\right\|_{\mathbf{W}}^2 \\
&\overset{(14)}{=} Q\left(\operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^{\tilde{d}}} \alpha\tilde{\psi}(\tilde{x}) + \left\|Q(\tilde{x}) - Q(\tilde{x}^k) + \alpha\left(h^k + \sum_{i\in\tilde{S}}\frac{1}{\tilde{p}_i}\left(\sum_{j\in R_i}\left(\frac{1}{n}\nabla_j\tilde{f}_i(\tilde{x}^k) - h^k_{(i-1)\tilde{d}+j}\right)e_{(i-1)\tilde{d}+j}\right)\right)\right\|_{\mathbf{W}}^2\right) \\
&= Q\left(\operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^{\tilde{d}}} \alpha\tilde{\psi}(\tilde{x}) + \frac{1}{n}\left\|n\tilde{x} - n\tilde{x}^k + \alpha\left(\sum_{i=1}^n h^k_{R_i} + \sum_{i\in\tilde{S}}\frac{1}{\tilde{p}_i}\left(\frac{1}{n}\nabla\tilde{f}_i(\tilde{x}^k) - h^k_{R_i}\right)\right)\right\|^2\right) \\
&\overset{(19)}{=} Q\left(\operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^{\tilde{d}}} \alpha\tilde{\psi}(\tilde{x}) + \frac{1}{n}\left\|n\tilde{x} - n\tilde{x}^k + \alpha\left(\frac{1}{n}\mathbf{J}^k\tilde{e} + \frac{1}{n}\sum_{i\in\tilde{S}}\frac{1}{\tilde{p}_i}\left((\nabla\tilde{f}_i(\tilde{x}^k) - \mathbf{J}^k_{:,i})\right)\right)\right\|^2\right) \\
&= Q\left(\operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^{\tilde{d}}} \tilde{\alpha}\tilde{\psi}(\tilde{x}) + \left\|\tilde{x} - \tilde{x}^k + \tilde{\alpha}\left(\frac{1}{n}\mathbf{J}^k\tilde{e} + \frac{1}{n}\sum_{i\in\tilde{S}}\frac{1}{\tilde{p}_i}\left((\nabla\tilde{f}_i(\tilde{x}^k) - \mathbf{J}^k_{:,i})\right)\right)\right\|^2\right) \\
&= Q\left(\tilde{x}^{k+1}\right). \qquad (20)
\end{aligned}
$$

It remains to notice that since $x^{k+1} = Q\left(\tilde{x}^k\right)$, we have $h^{k+1} = \frac{1}{n}\text{Vec}\left(\mathbf{J}^{k+1}\right)$ as desired.

B. Missing lemmas and proofs: ASVRCD

B.1. Technical lemmas

We first start with two key technical lemmas.

Lemma B.1 *Suppose that*

$$\eta \leq \frac{1}{2L}. \qquad (21)$$

Then, for all $x \in \text{Range}\,(\mathbf{W})$ the following inequality holds:

$$\frac{1}{\eta}\mathbb{E}\left[\left\langle x - x^k, x^k - y^{k+1}\right\rangle\right] \leq \mathbb{E}\left[P(x) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1} - x^k\right\|^2 + \frac{\eta}{2}\left\|g^k - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] - D_f(x, x^k). \qquad (22)$$

Proof: From the definition of y^{k+1} we get

$$y^{k+1} = x^k - \eta g^k - \eta\Delta,$$

where $\Delta \in \partial\psi(y^{k+1})$. Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{\eta}\left\langle x - x^k, x^k - y^{k+1}\right\rangle\right] &= \mathbb{E}\left[\left\langle x - x^k, g^k + \Delta\right\rangle\right] \\
&= \left\langle x - x^k, \nabla f(x^k)\right\rangle + \mathbb{E}\left[\left\langle x - y^{k+1}, \Delta\right\rangle + \left\langle y^{k+1} - x^k, \Delta\right\rangle\right] \\
&\leq f(x) - f(x^k) - D_f(x, x^k) + \mathbb{E}\left[\psi(x) - \psi(y^{k+1})\right] + \mathbb{E}\left[\left\langle y^{k+1} - x^k, \Delta\right\rangle\right] \qquad (23)
\end{aligned}
$$

Now, we use the fact that f is L-smooth over the set where iterates live (i.e., over $\{x^0 + \mathrm{Range}\,(\mathbf{W})\}$):

$$\begin{aligned} f(y^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), y^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|y^{k+1} - x^k\right\|^2 \\ &= f(x^k) + \left\langle \mathbf{W}\nabla f(x^k), y^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|y^{k+1} - x^k\right\|^2. \end{aligned} \tag{24}$$

Thus, we have

$$\begin{aligned} \mathbb{E}\left[\frac{1}{\eta}\left\langle x - x^k, x^k - y^{k+1}\right\rangle\right] &\overset{(23)+(24)}{\leq} \mathbb{E}\left[P(x) - P(y^{k+1}) + \left\langle y^{k+1} - x^k, \mathbf{W}(\Delta + \nabla f(x^k))\right\rangle + \frac{L}{2}\left\|y^{k+1} - x^k\right\|^2\right] \\ &\quad -D_f(x, x^k) \\ &= \mathbb{E}\left[P(x) - P(y^{k+1}) + \left\langle y^{k+1} - x^k, \mathbf{W}(\nabla f(x^k) - g^k)\right\rangle - \frac{1}{\eta}\left\|y^{k+1} - x^k\right\|^2\right] \\ &\quad +\mathbb{E}\left[\frac{L}{2}\left\|y^{k+1} - x^k\right\|^2\right] - D_f(x, x^k) \\ &\leq \mathbb{E}\left[P(x) - P(y^{k+1}) + \frac{\eta}{2}\left\|\nabla f(x^k) - g^k\right\|_{\mathbf{W}}^2 - \frac{1}{2\eta}\left\|y^{k+1} - x^k\right\|^2 + \frac{L}{2}\left\|y^{k+1} - x^k\right\|\right] \\ &\quad -D_f(x, x^k) \\ &\overset{(21)}{\leq} \mathbb{E}\left[P(x) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1} - x^k\right\|^2 + \frac{\eta}{2}\left\|\nabla f(x^k) - g^k\right\|_{\mathbf{W}}^2\right] - D_f(x, x^k), \end{aligned}$$

which concludes the proof.

Lemma B.2 *Suppose, the following choice of parameters is used:*

$$\eta = \frac{1}{4}\max\{\mathcal{L}', L\}^{-1}, \qquad \gamma = \frac{1}{\max\{2\mu, 4\theta_1/\eta\}}, \qquad \beta = 1 - \gamma\mu, \qquad \theta_2 = \frac{\mathcal{L}'}{2\max\{L, \mathcal{L}'\}}.$$

Then the following inequality holds:

$$\mathbb{E}\left[\left\|z^{k+1} - x^*\right\|^2 + \frac{2\gamma\beta}{\theta_1}\left[P(y^{k+1}) - P(x^*)\right]\right] \leq \beta\left\|z^k - x^*\right\|^2 + \frac{2\gamma\beta\theta_2}{\theta_1}\left[P(w^k) - P(x^*)\right] + \frac{2\gamma\beta(1 - \theta_1 - \theta_2)}{\theta_1}\left[P(y^k) - P(x^*)\right]. \tag{25}$$

Proof:

$$\begin{aligned}
\mathbb{E}\left[\left\|z^{k+1}-x^*\right\|^2\right] &= \mathbb{E}\left[\left\|\beta z^k + (1-\beta)x^k - x^* + \frac{\gamma}{\eta}(y^{k+1}-x^k)\right\|^2\right] \\
&\le \beta\left\|z^k-x^*\right\|^2 + (1-\beta)\left\|x^k-x^*\right\|^2 + \frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\quad + \frac{2\gamma}{\eta}\mathbb{E}\left[\left\langle y^{k+1}-x^k, \beta z^k + (1-\beta)x^k - x^*\right\rangle\right] \\
&= \beta\left\|z^k-x^*\right\|^2 + (1-\beta)\left\|x^k-x^*\right\|^2 + \frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\quad + \frac{2\gamma}{\eta}\mathbb{E}\left[\left\langle y^{k+1}-x^k, x^k-x^*\right\rangle\right] + \frac{2\gamma\beta}{\eta}\mathbb{E}\left[\left\langle y^{k+1}-x^k, z^k-x^k\right\rangle\right] \\
&= \beta\left\|z^k-x^*\right\|^2 + (1-\beta)\left\|x^k-x^*\right\|^2 + \frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] + \frac{2\gamma}{\eta}\mathbb{E}\left[\left\langle x^k-y^{k+1}, x^*-x^k\right\rangle\right] \\
&\quad + \frac{2\gamma\beta\theta_2}{\eta\theta_1}\mathbb{E}\left[\left\langle x^k-y^{k+1}, w^k-x^k\right\rangle\right] + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\eta\theta_1}\mathbb{E}\left[\left\langle x^k-y^{k+1}, y^k-x^k\right\rangle\right] \\
&\overset{(22)}{\le} \beta\left\|z^k-x^*\right\|^2 + (1-\beta)\left\|x^k-x^*\right\|^2 + \frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\quad + 2\gamma\mathbb{E}\left[P(x^*) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1}-x^k\right\|^2 - D_f(x^*,x^k) + \frac{\eta}{2}\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \\
&\quad + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\left[P(w^k) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1}-x^k\right\|^2 - D_f(w^k,x^k) + \frac{\eta}{2}\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \\
&\quad + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\left[P(y^k) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1}-x^k\right\|^2 + \frac{\eta}{2}\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \\
&\overset{(3)}{\le} \beta\left\|z^k-x^*\right\|^2 + (1-\beta-\gamma\mu)\left\|x^k-x^*\right\|^2 + \frac{\gamma^2}{\eta^2}\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] \\
&\quad + 2\gamma\beta\mathbb{E}\left[P(x^*) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1}-x^k\right\|^2\right] + \eta\gamma\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \\
&\quad + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\left[P(w^k) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1}-x^k\right\|^2 - D_f(w^k,x^k) + \frac{\eta}{2}\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \\
&\quad + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\left[P(y^k) - P(y^{k+1}) - \frac{1}{4\eta}\left\|y^{k+1}-x^k\right\|^2 + \frac{\eta}{2}\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right].
\end{aligned}$$

Using $\beta = 1-\gamma\mu$ we get

$$\begin{aligned}
\mathbb{E}\left[\left\|z^{k+1}-x^*\right\|^2\right] &\le \beta\left\|z^k-x^*\right\|^2 + \left[\frac{\gamma^2}{\eta^2} - \frac{\gamma\beta}{2\eta\theta_1}\right]\mathbb{E}\left[\left\|y^{k+1}-x^k\right\|^2\right] + \frac{\eta\gamma}{\theta_1}\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] - \frac{2\gamma\beta\theta_2}{\theta_1}D_f(w^k,x^k) \\
&\quad + 2\gamma\beta\mathbb{E}\left[P(x^*) - P(y^{k+1})\right] + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\left[P(w^k) - P(y^{k+1})\right] + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\left[P(y^k) - P(y^{k+1})\right].
\end{aligned}$$

Using stepsize $\gamma \le \frac{\beta\eta}{2\theta_1}$ we get

$$\begin{aligned}
\mathbb{E}\left[\left\|z^{k+1}-x^*\right\|^2\right] &\le \beta\left\|z^k-x^*\right\|^2 + \frac{\eta\gamma}{\theta_1}\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] - \frac{2\gamma\beta\theta_2}{\theta_1}D_f(w^k,x^k) \\
&\quad + 2\gamma\beta\mathbb{E}\left[P(x^*) - P(y^{k+1})\right] + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\left[P(w^k) - P(y^{k+1})\right] + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\left[P(y^k) - P(y^{k+1})\right].
\end{aligned}$$

Now, using the expected smoothness from inequality (16):

$$\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \le 2\mathcal{L}' D_f(w^k,x^k) \tag{26}$$

and stepsize $\eta \leq \frac{\beta\theta_2}{\mathcal{L}'}$ we get

$$\begin{aligned}
\mathbb{E}\left[\left\|z^{k+1}-x^*\right\|^2\right] &\leq \beta\left\|z^k-x^*\right\|^2 + \frac{2\mathcal{L}'\eta\gamma}{\theta_1}D_f(w^k,x^k) - \frac{2\gamma\beta\theta_2}{\theta_1}D_f(w^k,x^k) \\
&\quad + 2\gamma\beta\mathbb{E}\left[P(x^*)-P(y^{k+1})\right] + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\left[P(w^k)-P(y^{k+1})\right] + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\left[P(y^k)-P(y^{k+1})\right] \\
&\leq \beta\left\|z^k-x^*\right\|^2 + 2\gamma\beta\mathbb{E}\left[P(x^*)-P(y^{k+1})\right] + \frac{2\gamma\beta\theta_2}{\theta_1}\mathbb{E}\left[P(w^k)-P(y^{k+1})\right] \\
&\quad + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\mathbb{E}\left[P(y^k)-P(y^{k+1})\right] \\
&= \beta\left\|z^k-x^*\right\|^2 - \frac{2\gamma\beta}{\theta_1}\mathbb{E}\left[P(y^{k+1})-P(x^*)\right] + \frac{2\gamma\beta\theta_2}{\theta_1}\left[P(w^k)-P(x^*)\right] \\
&\quad + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\left[P(y^k)-P(x^*)\right].
\end{aligned}$$

It remains to rearrange the terms.

B.2. Proof of Theorem 4.1

One can easily show that

$$\mathbb{E}\left[P(w^{k+1})\right] = \rho P(y^k) + (1-\rho)P(w^k). \tag{27}$$

Using that we obtain

$$\begin{aligned}
\mathbb{E}\left[\Psi^{k+1}\right] &\overset{(25)+(27)}{\leq} \beta\left\|z^k-x^*\right\|^2 + \frac{2\gamma\beta\theta_2}{\theta_1}\left[P(w^k)-P(x^*)\right] + \frac{2\gamma\beta(1-\theta_1-\theta_2)}{\theta_1}\left[P(y^k)-P(x^*)\right] \\
&\quad + \frac{(2\theta_2+\theta_1)\gamma\beta}{\theta_1\rho}\left[\rho P(y^k) + (1-\rho)P(w^k) - P(x^*)\right] \\
&= \beta\left\|z^k-x^*\right\|^2 + \frac{2\gamma\beta(1-\theta_1/2)}{\theta_1}\left[P(y^k)-P(x^*)\right] + \frac{(2\theta_2+\theta_1)\gamma\beta}{\theta_1\rho}\left[1-\rho+\frac{2\rho\theta_2}{2\theta_2+\theta_1}\right]\left[P(w^k)-P(x^*)\right] \\
&\leq \max\left\{1-\frac{1}{\max\{2,4\theta_1/(\eta\mu)\}}, 1-\frac{\theta_1}{2}, 1-\frac{\rho\theta_1}{2\max\{2\theta_2,\theta_1\}}\right\}\Psi^k \\
&= \left[1-\max\left\{\frac{2}{\rho}, \frac{4}{\theta_1}\max\left\{\frac{1}{2},\frac{\theta_2}{\rho}\right\}, \frac{4\theta_1}{\eta\mu}\right\}^{-1}\right]\Psi^k.
\end{aligned}$$

Using $\theta_1 = \min\left\{\frac{1}{2}, \sqrt{\eta\mu\max\left\{\frac{1}{2},\frac{\theta_2}{\rho}\right\}}\right\}$ we get

$$\begin{aligned}
\mathbb{E}\left[\Psi^{k+1}\right] &\leq \left[1-\max\left\{\frac{2}{\rho}, 8\max\left\{\frac{1}{2},\frac{\theta_2}{\rho}\right\}, 4\sqrt{\frac{\max\left\{\frac{1}{2},\frac{\theta_2}{\rho}\right\}}{\eta\mu}}\right\}^{-1}\right]\Psi^k \\
&\leq \left[1-\frac{1}{4}\max\left\{\frac{1}{\rho}, \sqrt{\frac{2\max\left\{L,\frac{\mathcal{L}'}{\rho}\right\}}{\mu}}\right\}^{-1}\right]\Psi^k,
\end{aligned}$$

as desired.

B.3. Proof of Lemma 4.2

To establish that that we can choose $\mathcal{L}' = \mathcal{L}$, it suffices to see

$$\begin{aligned}
\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] &= \mathbb{E}\left[\left\|\sum_{i\in S}\frac{1}{p_i}(\nabla_i f(x^k) - \nabla_i f(w^k))e_i + \nabla f(w^k) - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] \\
&\leq \mathbb{E}\left[\left\|\sum_{i\in S}\frac{1}{p_i}(\nabla_i f(x^k) - \nabla_i f(w^k))e_i\right\|_{\mathbf{W}}^2\right] \\
&\overset{(7)}{\leq} \mathcal{L}\left\|\nabla f(x^k) - \nabla f(w^k)\right\|_{\mathbf{M}^{-1}}^2 \\
&\overset{(35)}{\leq} 2\mathcal{L}D_f(w^k, x^k).
\end{aligned}$$

Next, to establish $\mathcal{L} \geq L$, let $\mathbf{Q} \coloneqq \sum_{i\in S}\frac{1}{p_i}e_i{e_i}^\top\mathbf{W}$. Consequently, we get

$$\begin{aligned}
\mathcal{L} &\overset{(7)}{=} \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbb{E}\left[\sum_{i\in S}\frac{1}{p_i}e_i{e_i}^\top\mathbf{W}\sum_{i\in S}\frac{1}{p_i}e_i{e_i}^\top\right]\mathbf{M}^{\frac{1}{2}}\right) \\
&= \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbb{E}\left[\sum_{i\in S}\frac{1}{p_i}e_i{e_i}^\top\mathbf{W}^2\sum_{i\in S}\frac{1}{p_i}e_i{e_i}^\top\right]\mathbf{M}^{\frac{1}{2}}\right) \\
&= \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbb{E}\left[\mathbf{Q}\mathbf{Q}^\top\right]\mathbf{M}^{\frac{1}{2}}\right) \\
&\geq \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbb{E}\left[\mathbf{Q}\right]\mathbb{E}\left[\mathbf{Q}^\top\right]\mathbf{M}^{\frac{1}{2}}\right) \\
&= \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{W}^2\mathbf{M}^{\frac{1}{2}}\right) \\
&= \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{W}\mathbf{M}^{\frac{1}{2}}\right) \\
&= L,
\end{aligned}$$

as desired.

B.4. Proof of Lemma 4.3

Let us look first at $\mathbf{W} = \mathbf{I}$. In such case, it is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] &= \mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|d(\nabla_i f(x^k) - \nabla_i f(w^k))e_i\right\|^2\right] \\
&= d\left\|\nabla f(x^k) - \nabla f(w^k)\right\|^2 \\
&\leq 2d\lambda_{\max}\mathbf{M}D_f(w^k, x^k),
\end{aligned}$$

i. e. we can choose $\mathcal{L}' = d\lambda_{\max}\mathbf{M}$. Noting that $\lambda_{\max}\mathbf{M} \geq L$, the iteration complexity of Algorithm 3 is $\mathcal{O}\left(d\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$.

On the other hand, if $\mathbf{W} = \frac{1}{d}\tilde{e}^\top$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] &= \mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_{\frac{1}{d}ee^\top}^2\right] \\
&\leq \mathbb{E}\left[\left\|d(\nabla_i f(x^k) - \nabla_i f(w^k))e_i\right\|_{\frac{1}{d}ee^\top}^2\right] \\
&= \left\|\nabla f(x^k) - \nabla f(w^k)\right\|^2 \\
&\leq 2\lambda_{\max}\mathbf{M}D_f(w^k, x^k),
\end{aligned}$$

and therefore $\mathcal{L}' = \lambda_{\max}\mathbf{M}$, which yields $\mathcal{O}\left(\sqrt{\frac{d\lambda_{\max}\mathbf{M}}{\mu}}\log\frac{1}{\epsilon}\right)$ convergence rate.

C. Missing lemmas and proofs: L-Katyusha as a particular case of ASVRCD

C.1. Proof of Lemma 5.2

Let us proceed by induction. We will show the following for all $k \geq 0$ we have

$$\tilde{x}^k = x_{R_1}^k = \cdots = x_{R_n}^k,\ \tilde{y}^k = y_{R_1}^k = \cdots = y_{R_n}^k,\ \tilde{z}^k = z_{R_1}^k = \cdots = z_{R_n}^k \text{ and } \tilde{w}^k = w_{R_1}^k = \cdots = w_{R_n}^k. \tag{28}$$

Clearly, for $k = 0$, the above claim holds. Let us proceed with the second induction step and assume that (28) holds for some $k \geq 0$. First, the update rule on $\{x^k\}$ for ASVRCD together with the update rule on $\{\tilde{x}^k\}$ yields

$$\tilde{x}^{k+1} = x_{R_1}^{k+1} = \cdots = x_{R_n}^{k+1}. \tag{29}$$

To show

$$\tilde{y}^{k+1} = y_{R_1}^{k+1} = \cdots = y_{R_n}^{k+1}, \tag{30}$$

we essentially repeat the proof of Lemma 3.3. In particular, it is sufficient to repeat the sequence of inequalities (20) where variables $(x^{k+1}, \tilde{x}^{k+1}, h^k, \mathbf{J}^k\alpha, \tilde{\alpha})$ are replaced by $(y^{k+1}, \tilde{y}^{k+1}, \nabla f(w^k), [\nabla\tilde{f}_1(\tilde{w}^k), \ldots, \nabla\tilde{f}_n(\tilde{w}^k)], \eta, \tilde{\eta})$ respectively.

Next, $\tilde{z}^{k+1} = z_{R_1}^{k+1} = \cdots = z_{R_n}^{k+1}$ follows from (28), (29) and (30) together with the update rule (on $\{z^k\}$ and $\{\tilde{z}^k\}$) of both algorithms and the fact that $\frac{\gamma}{\eta} = \frac{\tilde{\gamma}}{\tilde{\eta}}$.

To finish the proof of the algorithms equivalence, we shall notice that $\tilde{w}^{k+1} = w_{R_1}^{k+1} = \cdots = w_{R_n}^{k+1}$ follows from (28), (30) together with the update rule (on $\{w^k\}$ and $\{\tilde{w}^k\}$) of both algorithms.

To show $\mathcal{L}' = \frac{\tilde{\mathcal{L}}}{n}$ it is sufficient to see

$$\begin{aligned}
\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_{\mathbf{W}}^2\right] &= \mathbb{E}\left[\left\|\sum_{i\in\tilde{S}} p_i^{-1}\left(\sum_{j\in R_i}\left(\nabla_j f(x^k) - \nabla_j f(w^k)\right)e_j\right) - \left(\nabla f(x^k) - \nabla f(w^k)\right)\right\|_{\mathbf{W}}^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{W}\left(\sum_{i\in\tilde{S}} p_i^{-1}\left(\sum_{j\in R_i}\left(\nabla_j f(x^k) - \nabla_j f(w^k)\right)e_j\right)\right) - \mathbf{W}\left(\nabla f(x^k) - \nabla f(w^k)\right)\right\|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left\|\left(\frac{1}{n}\sum_{i\in\tilde{S}}\tilde{p}_i^{-1}\left(\nabla\tilde{f}_i(\tilde{x}^k) - \nabla\tilde{f}_i(\tilde{w}^k)\right)\right) - \left(\nabla\tilde{f}(\tilde{x}^k) - \nabla\tilde{f}(\tilde{w}^k)\right)\right\|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\left\|\tilde{g}^k - \nabla\tilde{f}(\tilde{w}^k)\right\|^2\right] \\
&\leq 2\frac{\tilde{\mathcal{L}}}{n}D_{\tilde{f}}(\tilde{w}^k, \tilde{x}^k) \\
&= 2\frac{\tilde{\mathcal{L}}}{n}\left(\frac{1}{n}\sum_{i=1}^n D_{\tilde{f}_i}(w_{R_i}^k, x_{R_i}^k)\right) \\
&= 2\frac{\tilde{\mathcal{L}}}{n}D_f(w^k, x^k).
\end{aligned}$$

Lastly, if $x, y \in \text{Range}(\mathbf{W})$, there is $\tilde{x}, \tilde{y} \in \mathbb{R}^{\tilde{d}}$ such that $x = Q(\tilde{x}), y = Q(\tilde{y})$. Therefore we can write

$$\begin{aligned} f(x) = f(\mathbf{W}(x)) = \frac{1}{n}\sum_{j=1}^{n} \tilde{f}_j(\tilde{x}) &\leq \frac{1}{n}\sum_{j=1}^{n} \tilde{f}_j(\tilde{y}) + \left\langle \nabla\left(\frac{1}{n}\sum_{j=1}^{n} \tilde{f}_j(\tilde{y})\right), \tilde{x} - \tilde{y}\right\rangle + \frac{\tilde{L}}{2}\|\tilde{x} - \tilde{y}\|^2 \\ &= f(y) + \langle \nabla f(y), x - y\rangle + \frac{\tilde{L}}{2n}\|x - y\|^2 \end{aligned}$$

and thus $L = \lambda_{\max}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{W}\mathbf{M}^{\frac{1}{2}}\right) \leq n^{-1}\tilde{L}$.

D. Tighter rates for GJS (Hanzely & Richtárik, 2019) by exploiting prox and proof of Theorem 2.2

In this section, we show that specific nonsmooth function ψ might lead to faster convergence of variance reduced methods. We exploit the well-known fact that under some circumstances, a proximal operator might change the smoothness structure of the objective (Gutman & Pena, 2019). In particular, we consider Generalized Jacobian Sketching (GJS) from (Hanzely & Richtárik, 2019). We generalize Theorem 5.1 therein, which allows for a tighter rate if ψ has a specific structure.

D.1. GJS

Consider a the following objective:

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} f_i(x) + \psi(x).$$

and define Jacobian operator $\mathbf{G}: \mathbb{R}^d \to \mathbb{R}^{d\times n}$ as $\mathbf{G}(x) \coloneqq [\nabla f_1(x), \ldots, \nabla f_n(x)]$. Further, define $\mathcal{M}: \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ to be such linear operator that the following holds $(\mathcal{M}\mathbf{X})_{:j} = \mathbf{M}_j\mathbf{X}_{:j}$ for $j \in [n]$.

Suppose that $\mathcal{U}: \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ is a random linear operator such that $\mathbb{E}[\mathcal{U}]$ is identity, and $\mathcal{S}: \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ is a random projection operator. Given the (fixed) distribution over $\mathcal{U}, \mathcal{S}$, GJS is a variance reduced algorithm with the oracle access to $\mathcal{U}\mathbf{G}(x), \mathcal{S}\mathbf{G}(x)$.

Algorithm 5 Generalized JacSketch (GJS)

Parameters: Stepsize $\alpha > 0$, random projector $\mathcal{S}$ and unbiased sketch $\mathcal{U}$
Initialization: Choose solution estimate $x^0 \in \mathbb{R}^d$ and Jacobian estimate $\mathbf{J}^0 \in \mathbb{R}^{d\times n}$
for $k = 0, 1, \ldots$ **do**
  Sample realizations of $\mathcal{S}$ and $\mathcal{U}$, and perform sketches $\mathcal{S}\mathbf{G}(x^k)$ and $\mathcal{U}\mathbf{G}(x^k)$
  $\mathbf{J}^{k+1} = \mathbf{J}^k - \mathcal{S}(\mathbf{J}^k - \mathbf{G}(x^k))$ — update the Jacobian estimate
  $g^k = \frac{1}{n}\mathbf{J}^k e + \frac{1}{n}\mathcal{U}\left(\mathbf{G}(x^k) - \mathbf{J}^k\right)e$ — construct the gradient estimator
  $x^{k+1} = \text{prox}_{\alpha\varphi}(x^k - \alpha g^k)$ — perform the proximal SGD step
end for

Theorem D.1 (Extension of Theorem 5.1 from (Hanzely & Richtárik, 2019)) *Define $f(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. Let Assumption 1.1 hold and suppose that $\mathcal{M}^{\dagger\frac{1}{2}}$ commutes with $\mathcal{S}$. Next, let α and $\mathcal{B}$ are such that for every $\mathbf{X} \in \mathbb{R}^{d\times n}$ we have*

$$\frac{2\alpha}{n^2}\mathbb{E}\left[\|\mathcal{U}\mathbf{X}e\|^2_{\mathbf{W}}\right] + \left\|\left(\mathcal{I} - \mathbb{E}[\mathcal{S}]\right)^{\frac{1}{2}}\mathcal{B}\mathcal{M}^{\dagger}\mathbf{X}\right\|^2 \leq (1 - \alpha\sigma')\left\|\mathcal{B}\mathcal{M}^{\dagger}\mathbf{X}\right\|^2, \tag{31}$$

$$\frac{2\alpha}{n^2}\mathbb{E}\left[\|\mathcal{U}\mathbf{X}e\|^2_{\mathbf{W}}\right] + \left\|\left(\mathbb{E}[\mathcal{S}]\right)^{\frac{1}{2}}\mathcal{B}\mathcal{M}^{\dagger}\mathbf{X}\right\|^2 \leq \frac{1}{n}\left\|\mathcal{M}^{\dagger}\mathbf{X}\right\|^2 \tag{32}$$

and $\mathcal{B}$ commutes with $\mathcal{S}$. Then for all $k \geq 0$, we have $\mathbb{E}\left[\Psi^k\right] \leq (1 - \alpha\sigma')^k\Psi^0$, where

$$\Psi^k \coloneqq \left\|x^k - x^*\right\|^2 + \alpha\left\|\mathcal{B}\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)\right\|^2.$$

D.2. Towards the proof of Theorem D.1

Lemma D.2 *(Slight extension of Lemma from (Hanzely & Richtárik, 2019)) Let $\mathcal{U}$ be random linear operator which is identity in expectation. Let $\mathbf{G}(x)$ be Jacobian at x and $g^k = \frac{1}{n}\mathcal{U}(\mathbf{G}(x) - \mathbf{J}^k)e - \frac{1}{n}\mathbf{J}^k e$. Then for any $\mathbf{Q} \in \mathbb{R}^{d\times d}, \mathbf{Q} \succeq 0$ and all $k \geq 0$ we have*

$$\mathbb{E}\left[\left\|g^k - \nabla f(x^*)\right\|_{\mathbf{Q}}^2\right] \leq \frac{2}{n^2}\mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right] + \frac{2}{n^2}\mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right]. \tag{33}$$

Proof: Since $\nabla f(x^*) = \frac{1}{n}\mathbf{G}(x^*)e$, we have

$$g^k - \nabla f(x^*) = \underbrace{\frac{1}{n}\mathcal{U}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)e}_{a} + \underbrace{\frac{1}{n}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e - \frac{1}{n}\mathcal{U}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e}_{b}. \tag{34}$$

Applying the bound $\|a+b\|_{\mathbf{Q}}^2 \leq 2\|a\|_{\mathbf{Q}}^2 + 2\|b\|_{\mathbf{Q}}^2$ to (34) and taking expectations, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|g^k - \nabla f(x^*)\right\|_{\mathbf{Q}}^2\right] &\leq \mathbb{E}\left[\frac{2}{n^2}\left\|\mathcal{U}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right] \\
&\quad + \mathbb{E}\left[\frac{2}{n^2}\left\|\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e - \mathcal{U}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right] \\
&= \frac{2}{n^2}\mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right] \\
&\quad + \frac{2}{n^2}\mathbb{E}\left[\left\|(\mathcal{I} - \mathcal{U})\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right].
\end{aligned}$$

It remains to note that

$$\begin{aligned}
\mathbb{E}\left[\left\|(\mathcal{I} - \mathcal{U})(\mathbf{J}^k - \mathbf{G}(x^*))\,e\right\|_{\mathbf{Q}}^2\right] &= \mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right] - \left\|\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2 \\
&\leq \mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{Q}}^2\right].
\end{aligned}$$

Lemma D.3 *((Hanzely & Richtárik, 2019), Lemma E.3) Assume that function f_j are convex and $\mathbf{M}_j$-smooth. Then*

$$D_{f_j}(x,y) \geq \frac{1}{2}\left\|\nabla f_j(x) - \nabla f_j(y)\right\|_{\mathbf{M}_j^\dagger}^2, \quad \forall x, y \in \mathbb{R}^d, 1 \leq j \leq n. \tag{35}$$

If $x - y \in \mathrm{Null}\left(\mathbf{M}_j\right)$, then

(i)
$$f_j(x) = f_j(y) + \langle \nabla f_j(y), x - y\rangle, \tag{36}$$

(ii)
$$\nabla f_j(x) - \nabla f_j(y) \in \mathrm{Null}\left(\mathbf{M}_j\right), \tag{37}$$

(iii)
$$\langle \nabla f_j(x) - \nabla f_j(y), x - y\rangle = 0. \tag{38}$$

If, in addition, f_j is bounded below, then $\nabla f_j(x) \in \mathrm{Range}\left(\mathbf{M}_j\right)$ for all x.

Lemma D.4 *((Hanzely & Richtárik, 2019), Lemma E.5) Let $\mathcal{S}$ be a random projection operator and $\mathcal{A}$ any deterministic linear operator commuting with $\mathcal{S}$, i.e., $\mathcal{A}\mathcal{S} = \mathcal{S}\mathcal{A}$. Further, let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d\times n}$ and define $\mathbf{Z} = (\mathcal{I} - \mathcal{S})\mathbf{X} + \mathcal{S}\mathbf{Y}$. Then*

(i) $\mathcal{A}\mathbf{Z} = (\mathcal{I} - \mathcal{S})\mathcal{A}\mathbf{X} + \mathcal{S}\mathcal{A}\mathbf{Y}$,

(ii) $\|\mathcal{A}\mathbf{Z}\|^2 = \|(\mathcal{I} - \mathcal{S})\mathcal{A}\mathbf{X}\|^2 + \|\mathcal{S}\mathcal{A}\mathbf{Y}\|^2$,

(iii) $\mathbb{E}\left[\|\mathcal{A}\mathbf{Z}\|^2\right] = \left\|(\mathcal{I} - \mathbb{E}\left[\mathcal{S}\right])^{1/2}\mathcal{A}\mathbf{X}\right\|^2 + \left\|\mathbb{E}\left[\mathcal{S}\right]^{1/2}\mathcal{A}\mathbf{Y}\right\|^2$, *where the expectation is with respect to $\mathcal{S}$.*

Proof of Theorem D.1 For simplicity of notation, in this proof, all expectations are conditional on x^k, i.e., the expectation is taken with respect to the randomness of g^k. First notice that

$$\mathbb{E}\left[g^k\right] = \nabla f(x^k). \tag{39}$$

For any differentiable function h let $D_h(x,y)$ to be Bregman distance with kernel h, i.e., $D_h(x,y) \coloneqq h(x) - h(y) - \langle \nabla h(y), x-y\rangle$. Since

$$x^* = \operatorname{prox}_{\alpha\varphi}(x^* - \alpha\nabla f(x^*)), \tag{40}$$

and since the prox operator is non-expansive, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] &\overset{(40)}{=} \mathbb{E}\left[\left\|\operatorname{prox}_{\alpha\varphi}(x^k-\alpha g^k) - \operatorname{prox}_{\alpha\varphi}(x^*-\alpha\nabla f(x^*))\right\|^2\right] \\
&\overset{(5)+(4)}{\leq} \mathbb{E}\left[\left\|x^k - x^* - \alpha\mathbf{W}(g^k - \nabla f(x^*))\right\|^2\right] \\
&\overset{(39)}{=} \left\|x^k-x^*\right\|^2 - 2\alpha\left\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\right\rangle \\
&\qquad +\alpha^2\mathbb{E}\left[\left\|g^k - \nabla f(x^*)\right\|_{\mathbf{W}}^2\right] \\
&\leq (1-\alpha\sigma')\left\|x^k-x^*\right\|^2 + \alpha^2\mathbb{E}\left[\left\|g^k-\nabla f(x^*)\right\|_{\mathbf{W}}^2\right] \\
&\qquad -2\alpha D_f(x^k,x^*).
\end{aligned} \tag{41}$$

Since $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, in view of (35) we have

$$\begin{aligned}
D_f(x^k,x^*) &= \frac{1}{n}\sum_{i=1}^n D_{f_i}(x^k,x^*) \overset{(35)}{\geq} \frac{1}{2n}\sum_{i=1}^n \left\|\nabla f_i(x^k) - \nabla f_i(x^*)\right\|_{\mathbf{M}_i^\dagger}^2 \\
&= \frac{1}{2n}\left\|\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)\right\|^2.
\end{aligned} \tag{42}$$

By combining (41) and (42), we get

$$\begin{aligned}
\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] &\leq (1-\alpha\sigma')\left\|x^k-x^*\right\|^2 + \alpha^2\mathbb{E}\left[\left\|g^k-\nabla f(x^*)\right\|_{\mathbf{W}}^2\right] \\
&\qquad -\frac{\alpha}{n}\left\|\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{G}(x^k) - \mathbf{G}(x^*))\right\|^2.
\end{aligned}$$

Next, applying Lemma D.2 with $\mathbf{Q} = \mathbf{W}$ leads to the estimate

$$\begin{aligned}
\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] &\leq (1-\alpha\sigma')\left\|x^k-x^*\right\|^2 - \frac{\alpha}{n}\left\|\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)\right\|^2 \\
&\qquad +\frac{2\alpha^2}{n^2}\mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{W}}^2\right] \\
&\qquad +\frac{2\alpha^2}{n^2}\mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)e\right\|_{\mathbf{W}}^2\right].
\end{aligned} \tag{43}$$

Since, by assumption, both $\mathcal{B}$ and $\mathcal{M}^{\dagger\frac{1}{2}}$ commute with $\mathcal{S}$, so does their composition $\mathcal{A} \coloneqq \mathcal{B}\mathcal{M}^{\dagger\frac{1}{2}}$. Applying Lemma D.4, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathcal{B}\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{J}^{k+1} - \mathbf{G}(x^*)\right)\right\|^2\right] &= \left\|(\mathcal{I} - \mathbb{E}\left[\mathcal{S}\right])^{\frac{1}{2}}\mathcal{B}\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{J}^k - \mathbf{G}(x^*)\right)\right\|^2 \\
&\qquad +\left\|\mathbb{E}\left[\mathcal{S}\right]^{\frac{1}{2}}\mathcal{B}\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{G}(x^k) - \mathbf{G}(x^*)\right)\right\|^2.
\end{aligned} \tag{44}$$

Adding α-multiple of (44) for $\mathcal{C} = \mathcal{M}^{\dagger\frac{1}{2}}$ to (43) yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] + \alpha\mathbb{E}\left[\left\|\mathcal{B}\left(\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{J}^{k+1}-\mathbf{G}(x^*))\right)\right\|^2\right] \\
&\quad\le (1-\alpha\sigma')\left\|x^k-x^*\right\|^2 + \frac{2\alpha^2}{n^2}\mathbb{E}\left[\left\|\mathcal{U}\left(\mathbf{G}(x^k)-\mathbf{G}(x^*)\right)e\right\|_{\mathbf{W}}^2\right] + \\
&\qquad\frac{2\alpha^2}{n^2}\mathbb{E}\left[\|\mathcal{U}(\mathbf{J}^k-\mathbf{G}(x^*))e\|_{\mathbf{W}}^2\right] + \alpha\left\|(\mathcal{I}-\mathbb{E}\left[\mathcal{S}\right])^{\frac{1}{2}}\left(\mathcal{B}\left(\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{J}^k-\mathbf{G}(x^*))\right)\right)\right\|^2 \\
&\qquad+\alpha\left\|\mathbb{E}\left[\mathcal{S}\right]^{\frac{1}{2}}\left(\mathcal{B}\left(\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{G}(x^k)-\mathbf{G}(x^*))\right)\right)\right\|^2 - \frac{\alpha}{n}\left\|\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{G}(x^k)-\mathbf{G}(x^*))\right\|^2 \\
&\quad\overset{(31)}{\le} (1-\alpha\sigma')\left\|x^k-x^*\right\|^2 + (1-\alpha\sigma')\alpha\mathbb{E}\left[\left\|\mathcal{B}\left(\mathcal{M}^{\dagger\frac{1}{2}}\left(\mathbf{J}^k-\mathbf{G}(x^*)\right)\right)\right\|^2\right] \\
&\qquad+\frac{2\alpha^2}{n^2}\mathbb{E}\left[\|\mathcal{U}(\mathbf{G}(x^k)-\mathbf{G}(x^*))e\|_{\mathbf{W}}^2\right] + \alpha\left\|\mathbb{E}\left[\mathcal{S}\right]^{\frac{1}{2}}\left(\mathcal{B}\left(\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{G}(x^k)-\mathbf{G}(x^*))\right)\right)\right\|^2 \\
&\qquad-\frac{\alpha}{n}\left\|\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{G}(x^k)-\mathbf{G}(x^*))\right\|^2 \\
&\quad\overset{(32)}{\le} (1-\alpha\sigma')\left(\left\|x^k-x^*\right\|^2 + \alpha\mathbb{E}\left[\left\|\mathcal{B}\left(\mathcal{M}^{\dagger\frac{1}{2}}(\mathbf{J}^k-\mathbf{G}(x^*))\right)\right\|^2\right]\right).
\end{aligned}$$

Above, we have used (31) with $\mathbf{X} = \mathbf{J}^k - \mathbf{G}(x^*)$ and (32) with $\mathbf{X} = \mathbf{G}(x^k) - \mathbf{G}(x^*)$.

D.3. Proof of Theorem 2.2

First, due to our choice of $\mathcal{S}$ we have

$$\mathbb{E}\left[\mathcal{S}(x)\right] \quad = \quad \mathbf{D}(p)x$$

and at the same time $\mathcal{S}$ and $\mathcal{M}^{\dagger\frac{1}{2}}$ commute. Next, (6) implies

$$\mathbb{E}\left[\left\|\mathcal{U}(\mathbf{M}^{\frac{1}{2}}x)\right\|_{\mathbf{W}}^2\right] = \|x\|^2_{\mathbf{M}^{\frac{1}{2}}\mathbb{E}[\sum_{i\in S}p_i{}^{-1}e_ie_i{}^\top\mathbf{W}\sum_{i\in S}p_i{}^{-1}e_ie_i{}^\top]\mathbf{M}^{\frac{1}{2}}} \le \|x\|^2_{p^{-1}\circ w}.$$

In order to satisfy (31) and (32) it remains to have (we substituted $y = \mathbf{M}^{\dagger\frac{1}{2}}x$):

$$2\alpha\|y\|^2_{p^{-1}\circ w} + \left\|(\mathcal{I}-\mathbb{E}\left[\mathcal{S}\right])^{\frac{1}{2}}\mathcal{B}(y)\right\|^2 \le (1-\alpha\sigma)\|\mathcal{B}(y)\|^2, \tag{45}$$

$$2\alpha\|y\|^2_{p^{-1}\circ w} + \left\|(\mathbb{E}\left[\mathcal{S}\right])^{\frac{1}{2}}\mathcal{B}(y)\right\|^2 \le \|y\|^2. \tag{46}$$

Let us consider $\mathcal{B}$ to be the operator corresponding to the left multiplication with matrix $\mathbf{D}(b)$. Thus for satisfy (31) it suffices to have for all $i \in [d]$:

$$2\alpha m_i p_i{}^{-1} + b_i^2(1-p_i) \le b_i^2(1-\alpha\sigma) \qquad \Rightarrow \qquad 2\alpha m_i p_i{}^{-1} + b_i^2\alpha\sigma \le b_i^2 p_i.$$

For (32) it suffices to have for all $i \in [d]$

$$2\alpha m_i p_i{}^{-1} + b_i^2 p_i \le 1.$$

It remains to notice that choice $b_i^2 = \frac{1}{2p_i}$ and $\alpha = \min_i \frac{p_i}{4m_i+\sigma}$ is valid.