Proximal Splitting Meets Variance Reduction

Fabian Pedregosa
ETH Zürich and UC Berkeley[1]
USA and Switzerland

Kilian Fatras
Univ. Bretagne-Sud, CNRS, IRISA
INRIA Rennes and OBELIX

Mattia Casotto
Akur8
France

Abstract

Despite the raise to fame of stochastic variance reduced methods like SAGA and ProxSVRG, their use in non-smooth optimization is still limited to a few simple cases. Existing methods require to compute the proximal operator of the non-smooth term at each iteration, which, for complex penalties like the total variation, overlapping group lasso or trend filtering, is an iterative process that becomes unfeasible for moderately large problems. In this work we propose and analyze VR-TOS, a variance-reduced method to solve problems with an arbitrary number of non-smooth terms. Like other variance reduced methods, it only requires to evaluate one gradient per iteration and converges with a constant step size, and so is ideally suited for large scale applications. Unlike existing variance reduced methods, it admits multiple non-smooth terms whose proximal operator only needs to be evaluated once per iteration. We provide a convergence rate analysis for the proposed methods that achieves the same asymptotic rate as their full gradient variants and illustrate its computational advantage on 4 different large scale datasets.

1 Introduction

Stochastic variance reduced methods (Le Roux et al., 2012; Johnson and Zhang, 2013; Shalev-Shwartz and Zhang, 2013) have been recently proposed as an improved alternative to the venerable stochastic gradient descent (SGD) method (Robbins and Monro, 1951). As SGD, these methods only require to visit a small batch of random examples per iteration. This makes them ideally suited for large scale machine learning problems. Unlike SGD, the variance of the updates decreases to zero –hence the name– and converge with non-decreasing step sizes.

While initial stochastic variance reduced methods only considered smooth objectives, variants with support for a non-smooth term like ProxSVRG (Xiao and Zhang, 2014) and SAGA (Defazio et al., 2014) were soon developed. These methods are highly efficient whenever the nonsmooth part is *proximal*, that is, its proximal operator is available in closed form or at least fast to compute. This includes penalties such as the ℓ_1 or group lasso norm, but not more complex ones like the overlapping group lasso (Jacob et al., 2009), multidimensional total variation (Barbero and Sra, 2014) or trend filtering (Kim et al., 2009), to name a few.

A key observation is that many of these complex penalties can be decomposed as a sum of proximal terms. Proximal splitting methods like the three operator splitting (Davis and Yin, 2017) or the Condat-Vũ algorithm (Condat, 2013b; Vũ, 2013) then provide a principled approach to incorporate these penalties into the optimization. However, these methods require to compute the full gradient of the smooth term at each iteration, which can become costly in the context of large scale machine learning problems as it involves a full pass over the dataset. A question of key practical interest is whether these proximal splitting methods can be accelerated through the use of stochastic variance reduction techniques.

Our **main contribution** is the development and analysis of VR-TOS, a stochastic variance reduced method that can solve problems with a sum of proximal terms.

The proposed method bridges two previously distant families of algorithms and inherit the best of both: like the three operator splitting of Davis and Yin (2017), it can solve problems with multiple proximal terms, and like variance reduced stochastic methods its cost is independent on the number of smooth terms, converges with a fixed step size, and reaches the same asymptotic

[1]Currently at Google AI, Canada

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

convergence rate than full gradient methods. Furthermore, we also develop a sparse variant of the proposed algorithm which can take advantage of the sparsity in the input data. The paper is organized as follows:

- *Method.* §2 describes the VR-TOS algorithm, and extends it in §2.1 to leverage sparsity in the input data. §2.2 extends these methods to the case of an arbitrary number of proximal terms.
- *Analysis.* In §4 we provide a non-asymptotic convergence analysis of the proposed method. We show that, like other variance reduced methods, it converges with a fixed step size and can achieve the same asymptotic rate as the full gradient variants.
- *Experiments.* In §5 we compare the proposed method and related algorithms on a logistic regression problem with overlapping group lasso penalty on 4 datasets.

1.1 Definitions and notation

By convention, we denote vectors and vector-valued functions in lowercase boldface (e.g. $\boldsymbol{x}$) and matrices in uppercase boldface letters (e.g. $\boldsymbol{D}$). The proximal operator of a convex lower semicontinuous function h is defined as $\mathbf{prox}_{\gamma h}(\boldsymbol{x}) \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{z}\in\mathbb{R}^p}\{h(\boldsymbol{z}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{z}\|^2\}$. We say a function f is L-smooth if it is differentiable and its gradient is L-Lipschitz, while it is μ-strongly convex if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

We denote by $[\,\boldsymbol{x}\,]_b$ the b-th coordinate in $\boldsymbol{x}$. This notation is overloaded so that for a collection of blocks $T = \{B_1, B_2, \ldots\}$, $[\boldsymbol{x}]_T$ denotes the vector $\boldsymbol{x}$ restricted to the coordinates in the blocks of T. For convenience, when T consists of a single block B we use $[\boldsymbol{x}]_B$ as a shortcut of $[\boldsymbol{x}]_{\{B\}}$. Finally, we distinguish $\mathbb{E}$, the full expectation taken with respect to all the randomness in the system, from $\mathbf{E}$, the conditional expectation with respect to the random index sampled at iteration t, conditioned on all randomness up to iteration t.

2 Methods

In this section we present our main contribution, the variance reduced three operator splitting method. We will first consider problems with only two non-smooth terms, and generalize this formulation to an arbitrary number in §2.2.

We consider the following optimization problem

$$\begin{aligned}&\underset{\boldsymbol{x}\in\mathbb{R}^p}{\text{minimize}}\ f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x})\,,\\ &\text{with } f(\boldsymbol{x}) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \psi_i(\boldsymbol{x}) + \omega(\boldsymbol{x})\end{aligned} \tag{OPT}$$

Algorithm 1: Variance Reduced TOS (VR-TOS)

Input: $\boldsymbol{y}_0 \in \mathbb{R}^p$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^{n\times p}$, $\gamma > 0$
Temporary storage: $\boldsymbol{z}_t$, $\boldsymbol{v}_t$ and $\boldsymbol{x}_t$, all in $\mathbb{R}^p$
Result: approximate solution to (OPT)
for $t = 0, 1, \ldots$ **do**
  $\boldsymbol{z}_t = \mathbf{prox}_{\gamma h}(\boldsymbol{y}_t)$
  Sample $i \in \{1, \ldots, n\}$ uniformly at random
  $\boldsymbol{v}_t = \nabla\psi_i(\boldsymbol{z}_t) - \boldsymbol{\alpha}_{i,t} + \overline{\boldsymbol{\alpha}}_t + \nabla\omega(\boldsymbol{z}_t)$
  $\boldsymbol{x}_t = \mathbf{prox}_{\gamma g}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t)$
  $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \boldsymbol{x}_t - \boldsymbol{z}_t$
  Update $\boldsymbol{\alpha}_{t+1}$ according to (1)
return $\boldsymbol{z}_t$

where each ψ_i is convex and L_ψ-smooth, ω is convex and L_ω-smooth and g, h are *proximal*, i.e., convex and we have access to their proximal operator.

This formulation allows to express a broad range of problems arising in machine learning and signal processing: the finite-sum includes common loss functions such as least squares or logistic loss; the two proximal terms g, h can be extended to an arbitrary number and include penalties such as the group lasso with overlap, total variation, ℓ_1 trend filtering, etc. Furthermore, the proximal terms can be extended-valued, thus allowing for convex constraints through the use of the indicator function. With respect to previous work, this significantly enlarges the class of functions stochastic variance reduced methods can solve efficiently.

We allow the terms inside the finite sum to be an addition of two terms: ψ_i and ω. This might seem superfluous since it is not more general than the standard formulation with a single term. However, in practice ψ_i (e.g., a least squares or logistic loss, see Appendix F.1) can be highly structured and allow for reduced storage schemes and/or have sparse gradients (see §2.1), properties which might not be shared by ω, (e.g., an ℓ_2 regularization term).

Central to our algorithm is the concept of q-memorization (Hofmann et al., 2015), which we recall below. It provides a convenient abstraction over common gradient memorization techniques like the ones in SAGA and SVRG.

Definition 1. *A* uniform q-memorization *algorithm selects at each iteration t a random index set J_t of memory terms to update according to*

$$\boldsymbol{\alpha}_{j,t+1} = \begin{cases}\nabla f_j(\boldsymbol{z}_t) & \text{if } j \in J_t\\ \boldsymbol{\alpha}_{j,t} & \text{otherwise}\,,\end{cases} \tag{1}$$

such that any j has the same probability q/n of being updated.

We now introduce the variance-reduced three operator

splitting (Vr-Tos), a method to solve problems of the form (OPT). It is specified in Algorithm 1 and takes as input a vector of coefficients $\boldsymbol{y}_0 \in \mathbb{R}^p$, a table $\boldsymbol{\alpha}_0 \in \mathbb{R}^{n\times p}$ to store previous gradients and a step size $\gamma > 0$. Although in the general case this table is required to be of size $n \times p$, for linearly-parametrized loss functions like the logistic or least squares loss this can be reduced to size n (Appendix F.1). Furthermore, the Svrg-like update detailed below avoids the need for this storage at the expense of a lightly increased per iteration cost.

The proposed method performs one evaluation of each of the proximal terms and builds the gradient estimator $\boldsymbol{v}_t$ from the table of previous gradients $\boldsymbol{\alpha}_t$ and the index i sampled uniformly at random. It is easy to see that $\boldsymbol{v}_t$ is an unbiased estimate of the gradient, that is, $\mathbf{E}\, \boldsymbol{v}_t = \nabla f(\boldsymbol{z}_t)$.

This method allows the memory terms to be updated using any scheme that verifies the q-memorization framework (line 8). Some common schemes are:

- Saga-*like update*. At each iteration, the algorithm updates the same coefficient that has been sampled, i.e. $J_t = \{i\}$. In this scheme each memory term has probability $1/n$ of being updated, and so $q = 1$.

- Svrg-*like update*. Fix parameter $q > 0$ and draw at each iteration r from a uniform distribution in the $[0, 1]$ interval. If $r < q/n$, the algorithm performs a complete update $\boldsymbol{\alpha}_{j,t+1} = \nabla\psi_j(\boldsymbol{z}_t)$ for all j, otherwise they are left unchanged.

  Like in the Svrg algorithm (Johnson and Zhang, 2013), it is possible to avoid storing the memory terms since the $\overline{\boldsymbol{\alpha}}_t$ is constant unless a full refresh is triggered. In this setting, only the p-dimensional vectors $\overline{\boldsymbol{\alpha}}_t$ and $\widetilde{\boldsymbol{z}}_t$ needs to be stored, where $\widetilde{\boldsymbol{z}}_t$ is the value of $\boldsymbol{z}_t$ last time a full refresh was triggered. This variant avoids the need to store $\boldsymbol{\alpha}_t$, at the cost of a slight per iteration cost, as $\boldsymbol{\alpha}_i = \nabla f_i(\widetilde{\boldsymbol{z}}_t)$ needs to be computed at each iteration.

  This memory update scheme was proposed by Hofmann et al. (2015), and unlike the original Svrg algorithm the number of iterates between two full regresh is a random variable instead of a fixed number of iterations.

2.1 Sparse Vr-Tos

Need for a sparse variant. Modern web-scale optimization problems that arise in machine learning are not only large, they are also often *sparse*. For example, in the `LibSVM` datasets suite[2], 8 out of the 11 datasets with more than a million samples have a density below 0.01%, and the largest one in number of samples has a density below 1 per million. Linearly-parametrized loss functions of the form $\psi_i(\boldsymbol{x}) = l_i(\boldsymbol{a}_i^T\boldsymbol{x})$ have a gradient of the form $\nabla\psi_i(\boldsymbol{x}) = \boldsymbol{a}_i l_i'(\boldsymbol{a}_i^T\boldsymbol{x})$, which inherits the same sparsity pattern as the data $\boldsymbol{a}_i$. Since the data might be extremely sparse, it is hence of great practical interest to leverage sparsity in the partial gradients. This is the case in generalized linear models such as least squares or logistic regression, where $\boldsymbol{a}_i$ are the rows of a data matrix.

In this subsection we assume that g and ω are block separable, i.e., can be decomposed block coordinate-wise as $g(\boldsymbol{x}) = \sum_{B\in\mathcal{B}} g_B([\boldsymbol{x}]_B)$ and $\omega(\boldsymbol{x}) = \sum_{B\in\mathcal{B}}\omega_B([\boldsymbol{x}]_B)$, where $\mathcal{B}$ is a partition of the coefficients into subsets which will call *blocks* and g_B, ω_B only depends on coordinates in block B. Furthermore, we will make use of the following notation:

- *Extended support.* We define the extended support of $\nabla\psi_i$, denoted T_i as the set of blocks of $\mathcal{B}$ that intersect with its support, formally defined as $T_i \overset{\text{def}}{=} \{B : \operatorname{supp}(\nabla f_i) \cap B \neq \varnothing,\, B \in \mathcal{B}\}$. For totally separable penalties such as the ℓ_1 norm, the blocks are individual coordinates and so the extended support covers the same coordinates as the support.

- *Reweighting constants.* Let $\boldsymbol{P}_i$ be the projection onto the extended support, i.e., the diagonal matrix where $[\boldsymbol{P}_i]_{B,B}$ is the identity if $B \in T_i$ and zero otherwise. For simplicity we assume that each block appears in at least one T_i, as otherwise the problem can be reformulated without it. For each block $B \in \mathcal{B}$ we define d_B as the inverse frequency of that block in the extended support, i.e. $d_B = 1/(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{B \in T_i\})^{-1}$. For notational convenience we define the block-diagonal matrix $\boldsymbol{D}$ as $[\boldsymbol{D}]_{B,B} = d_B\boldsymbol{I}$ for each block $B \in \mathcal{B}$. Note that by definition $\frac{1}{n}\sum_{i=1}^n \boldsymbol{P}_i = \boldsymbol{D}^{-1}$. Computation of this diagonal matrix should be done as a preprocessing step of the algorithm.

- The *scaled proximal operator* is defined for a function φ, step size $\gamma > 0$, positive definite matrix $\boldsymbol{H}$ and norm $\|\cdot\|_{\boldsymbol{H}}^2 \overset{\text{def}}{=} \langle\cdot, \boldsymbol{H}\cdot\rangle$ as

$$\mathbf{prox}_{\gamma\varphi}^{\boldsymbol{H}}(\boldsymbol{x}) \overset{\text{def}}{=} \operatorname*{arg\,min}_{\boldsymbol{z}\in\mathbb{R}^p} \left\{ \varphi(\boldsymbol{z}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{z}\|_{\boldsymbol{H}}^2 \right\} \quad (2)$$

We now have all necessary ingredients to present the sparse variant of Vr-Tos. This is specified in Algorithm 2. In this variant, all operations are restricted to the extended support.

The algorithm requires to compute the scaled proximal operators of g and h. By block separability of g its scaled proximal operator can be computed in

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Algorithm 2: Sparse VR-TOS

Input: $\boldsymbol{y}_0 \in \mathbb{R}^p$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^{n\times p}$, $\gamma > 0$

Temporary storage: $\boldsymbol{z}_t$, $\boldsymbol{v}_t$ and $\boldsymbol{x}_t$, all in $\mathbb{R}^p$

Result: approximate solution to (OPT)

for $t = 0, 1, \ldots$ **do**

- Sample $i \in \{1, \ldots, n\}$ uniformly at random
- T_i = extended support of $\nabla\psi_i$
- $[\boldsymbol{z}_t]_{T_i} = [\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}_t)]_{T_i}$
- $[\boldsymbol{v}_t]_{T_i} = [\nabla\psi_i(\boldsymbol{z}_t) - \boldsymbol{\alpha}_{i,t} + \boldsymbol{D}(\overline{\boldsymbol{\alpha}}_t + \nabla\omega(\boldsymbol{z}_t))]_{T_i}$
- $[\boldsymbol{x}_t]_{T_i} = [\mathbf{prox}_{\gamma\varphi_i}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t)]_{T_i}$
- $[\boldsymbol{y}_{t+1}]_{T_i} = [\boldsymbol{y}_t + \boldsymbol{x}_t - \boldsymbol{z}_t]_{T_i}$
- Update $\boldsymbol{\alpha}_{t+1}$ according to (1)

return $\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}_t)$

block-wise as $[\mathbf{prox}_{\gamma g}^{\boldsymbol{D}^{-1}}(\boldsymbol{x})]_B = [\mathbf{prox}_{(d_B\gamma)h}(\boldsymbol{x})]_B$ for all $B \in \mathcal{B}$. Hence the cost of computing $[\boldsymbol{x}_t]_{T_i}$ will depend on the extended support size and not on the dimensionality.

We can unfortunately not guarantee the same complexity for $[\boldsymbol{z}_t]_{T_i}$ since we do not have a closed form for the scaled proximal operator of h in general. We review some specific cases in which it is possible to compute this scaled proximal operator in Appendix D. Alternatively, in the next subsection we propose a reformulation that avoids the need to compute this scaled proximal operator at the expense of higher memory usage.

In the case that one proximal term is zero, the proposed algorithm with SAGA-like update of the memory terms defaults to the Sparse SAGA variant of Pedregosa et al. (2017). With SVRG-like update of the memory terms it instead yields a novel sparse variant of ProxSVRG (Xiao and Zhang, 2014). For both of the proposed algorithms, when input is dense, $\boldsymbol{P}_i = \boldsymbol{D} = \boldsymbol{I}$ and we recover Algorithm 1.

2.2 Extension to an arbitrary number of proximal terms

The proposed method can be easily extended to the more general setting of an objective function with an arbitrary number of proximal terms of the form

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^p}{\text{minimize}}\ f(\boldsymbol{x}) + \textstyle\sum_{j=1}^k g_j(\boldsymbol{x})\,, \\ &\text{with } f(\boldsymbol{x}) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \psi_i(\boldsymbol{x}) + \omega(\boldsymbol{x})\ , \end{aligned} \tag{OPT-k}$$

where ψ_i and ω are as in (OPT) and $g_1, \ldots, g_k$ are proximal. This is done by expressing the above as a problem of the form (OPT) in an enlarged space and then applying the proposed algorithm to this reformulation. For this, we will introduce k new variables which we will constrain to be equal via an indicator function. The above problem can be written equivalently as follows,

$$\min_{\boldsymbol{X}\in\mathbb{R}^{k\times p}} f(\overline{\boldsymbol{X}}) + \underbrace{\textstyle\sum_{j=1}^k g_j(\boldsymbol{X}_j)}_{\stackrel{\text{def}}{=} g(\boldsymbol{X})} + \underbrace{\imath\{\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_k\}}_{\stackrel{\text{def}}{=} h(\boldsymbol{X})}\,,$$

where we have split the original variable into k variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ and constrained them to be equal using an indicator function in the last term. In this formulation the first term is smooth, and the other two terms are proximal. The second term is proximal since the variables in g_i are decoupled, each g_i is proximal by assumption and the last term is an indicator function over a linear subspace, and hence its scaled proximal operator can be computed in closed form as follows (Lemma 15):

$$\begin{aligned} &[\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{X})]_{i,j} = \left(\textstyle\sum_{i=1}^n a_{i,j}\boldsymbol{X}_{i,j}\right) / \left(\textstyle\sum_{i=1}^n a_{i,j}\right) \\ &\text{with } a_{i,j} = \boldsymbol{D}^{-1}_{ip+j,ip+j}\ , \end{aligned} \tag{3}$$

Hence, the problem with multiple proximal terms (OPT-k) can be formulated as a problem with two proximal terms (OPT) and so it is possible to apply the proposed method defined in the previous subsections. This gives a variance reduced method for problems with an arbitrary number of proximal term. It is worth noting that for the sparse variants this formulation avoids the potentially difficult computation of the scaled proximal operator of h.

3 Related work

We comment on the most closely related ideas, summarized in Table 1.

Methods that support objective functions of the form (OPT) with two or more proximal terms and a smooth term accessed via its gradient have recently been proposed. Examples are the the primal-dual hybrid gradient method (also known as the Condat-Vũ) (Condat, 2013a; Vũ, 2013),[3] the generalized forward-backward splitting (Raguet et al., 2013) or the three operator splitting (Davis and Yin, 2017). Due to its excellent empirical performance and amenability to sparse updates we have chosen this last method as the basis for the proposed method. The proposed VR-TOS method can be seen as a generalization of this last method, as both method are identical when $n = 1$.

A different stochastic variant of the three operator splitting was proposed by Yurtsever et al. (2016) for the slightly more general case in which f is given by

[3] We note that this method can optimize the more general objective function $f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{L}\boldsymbol{x})$, for an arbitrary linear operator $\boldsymbol{L}$ that is fixed to the identity in our setting.

Methods	incremental updates	non-decreasing step size	multiple non-smooth terms	sparse updates
Vr-Tos (*this work*)	✓	✓	✓	✓
Saga (Defazio et al., 2014)	✓	✓	✗	✓ (Pedregosa et al., 2017)
ProxSvrg (Xiao and Zhang, 2014)	✓	✓	✗	✗ †
Tos (Davis and Yin, 2017)	✗	✓	✓	N/A
Stochastic Tos (Yurtsever et al., 2016)	✓	✗	✓	✗

Table 1: **Comparison with related work.** The proposed method is unique in that it combines the advantages of variance-reduced methods (incremental updates, non-decreasing step sizes and sparse updates) with the advantages of proximal splitting (support for multiple non-smooth terms). †: a sparse variant of ProxSvrg follows as a special case of Algorithm 2 with $h = 0$ and the SVRG-like update of the memory terms.

an expectation. Like the proposed algorithms, this method only needs to evaluate the gradient of one element in the finite sum per iteration. Unlike the proposed methods, the variance of the updates does not decrease to zero and requires –as other non-variance reduced method– a decreasing step size. Furthermore, all updates are dense even in the presence of sparse gradients so the method performs poorly on large sparse problems.

(Balamurugan and Bach, 2016) proposed a variance-reduced method to solve problems a general class of saddle point problems including $\min_{\boldsymbol{x}} \max_{\boldsymbol{u}} \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x}) + M(\boldsymbol{x}, \boldsymbol{u})$, where $M(\cdot)$ is proximal. With $M(\boldsymbol{x}, \boldsymbol{u}) = g(\boldsymbol{x}) + \langle \boldsymbol{x}, \boldsymbol{u}\rangle - h^*(\boldsymbol{u})$, this is equivalent to the problem in (OPT). However, the method requires M to be strongly concave in $\boldsymbol{u}$, which is equivalent to h being smooth, and so is not applicable to the same class of problems as the proposed method. We note that this requirement is not merely an artifact of the theory, as the algorithm requires knowledge of this smoothness parameter.

Stochastic variance-reduced variants of ADMM have also been recently proposed, see e.g. (Zheng and Kwok, 2016; Yu and Huang, 2017). Compared to the proposed methods, none of the existing variants support sparse updates and require tuning more than one step-size parameter.

4 Analysis

In this section we provide a non-asymptotic convergence rate analysis for the proposed method:

- All the proposed variants converge with a step size $1/(3L_f)$, with $L_f \stackrel{\text{def}}{=} L_\psi + d_{\max} L_\omega$, where $d_{\max}$ is the maximum element in the diagonal matrix $\boldsymbol{D}$ ($d_{\max} = 1$ for non-sparse variants).
- For Vr-Tos (Algorithm 1) we obtain convergence rates that asymptotically match those of the full-gradient variant, i.e., $\mathcal{O}(1/t)$ convergence rate for convex problems (Theorem 1) and a linear convergence rate under strong convexity of f and smoothness of h (Theorem 3).
- For the sparse variant, Sparse Vr-Tos (Algorithm 2), we obtain a linear convergence rate under the same assumptions (Theorem 3). However, for general convex objectives we could only obtain a worse $\mathcal{O}(1/\sqrt{t})$ convergence rate (Theorem 2).

In this section we will use the following **extra notation**. We define the following primal ($\mathcal{P}$), and dual function ($\mathcal{D}$) as:

$$\mathcal{P}(\boldsymbol{x}) \stackrel{\text{def}}{=} f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x}) \,,$$
$$\mathcal{D}(\boldsymbol{u}) \stackrel{\text{def}}{=} (f + g)^*(-\boldsymbol{u}) + h^*(\boldsymbol{u}) \,, \tag{4}$$

where * denotes the Fenchel conjugate. We denote by $\boldsymbol{x}^\star$ an arbitrary minimizer of the primal objective and define the "dual iterate" $\boldsymbol{u}_t \stackrel{\text{def}}{=} \boldsymbol{D}^{-1}(\boldsymbol{y}_t - \boldsymbol{z}_t)/\gamma$ ($\boldsymbol{D} = \boldsymbol{I}$ for the dense variants). We also define the following generalized three operator splitting operator:

$$\boldsymbol{G}_\gamma(\boldsymbol{y}) \stackrel{\text{def}}{=} \boldsymbol{y} - \boldsymbol{z}_{\boldsymbol{y}} + \mathbf{prox}_{\gamma g}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z}_{\boldsymbol{y}} - \boldsymbol{y} - \gamma \boldsymbol{D}\nabla f(\boldsymbol{z}_{\boldsymbol{y}})) \,,$$
$$\text{with } \boldsymbol{z}_{\boldsymbol{y}} = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}) \,, \tag{5}$$

and its set of fixed points, which we denote $\mathrm{Fix}(\boldsymbol{G}_\gamma)$. Another quantity that will appear often in the analysis is $H_0 \stackrel{\text{def}}{=} 1/(2nL_f)\sum_{i=1}^n \|\boldsymbol{\alpha}_{i,0} - \psi_i(\boldsymbol{x}^\star)\|^2$.

Throughout this section we make the following two technical assumptions:

Assumption 1: Regularity. We assume each ψ_i is L_ψ-smooth, ω is L_ω-smooth, g and h are proper (i.e., have nonempty domain), lower semicontinuous (i.e., its sublevel sets are closed) convex functions. We recall

that lower semicontinuity is a weak form of continuity that allows extended-valued functions with domain over a closed set.

Assumption 2: Qualification conditions. We assume the relative interior of $\operatorname{dom} g$ and $\operatorname{dom} h$ have a non-empty intersection. This is a very weak and standard assumption, which allows to rule out pathological cases such as disjoint domains and allows to relate the primal and dual optimal objective (see e.g.(Bauschke and Combettes, 2017, Proposition 15.13) or (Bertsekas, 2015, Proposition 5.3.8)), a property sometimes referred to as strong or total duality.

Sublinear convergence. The following theorem shows a $\mathcal{O}(1/t)$ convergence rate for VR-TOS on arbitrary convex objectives.

One of the issues when analyzing the convergence of the three operator splitting is that the objective function might be $+\infty$, for example when both proximal terms are an indicator function. Following Chambolle and Pock (2015); Pedregosa and Gidel (2018), we will state the convergence rate for general functions in terms of the *saddle point suboptimality*, defined as

$$\mathcal{L}(\widetilde{\boldsymbol{x}}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \widetilde{\boldsymbol{u}}) \,, \quad \text{with}$$
$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{u}) \stackrel{\text{def}}{=} f(\boldsymbol{x}) + g(\boldsymbol{x}) + \langle \boldsymbol{x}, \boldsymbol{u} \rangle - h^*(\boldsymbol{u}) \,, \tag{6}$$

where $\mathcal{L}$ is the Lagrangian associated with $\mathcal{P}$ and $\mathcal{D}$. As Davis and Yin (2017), we will also state convergence rates in terms of the objective suboptimality under a Lipschitz assumption on h in (8).

Theorem 1. *Let $\overline{\boldsymbol{x}}_t$ denote the averaged (also known as ergodic) iterate, i.e., $\overline{\boldsymbol{x}}_t = (\sum_{k=0}^t \boldsymbol{x}_k)/(t+1)$ and $\overline{\boldsymbol{u}}_t = (\sum_{k=0}^t \boldsymbol{u}_k)/(t+1)$. Then the* VR-TOS *method (Algorithm 1) converges for any step size $\gamma \leq 1/(3L_f)$, and for $\gamma = 1/(3L_f)$ we have the following bound for all $(\boldsymbol{x}, \boldsymbol{u}) \in \operatorname{dom} g \times \operatorname{dom} h^*$:*

$$\mathbb{E}\left[\mathcal{L}(\overline{\boldsymbol{x}}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \overline{\boldsymbol{u}}_t)\right] \leq \frac{10n}{q(t+1)} C_0, \tag{7}$$

with $\boldsymbol{y} = \boldsymbol{x} + \gamma \boldsymbol{u}$, $\boldsymbol{y}^\star \in \operatorname{Fix}(\boldsymbol{G}_\gamma)$, and $C_0 = \left[\frac{3L_f q}{20n}\|\boldsymbol{y}_0 - \boldsymbol{y}\|^2 + \frac{3L_f q}{2n}\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + H_0\right]$, where we recall $H_0 = 1/(2nL_f)\sum_{i=1}^n \|\boldsymbol{\alpha}_{i,0} - \psi_i(\boldsymbol{x}^\star)\|^2$.

Furthermore, if h is β_h-Lipschitz we have the following rate in terms of the primal objective:

$$\mathcal{P}(\overline{\boldsymbol{x}}_t) - \mathcal{P}(\boldsymbol{x}^\star) \leq \frac{10n}{q(t+1)} \widetilde{C}_0 \,, \tag{8}$$

with $\widetilde{C}_0 = \frac{6L_f q}{20n}\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{3L_f q}{2n}\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + \frac{q}{15nL_f}\beta_h^2 + H_0$.

The previous theorem gives a $\mathcal{O}(1/t)$ convergence rate in terms of the saddle point suboptimality for arbitrary convex functions and $\mathcal{O}(1/t)$ rate in function suboptimality under a Lipschitz assumption on h, matching the strongest bounds of SAGA (Defazio et al., 2014).

For their sparse variants, however, we have only been able to prove a slower $\mathcal{O}(1/\sqrt{t})$ rate on the operator residual, despite the fact that in practice the algorithm exhibits a much faster empirical convergence (see §5). Appendix B contains a characterization of the fixed points of this operator that justifies why this is a meaningful suboptimality criterion for (OPT). Although there is no direct correspondence between rates on the gradient and on objective values, lower bounds are asymptotically equivalent (Nesterov, 2012).

Theorem 2. *Sparse* VR-TOS *(Algorithm 2) converges for every step size $\gamma \leq 1/(3L_f)$. In particular, for $\gamma = 1/(3L_f)$ and $\boldsymbol{y}_t$ obtained after $t \geq 1$ updates we have the bound*

$$\min_{k=0,\dots,t} \{\mathbb{E}\|\boldsymbol{y}_k - \boldsymbol{G}_\gamma(\boldsymbol{y}_k)\|\} \leq \sqrt{\frac{C_0}{Lq(t+1)}} = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right), \tag{9}$$

with $C_0 = \frac{5d_{\max}n}{Lq(t+1)}\left[(2Lq/n)\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + H_0\right]$.

Linear convergence. The three operator splitting has been shown to have a linear convergence rate under the assumption of strong convexity of the smooth term and smoothness of one of the proximal terms (Davis and Yin, 2015, §4.4). Although this last condition is rarely verified in practice since its main application is on non-smooth proximal terms, it is instructive to see that the proposed method –despite the reduced cost per iteration– also enjoys a linear convergence rate under the same assumptions.

Theorem 3 (Linear convergence)**.** *Let ψ_i be μ_ψ-strongly convex and ω be μ_ω-strongly convex, where $\mu_\psi + \mu_\omega > 0$. Furthermore, let h be L_h-smooth. Then for any step size $\gamma \leq 1/(3L_f)$, all the proposed methods converge geometrically in expectation. For $\gamma = 1/(3L_f)$, we have the following bound for Algorithm 1 ($d_{\max} = 1$ in this case) and Algorithm 2:*

$$\mathbb{E}\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \min\left\{\frac{q}{4n}, \frac{1}{3d_{\max}^3\delta^2\kappa}\right\}\right)^t D_0 \quad , \tag{10}$$

with $D_0 \stackrel{def}{=} d_{\max}\left[\frac{q}{2\gamma(1-\gamma\mu)n}\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + H_0\right]$, $\delta = (1 + L_h/(3L_f))$, $\kappa = L_f/\mu$ and $\boldsymbol{y}^\star \in \operatorname{Fix}(\boldsymbol{G}_\gamma)$.

4.1 Discussion

Comparison of convergence rates. We summarize the obtained convergence rates for the proposed methods and compare them against the best known rates for related stochastic methods in Table 2. In the linearly-convergent regime, we obtain rates that are

	Method	step size	Proximal oracle	Convergence rate	Extra assumptions
Geometric	SAGA (Defazio et al., 2014)	$1/3L_f$	$\mathbf{prox}_{\gamma(g+h)}$	$\left(1 - \min\left\{\frac{1}{4n}, \frac{1}{3\kappa}\right\}\right)^t C_0$	Each ψ_i is μ-cvx
	ProxSVRG (Xiao and Zhang, 2014)	$1/10L_f$	$\mathbf{prox}_{\gamma(g+h)}$	$\left(\frac{1}{\kappa 0.6m} + \frac{2}{3}\right)^t C_0$	f is μ-cvx
	VR-TOS (this work)	$1/3L_f$	$\mathbf{prox}_{\gamma g}$, $\mathbf{prox}_{\gamma h}$	$\left(1 - \min\left\{\frac{q}{4n}, \frac{1}{3d_{\max}^2 \delta^2 \kappa}\right\}\right)^t C_0$	Each ψ_i is μ-cvx and h is L_h-smooth
Sublinear	SAGA (Defazio et al., 2014)	$1/3L_f$	$\mathbf{prox}_{\gamma(g+h)}$	$\mathcal{O}(1/t)$	None
	Stochastic TOS (Yurtsever et al., 2016)	$\mathcal{O}(1/t)$	$\mathbf{prox}_{\gamma g}$, $\mathbf{prox}_{\gamma h}$	$\mathcal{O}(1/t)$	f is μ-cvx + bound on gradients
	VR-TOS (this work, dense/sparse variant)	$1/3L_f$	$\mathbf{prox}_{\gamma g}$, $\mathbf{prox}_{\gamma h}$	$\mathcal{O}(1/t)$ / $\mathcal{O}(1/\sqrt{t})$	None

Table 2: **Assumptions and properties of related incremental methods.** In every case, we take the step size recommended by the theory, where we assume $\omega = 0$ to make them comparable. Proximal oracle is the proximal operators that are needed by the algorithm. Extra assumptions refer to those other than Assumptions 1 and 2. The linear rates use the quantities $\delta = (1 + \gamma L_h)$, $\kappa = L_f/\mu$. For ProxSVRG, m denotes the epoch size and the convergence rate is relative to the number of epochs and not iterations like the rest.

similar to SAGA but with the rate factor multiplied by $1/(\delta^2 d_{\max}^3)$, quantity that depends on the smoothness of g and the sparsity of the gradients.

An improved ProxSVRG variant. The analysis of ProxSVRG (Xiao and Zhang, 2014) requires that the step size verifies an implicit equation that depends among other things on the strong convexity parameter. For typical choices of the parameters this is $1/(10L_f)$ (Xiao and Zhang, 2014, Theorem 1). In contrast, Sparse VR-TOS with SVRG-like sampling with $h = 0$ yields a variant of ProxSVRG with more favorable properties. First, none of its parameters depend on the strong convexity constant (while still obtaining a linear convergence rate since $L_h = 0$ in this case), which is most often unknown. Second, it admits the much larger step size $1/(3L_f)$, which is, to the best of our knowledge, the largest step size of any SVRG variant. Third, it can leverage sparsity in the input data through sparse updates.

Linear convergence without smoothness of the proximal term. Theorem 3 requires smoothness of one of the proximal terms to guarantee linear convergence. Despite this, linear convergence is observed in practice without this assumption (Figure 1). This has also been observed in the case of the original (non-variance reduced) three operator splitting (Davis and Yin, 2017; Pedregosa and Gidel, 2018), although an explanation for this is still an open problem. Furthermore, the lack of linear convergence when both proximal terms are non-smooth does not seem to be a limitation of the proof, as a counterexample was provided in (Davis and Yin, 2015, Appendix D.6). In this work, the authors constructed a strongly monotone operator with a sublinear convergence.

Step size adaptivity to linear convergence. A practical consequence of the above theorems is that using the same step size $\gamma = 1/(3L_f)$ we obtain a sublinear convergence by Theorem 1 and a linear rate (under additional assumptions) by Theorem 3. That is, one can use the "universal" step size $1/(3L_f)$ and automatically obtain linear convergence whenever the assumptions of Theorem 3 are verified.

Limitations. The following are some scenarios under which the proposed method is expected to perform poorly. The cost in computation and storage scales linearly with the number of proximal terms, hence it cannot cope with other scenarios with many non-smooth terms such as empirical risk minimization with the hinge loss or group lasso with overlap with a large number of overlaps (for instance > 100). Also, there are still penalties that cannot be reduced to a sum of proximal terms, such as the nuclear norm. Algorithms based on Frank-Wolfe (Jaggi, 2013) or with approximate proximal operators (Schmidt et al., 2011) might be better suited in such regimes.

5 Experiments

Although the proposed methods can be applied more broadly, we consider for the experiments a logistic regression problem with squared ℓ_2 regulrization and an overlapping group lasso penalty (Jacob et al., 2009). Following Jacob et al. (2009) we choose groups of 10 variables with 2 variables of overlap between two successive groups: $\{\{1, \dots, 10\}, \{8, \dots, 18\}, \{16, \dots, 26\}, \dots\}$. The amount of group regularization was chosen such that the solution has roughly 10% of non-zero coefficients and the of ℓ_2 regularization was fixed to $1/n$. We consider the following methods:

Dataset	#samples	#dimensions	density	L_f/μ
RCV1 (full) (Lewis et al., 2004)	697,641	47,236	1.5×10^{-3}	2.50×10^{4}
URL (Ma et al., 2009)	2,396,130	3,231,961	3.5×10^{-5}	1.28×10^{7}
KDD10 (Yu et al., 2010)	19,264,097	29,890,095	9.8×10^{-7}	5.2×10^{8}
Criteo (Juan et al., 2016)	45,840,617	1,000,000	3.8×10^{-5}	1.1×10^{7}

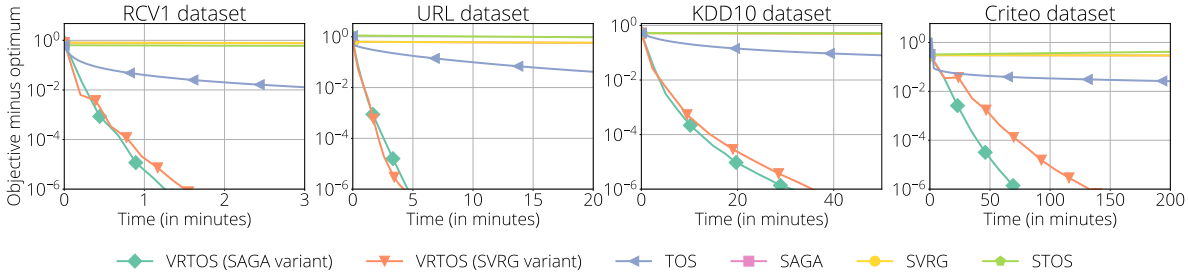

Figure 1: **Top**: Description of considered datasets. **Bottom:** Suboptimality vs time of different algorithms on a logistic regression with overlapping group lasso penalty problem.

- The proposed method Sparse Vr-Tos (Algorithm 2), where the overlapping group lasso penalty is split as a sum of two non-overlapping group lasso penalties, for which the proximal operator is available in closed form. We used the formulation with 3 proximal terms of §2.2 to better leverage sparsity in the dataset and consider Saga and Svrg-like updates, denoted Vr-Tos (Saga variant) and Vr-Tos (Svrg variant) respectively. This implementation is publicly available in the C-OPT package.[4]

  It is worth noting that while original penalty is *not* block separable, each of the terms in the splitting as two group lasso penalties *is* block separable. This will allow us to make a much more efficient use of sparsity than what is possible on on methods like Saga and ProxSvrg.

- The three operator splitting (denoted Tos), in its recently proposed variant with adaptive step size (Pedregosa and Gidel, 2018).

- The stochastic three operator spitting of (Yurtsever et al., 2016) with the same splitting as Vr-Tos, denoted Stos.

- Saga and ProxSvrg, where the proximal operator is evaluated approximately using 10 iterations of the Douglas-Rachford method.

The above methods were compared on 4 large-scale datasets described in the table of Figure 1. Further details and implementation aspects are discussed in Appendix F.1.

The best performing algorithms overall are the proposed Vr-Tos variants, which are over an order of magnitude faster than the second best method, the adaptive three operator splitting. The stochastic three operator splitting, not being able to take advantage of the sparsity in the gradients, performs poorly in this benchmark, appearing as a straight line. Saga and ProxSvrg were the slowest since they require to compute a costly proximal operator at each iteration and are unable to leverage the sparsity of the dataset due to the non-block-separability of the non-smooth term.

It is worth noting from Figure 1 that the two variants of Sparse Vr-Tos exhibit an empirical linear convergence, despite the fact that the theory only predicts in this regime a much slower $\mathcal{O}(1/\sqrt{t})$ convergence rate (Theorem 1).

We provide extra experiments in Appendix F.2.

6 Future work

This work can be extended in several ways. As highlighted in §4.1, a theoretical explanation for the empirical linear convergence without smoothess of any proximal term, even for the full gradient algorithm, is lacking. We conjecture *partly smooth* is a sufficient condition on the penalties to ensure local linear convergence, as recently proven for related methods (Liang et al., 2018). Second, we conjecture that the convergence rate of the sparse variant can be improved up to to $\mathcal{O}(1/t)$. A third direction for future work would be the development an extension that allow for a linear operator inside one of the proximal terms, as in (Condat, 2013b; Zhao and Cevher, 2018; Yan, 2018).

[4] http://openopt.github.io/copt/

Acknowledgements

The authors warmly thank Vincent Roulet, Vlad Niculae, Rémi Leblond and Federico Vaggi for their feedback on the manuscript, as well as Adrien Taylor, Alexandre D'Aspremont, Gabriel Peyré, Guillaume Obozinski, P. Balamurugan, Francis Bach and Marwa El Halabi for fruitful discussions.

This work has been done while FP was under funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodorowska-Curie grant agreement 748900. KF is funded through the project OATMIL ANR-17-CE23-0012 of the French National Research Agency (ANR). Computing time on was donated by Amazon through the program "AWS Cloud Credits for Research".

References

Balamurugan, P. and Bach, F. (2016). Stochastic Variance Reduction Methods for Saddle-Point Problems. *Advances in Neural Information Processing Systems*.

Barbero, Á. and Sra, S. (2014). Modular proximal optimization for multidimensional total-variation regularization. *arXiv preprint arXiv:1411.0589*.

Bauschke, H. H., Boţ, R. I., Hare, W. L., and Moursi, W. M. (2012). Attouch–Théra duality revisited: paramonotonicity and operator splitting. *Journal of Approximation Theory*.

Bauschke, H. H. and Combettes, P. L. (2017). *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media.

Bertsekas, D. P. (2015). *Convex optimization algorithms*. Athena Scientific Belmont.

Chambolle, A. and Pock, T. (2015). On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*.

Condat, L. (2013a). A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*.

Condat, L. (2013b). A direct algorithm for 1D total variation denoising. *IEEE Signal Processing Letters*.

Davis, D. and Yin, W. (2015). A three-operator splitting scheme and its optimization applications. *preprint arXiv:1504.01032v1*.

Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*.

Giselsson, P. and Boyd, S. (2016). Linear Convergence and Metric Selection in Douglas-Rachford Splitting and ADMM. *IEEE Transactions on Automatic Control*.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. (2015). Variance Reduced Stochastic Gradient Descent with Neighbors. In *Advances in Neural Information Processing Systems*.

Iusem, A. N. (1998). On Some Properties of Generalized Proximal Point Methods for Variational Inequalities. *Journal of Optimization Theory and Applications*.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*. ACM.

Jaggi, M. (2013). Revisiting Frank-Wolfe: projection-free sparse convex optimization. In *International Conference on Machine Learning*.

Johnson, N. (2013). A dynamic programming algorithm for the fused lasso and L_0-segmentation. *Journal of Computational and Graphical Statistics*.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*.

Juan, Y., Zhuang, Y., Chin, W.-S., and Lin, C.-J. (2016). Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM.

Kim, S.-J., Koh, K., Boyd, S., et al. (2009). ℓ_1 trend filtering. *SIAM review*.

Le Roux, N., Schmidt, M., and Bach, F. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Liang, J., Fadili, J., and Peyré, G. (2018). Local linear convergence analysis of primal–dual splitting methods. *Optimization*.

Ma, J., Saul, L. K., et al. (2009). Identifying suspicious URLs: an application of large-scale online learning. In *Proceedings 26th ACM international conference on machine learning*.

Nesterov, Y. (2004). *Introductory lectures on convex optimization*. Springer.

Nesterov, Y. (2012). How to make the gradients small. *Optima*.

Pedregosa, F. and Gidel, G. (2018). Adaptive Three Operator Splitting. *Proceedings of the 35th International Conference on Machine Learning*.

Pedregosa, F., Leblond, R., and Lacoste-Julien, S. (2017). Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization. *Advances in Neural Information Processing System 30 (NIPS)*.

Raguet, H., Fadili, J., and Peyré, G. (2013). A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Ann. Math. Statist.*

Rockafellar, R. T. (1997). Convex analysis.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational analysis*. Springer.

Schmidt, M., Le Roux, N., and Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems 24*.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*.

Vũ, B. C. (2013). A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*.

Xiao, L. and Zhang, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*.

Yan, M. (2018). A New Primal–Dual Algorithm for Minimizing the Sum of Three Functions with a Linear Operator. *Journal of Scientific Computing*.

Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, J.-K., McKenzie, T. G., Chou, J.-W., Chung, P.-H., Ho, C.-H., Chang, C.-F., Wei, Y.-H., et al. (2010). Feature engineering and classifier ensemble for KDD cup 2010. In *KDD Cup*.

Yu, Y. and Huang, L. (2017). Fast stochastic variance reduced admm for stochastic composition optimization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*.

Yurtsever, A., Vu, C. B., and Cevher, V. (2016). Stochastic Three-Composite Convex Minimization. In *Advances in Neural Information Processing Systems*.

Zhao, R. and Cevher, V. (2018). Stochastic Three-Composite Convex Minimization with a Linear Operator. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*.

Zheng, S. and Kwok, J. T. (2016). Stochastic variance-reduced ADMM. *arXiv preprint arXiv:1604.07070*.

Proximal Splitting Meets Variance Reduction

Supplementary material

Outline. The supplementary material of this paper is organized as follows.

- Appendix A presents basic definitions and properties that will be used throughout the proofs but which are not specific to our methods. Most of these can be found in convex optimization textbooks, such as (Bauschke and Combettes, 2017; Nesterov, 2004).
- Appendix B give a characterization of the fixed points of the three operator splitting, relating the set of fixed points of the three operator splitting to the solutions of primal and dual objectives. This is a stronger result than the one stated in (Davis and Yin, 2017) and used in some of our proofs.
- Appendix C gives the proofs of those results in the Analysis section of the paper.
- Appendix D discusses splitting strategies for different penalties and examines some cases in which the scaled proximal operator can be computed in closed form.
- Appendix F discusses implementation aspects of the proposed algorithms.

Appendix A Basic definitions and properties

Definition 2 (proper function). *A function $f : \mathcal{X} \subseteq \mathbb{R}^p \to]-\infty, \infty]$ is said to be proper if its domain is not empty.*

Definition 3 (Fenchel conjugate). *The Fenchel conjugate of a function $f : \mathcal{X} \subseteq \mathbb{R}^p \to]-\infty, \infty]$ is defined as*

$$f^*(\boldsymbol{x}^\star) = \sup \{ \langle \boldsymbol{x}^\star, \boldsymbol{x} \rangle - f(\boldsymbol{x}) | \, \boldsymbol{x} \in \mathcal{X} \}. \tag{11}$$

Definition 4 (lower semicontinuity). *We say that a proper convex function f is lower-semicontinuous if all of its levelsets $\{\boldsymbol{x} \in \operatorname{dom}(f) \,|\, f(\boldsymbol{x}) \leq \alpha\}$ are closed.*

Definition 5 (relative interior). *The relative interior of a convex set $C \subseteq \mathbb{R}^p$ is defined as*

$$\operatorname{relint}(C) \stackrel{def}{=} \{\boldsymbol{x} \in C : \forall \boldsymbol{y} \in C \; \exists \lambda > 1 : \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in C\} \tag{12}$$

Definition 6 (Bregman divergence). *The Bregman divergence associated with a convex function f for points $\boldsymbol{x}, \boldsymbol{y}$ in its domain is defined as:*

$$B_f(\boldsymbol{x}, \boldsymbol{y}) \stackrel{def}{=} f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$$

Note that this is always positive due to the convexity of f.

Definition 7 (Proximal operators). *Here, we redefine 2 variants of a critical notion. The proximal operator is defined for a function φ, step size $\gamma > 0$ as:*

$$\mathbf{prox}_{\gamma\varphi}(\boldsymbol{x}) \stackrel{def}{=} \operatorname*{arg\,min}_{\boldsymbol{z} \in \mathbb{R}^p} \Big\{ \varphi(\boldsymbol{z}) + \frac{1}{2\gamma} \|\boldsymbol{x} - \boldsymbol{z}\| \Big\} \tag{13}$$

The scaled proximal opeartor *is defined for a function φ, step size $\gamma > 0$ and positive definite matrix $\boldsymbol{H}$ as the solution of the following optimization problem*

$$\mathbf{prox}^{\boldsymbol{H}}_{\gamma\varphi}(\boldsymbol{x}) \stackrel{def}{=} \operatorname*{arg\,min}_{\boldsymbol{z} \in \mathbb{R}^p} \Big\{ \varphi(\boldsymbol{z}) + \frac{1}{2\gamma} \|\boldsymbol{x} - \boldsymbol{z}\|^2_{\boldsymbol{H}} \Big\} \quad \text{with } \|\cdot\|^2_{\boldsymbol{H}} \stackrel{def}{=} \langle \cdot, \boldsymbol{H} \cdot \rangle. \tag{14}$$

Lemma 1 (subgradient characterization of proximal operator)**.** *Let g be a convex proper lower semicontinuous function. Then for any $\boldsymbol{x}$, $\boldsymbol{H}$ positive definite matrix and any $\gamma > 0$ we have the following characterization of proximal operator:*

$$\boldsymbol{z} = \mathbf{prox}_{\gamma g}^{\boldsymbol{H}}(\boldsymbol{x}) \iff \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{z})/\gamma \in \partial g(\boldsymbol{z}) \tag{15}$$

Proof. By the definition of proximal operator we have that $\boldsymbol{z} = \mathbf{prox}_{\gamma g}^{\boldsymbol{H}}(\boldsymbol{x})$ is equivalent to

$$\boldsymbol{z} \in \operatorname*{arg\,min}_{\boldsymbol{z}' \in \mathbb{R}^p} g(\boldsymbol{z}') + \frac{1}{2\gamma}\|\boldsymbol{z}' - \boldsymbol{x}\|_{\boldsymbol{H}}^2 \tag{16}$$

$$\iff 0 \in \partial g(\boldsymbol{z}) + \frac{\boldsymbol{H}}{\gamma}(\boldsymbol{z} - \boldsymbol{x}) \tag{17}$$

$$\iff \frac{\boldsymbol{H}}{\gamma}(\boldsymbol{x} - \boldsymbol{z}) \in \partial g(\boldsymbol{z}) \tag{18}$$

where the first equivalence is a consequence of the first order optimality conditions. □

Lemma 2 (Conjugate-inverse identity)**.** *Let h be a convex, proper lower semicontinuous function. Then*

$$\boldsymbol{u} \in \partial h(\boldsymbol{z}) \iff \boldsymbol{z} \in \partial h^*(\boldsymbol{u}) \ . \tag{19}$$

In other words, $(\partial h)^{-1} = \partial h^*$.

Proof. See e.g. (Bauschke and Combettes, 2017, Corollary 16.30) or (Rockafellar and Wets, 1998, Proposition 11.3). □

Lemma 3 (Generalized variance decomposition)**.** *Let ζ_i be a random variable and let $\mathbf{E}$ the expectation with respect to this random variable. Furthermore, let $\boldsymbol{Q}_i$ be an orthogonal projection such that $\boldsymbol{Q}_i\boldsymbol{\zeta}_i = \boldsymbol{\zeta}_i$, $\mathbf{E}\boldsymbol{Q}_i$ is invertible and $\boldsymbol{A} = (\mathbf{E}\boldsymbol{Q}_i)^{-1}$. Then we have*

$$\mathbf{E}\|\boldsymbol{\zeta}_i - \boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\|^2 = \mathbf{E}\|\boldsymbol{\zeta}_i\|^2 - \|\mathbf{E}\boldsymbol{\zeta}_i\|_{\boldsymbol{A}}^2 \ . \tag{20}$$

Proof. The assumption of $\boldsymbol{Q}_i$ being an orthogonal projection implies that it is symmetric and idempotent. Developing the square we have

$$\mathbf{E}\|\boldsymbol{\zeta}_i - \boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\|^2 = \mathbf{E}\|\boldsymbol{\zeta}_i\|^2 + \mathbf{E}\langle \boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i, \boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle - 2\mathbf{E}\langle \boldsymbol{\zeta}_i, \boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle \tag{21}$$

$$= \mathbf{E}\|\boldsymbol{\zeta}_i\|^2 + \mathbf{E}\langle \boldsymbol{Q}_i\boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i, \boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle - 2\mathbf{E}\langle \boldsymbol{Q}_i\zeta_i, \boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle \tag{22}$$

(by symmetry of $\boldsymbol{Q}_i$)

$$= \mathbf{E}\|\boldsymbol{\zeta}_i\|^2 + \mathbf{E}\langle \boldsymbol{Q}_i\boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i, \boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle - 2\mathbf{E}\langle \boldsymbol{\zeta}_i, \boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle \tag{23}$$

(idempotence of $\boldsymbol{Q}_i$ and assumption $\boldsymbol{Q}_i\boldsymbol{\zeta}_i = \boldsymbol{\zeta}_i$ respectively)

$$= \mathbf{E}\|\boldsymbol{\zeta}_i\|^2 + \langle \mathbf{E}\boldsymbol{\zeta}_i, \boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle - 2\langle \mathbf{E}\boldsymbol{\zeta}_i, \boldsymbol{A}\mathbf{E}\boldsymbol{\zeta}_i\rangle \tag{24}$$

(taking expectations)

$$= \mathbf{E}\|\boldsymbol{\zeta}_i\|^2 - \|\mathbf{E}\boldsymbol{\zeta}_i\|_{\boldsymbol{A}}^2 \tag{25}$$

□

Lemma 4 (Smooth inequality 1)**.** *Let f_i be L_f-smooth and convex for $i = 1, \dots, n$. Then it is verified that*

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|^2 \leq 2L_f B_f(\boldsymbol{x}, \boldsymbol{y}) \ . \tag{26}$$

Proof. Since each f_i is L_f-smooth, it is verified (Nesterov, 2004, Theorem 2.1.5) that

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|^2 \leq 2L_f(f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) - \langle \nabla f_i(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle) \tag{27}$$

The result is obtained by averaging over i. □

Lemma 5 (Bound on matrix norm)**.** *Let $\boldsymbol{D}$ be a diagonal matrix with strictly positive diagonal elements, let $d_{\max}$ and $d_{\min}$ denote its maximum and minimum diagonal entry respectively. Then for any $\boldsymbol{x} \in \mathbb{R}^p$ we have the following inequalities*

$$d_{\min} \|\boldsymbol{x}\|^2 \leq \|\boldsymbol{x}\|_{\boldsymbol{D}}^2 \leq d_{\max} \|\boldsymbol{x}\|^2 \tag{28}$$

Proof. By definition of the $\boldsymbol{D}$-norm we have

$$\|\boldsymbol{x}\|_{\boldsymbol{D}}^2 = \sum_{i=1}^n \boldsymbol{D}_{i,i} \boldsymbol{x}_i^2 \leq \sum_{i=1}^n d_{\max} \boldsymbol{x}_i^2 = d_{\max} \|\boldsymbol{x}\|^2 \tag{29}$$

$$\|\boldsymbol{x}\|_{\boldsymbol{D}}^2 = \sum_{i=1}^n \boldsymbol{D}_{i,i} \boldsymbol{x}_i^2 \geq \sum_{i=1}^n d_{\min} \boldsymbol{x}_i^2 = d_{\min} \|\boldsymbol{x}\|^2 \tag{30}$$

The result follows from chaining both inequalities □

Lemma 6 (Properties of proximal operator)**.** *Let g be a convex lower semicontinuous function and $\boldsymbol{H}$ a symmetric positive definite matrix. Then for all $\boldsymbol{y}, \widetilde{\boldsymbol{y}}$ we have the following inequality, often referred to as firm nonexpansiveness:*

$$\|\operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\boldsymbol{y}) - \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\widetilde{\boldsymbol{y}})\|_{\boldsymbol{H}}^2 \leq \langle \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\boldsymbol{y}) - \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\widetilde{\boldsymbol{y}}), \boldsymbol{y} - \widetilde{\boldsymbol{y}} \rangle_{\boldsymbol{H}} \tag{31}$$

Furthermore, if g is L_g-smooth and $\boldsymbol{H}$ has smallest singular value $\sigma_{\min}$ and largest singular value $\sigma_{\max}$, then we also have the following bound:

$$\|\boldsymbol{y} - \widetilde{\boldsymbol{y}}\| \leq (\sigma_{\max}/\sigma_{\min} + \gamma L \sigma_{\max}) \|\operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\boldsymbol{y}) - \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\widetilde{\boldsymbol{y}})\| \tag{32}$$

Proof. *First inequality*. Let $\boldsymbol{z} \overset{\text{def}}{=} \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\boldsymbol{y})$, $\widetilde{\boldsymbol{z}} \overset{\text{def}}{=} \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\widetilde{\boldsymbol{y}})$. By the subgradient characterization of Lemma 1 we have

$$\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{z})/\gamma \in \partial g(\boldsymbol{z}) \quad , \quad \boldsymbol{H}^{-1}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{z}})/\gamma \in \partial g(\widetilde{\boldsymbol{z}}) \tag{33}$$

Since the subdifferential of a convex function is monotonous, in particular ∂g is monotonous, and so we have

$$\langle \boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{z})/\gamma - \boldsymbol{H}^{-1}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{z}})/\gamma, \boldsymbol{z} - \widetilde{\boldsymbol{z}} \rangle \geq 0 \iff \langle (\boldsymbol{y} - \boldsymbol{z}) - (\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{z}}), \boldsymbol{z} - \widetilde{\boldsymbol{z}} \rangle_{\boldsymbol{H}^{-1}} \geq 0 \tag{34}$$

$$\iff \langle \boldsymbol{y} - \widetilde{\boldsymbol{y}}, \boldsymbol{z} - \widetilde{\boldsymbol{z}} \rangle_{\boldsymbol{H}^{-1}} \geq \|\boldsymbol{z} - \widetilde{\boldsymbol{z}}\|_{\boldsymbol{H}^{-1}} \tag{35}$$

which proves the first part of the lemma (firm nonexpansive).

Second inequality. To prove the second inequality we will use a generalization of the argument from Giselsson and Boyd (2016, Proposition 1). Let g_γ be defined as

$$g_\gamma \overset{\text{def}}{=} \gamma g + \frac{1}{2} \|\cdot\|_{\boldsymbol{H}^{-1}}^2 \quad . \tag{36}$$

By the subgradient characterization of the proximal operator (Lemma 1) and the conjugate-inverse identity (Lemma 2), we have

$$\boldsymbol{z} = \operatorname{\mathbf{prox}}_{\gamma g}^{\boldsymbol{H}^{-1}}(\boldsymbol{y}) \tag{37}$$

$$\iff \boldsymbol{z} \in \{\boldsymbol{x} \,|\, \boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{x}) \in \gamma \nabla g(\boldsymbol{x})\} \qquad \text{(Lemma 1)} \tag{38}$$

$$\iff \boldsymbol{z} \in \{\boldsymbol{x} \,|\, \boldsymbol{H}^{-1}\boldsymbol{y} \in \nabla g_\gamma(\boldsymbol{x})\} \tag{39}$$

$$\iff \boldsymbol{z} \in (\nabla g_\gamma)^{-1}(\boldsymbol{H}^{-1}\boldsymbol{y}) \tag{40}$$

$$\iff \boldsymbol{z} = \nabla g_\gamma^*(\boldsymbol{H}^{-1}\boldsymbol{y}) \quad \text{(Lemma 2)} \tag{41}$$

where g_γ^* denotes the convex conjugate of g_γ. Note that we can write ∇ in the last term instead of ∂ for g_γ^* because this function is 1-smooth with respect to the $\boldsymbol{H}^{-1}$-norm by the strong convexity of g_γ.

The term $\frac{1}{2}\|\cdot\|^2_{\boldsymbol{H}^{-1}}$ is $\sigma_{\min}^{-1}$-smooth and so g_γ is $(\sigma_{\min}^{-1}+\gamma L_g)$-smooth. By the duality between Lipschitz gradient and strong convexity (see e.g., Rockafellar and Wets (1998)), g_γ^* is $1/(\sigma_{\min}^{-1}+\gamma L)$-strongly convex. Then for arbitrary $\boldsymbol{y}$ and $\widetilde{\boldsymbol{y}}$ we have

$$\|\nabla g_\gamma^*(\boldsymbol{H}^{-1}\boldsymbol{y}) - \nabla g_\gamma^*(\boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}})\|\|\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}}\| \tag{42}$$

$$\geq \langle \nabla g_\gamma^*(\boldsymbol{H}^{-1}\boldsymbol{y}) - \nabla g_\gamma^*(\boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}}), \boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}}\rangle \qquad \text{(by Cauchy-Schwarz)} \tag{43}$$

$$\geq \frac{1}{1/\sigma_{\min}+\gamma L_g}\|\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}}\|^2 \qquad \text{(by strong convexity)} \tag{44}$$

The result is trivial when $\boldsymbol{y} = \widetilde{\boldsymbol{y}}$. We can then assume $\boldsymbol{y} \neq \widetilde{\boldsymbol{y}}$, and dividing both sides by $\|\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}}\|$ (non-zero by assumption) we obtain

$$\|\nabla g_\gamma^*(\boldsymbol{H}^{-1}\boldsymbol{y}) - \nabla g_\gamma^*(\boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}})\| \geq \frac{1}{1/\sigma_{\min}+\gamma L_g}\|\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}}\| \tag{45}$$

$$\geq \frac{\sigma_{\max}^{-1}}{1/\sigma_{\min}+\gamma L_g}\|\boldsymbol{y} - \widetilde{\boldsymbol{y}}\| \,. \tag{46}$$

where the last inequality we have used that $\boldsymbol{y} \to \frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y}$ is strongly convex with strong convexity parameter $\sigma_{\max}^{-1}$ and $\boldsymbol{H}^{-1}\boldsymbol{y}$ is the gradient of this function.

Using now the equivalence between $\nabla g_\gamma^*(\boldsymbol{H}^{-1}\cdot)$ of Eq. (41) and the proximal operator we finally have the claimed bound:

$$\|\boldsymbol{y} - \widetilde{\boldsymbol{y}}\| \leq (\sigma_{\max}/\sigma_{\min} + \gamma L_g\sigma_{\max})\|\nabla g_\gamma^*(\boldsymbol{H}^{-1}\boldsymbol{y}) - \nabla g_\gamma^*(\boldsymbol{H}^{-1}\widetilde{\boldsymbol{y}})\| \tag{47}$$

$$= (\sigma_{\max}/\sigma_{\min} + \gamma L_g\sigma_{\max})\|\operatorname{\mathbf{prox}}_{\gamma g}(\boldsymbol{y}) - \operatorname{\mathbf{prox}}_{\gamma g}(\widetilde{\boldsymbol{y}})\| \,. \tag{48}$$

□

Lemma 7 (Block firm non-expansiveness)**.** *Let $\boldsymbol{x}, \widetilde{\boldsymbol{x}}$ be two arbitrary vectors in $\mathbb{R}^p$, g be a block-separable convex lower semicontinuous function with blocks $\mathcal{B}$. Let $\boldsymbol{z} \overset{def}{=} \operatorname{\mathbf{prox}}_{\gamma g}(\boldsymbol{x})$, $\widetilde{\boldsymbol{z}} \overset{def}{=} \operatorname{\mathbf{prox}}_{\gamma g}(\widetilde{\boldsymbol{x}})$. Then for any subset $\mathcal{A} \subseteq \mathcal{B}$ it is verified that:*

$$\langle [\boldsymbol{z} - \widetilde{\boldsymbol{z}}]_{\mathcal{A}}, [\boldsymbol{x} - \widetilde{\boldsymbol{x}}]_{\mathcal{A}}\rangle \geq \|[\boldsymbol{z} - \widetilde{\boldsymbol{z}}]_{\mathcal{A}}\|^2 \,. \tag{49}$$

Proof. By the block-separability of g, the proximal operator is the concatenation of the proximal operators of the blocks. In other words, for any block $A \in \mathcal{A}$ we have:

$$[\boldsymbol{z}]_A = \operatorname{\mathbf{prox}}_{\gamma g_A}([\boldsymbol{x}]_A) \,, \quad [\widetilde{\boldsymbol{z}}]_A = \operatorname{\mathbf{prox}}_{\gamma g_A}([\widetilde{\boldsymbol{x}}]_A) \,, \tag{50}$$

where g_A is the restriction of g to A. By firm non-expansiveness of the proximal operator (see e.g. Bauschke and Combettes (2017, Proposition 4.2)) we have that:

$$\langle [\boldsymbol{z}]_A - [\widetilde{\boldsymbol{z}}]_A, [\boldsymbol{x}]_A - [\widetilde{\boldsymbol{x}}]_A\rangle \geq \|[\boldsymbol{z}]_A - [\widetilde{\boldsymbol{z}}]_A\|^2 \,.$$

Summing over the blocks in $\mathcal{A}$ yields the desired result. □

Appendix B Fixed point characterization

In this subsection we provide a characterization of the fixed points of the three operator splitting.

The following theorem characterizes the set of fixed points $\boldsymbol{G}_\gamma$ (defined in (5)) as the weighted Minkowski sum of primal and dual solutions. We will denote by $\mathrm{Fix}(\boldsymbol{G}_\gamma)$ the set of fixed points of $\boldsymbol{G}_\gamma$. This characterization seems to be new, and it will be used in some of the later proofs.

Theorem 4 (Fixed point for operator splitting)**.** *Let $\mathcal{P}^\star$ denote the set of minimizers of the primal objective and $\mathcal{D}^\star$ the set of minimizers of the dual objective. Then the set of fixed points of the three splitting is*

$$\mathrm{Fix}(\boldsymbol{G}_\gamma) = \mathcal{P}^\star + \gamma\, \boldsymbol{D}\mathcal{D}^\star = \{\boldsymbol{x} + \gamma \boldsymbol{D}\boldsymbol{u} \,|\, \boldsymbol{x} \in \mathcal{P}^\star, \boldsymbol{u} \in \mathcal{D}^\star\} \ . \tag{51}$$

Proof. We first characterize the fixed points of $\boldsymbol{G}_\gamma$ by a subdifferential inclusion. Given $\boldsymbol{y} \in \mathbb{R}^p$, let $\boldsymbol{z} \stackrel{\text{def}}{=} \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y})$ and $\boldsymbol{u} \stackrel{\text{def}}{=} \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{z})/\gamma$. Consider the following sequence of equivalences:

$$\boldsymbol{y} = \boldsymbol{G}_\gamma(\boldsymbol{y}) \iff \begin{cases} \boldsymbol{z} = \mathbf{prox}_{\gamma g}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z} - \boldsymbol{y} - \gamma \boldsymbol{D}\nabla f(\boldsymbol{z})) \\ \boldsymbol{z} = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}) \end{cases} \quad \text{(by definition of } \boldsymbol{G}_\gamma) \tag{52}$$

$$\iff \begin{cases} \boldsymbol{D}^{-1}(-\frac{\gamma}{\gamma}(\boldsymbol{y} - \boldsymbol{z}) - \gamma \boldsymbol{D}\nabla f(\boldsymbol{z}))/\gamma \in \partial g(\boldsymbol{z}) \\ \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{z})/\gamma \in \partial h(\boldsymbol{z}) \end{cases} \quad \text{(by Lemma 1)} \tag{53}$$

$$\iff \begin{cases} -\boldsymbol{u} \in \partial(f + g)(\boldsymbol{z}) \\ \boldsymbol{u} \in \partial h(\boldsymbol{z}) \end{cases} \tag{54}$$

$$\iff \begin{cases} \boldsymbol{z} \in \partial(f + g)^*(-\boldsymbol{u}) \\ \boldsymbol{z} \in \partial h^*(\boldsymbol{u}) \end{cases} \quad \text{(by Lemma 2)} \tag{55}$$

The rest of the proof is divided in two parts, proving in the first part that $\mathrm{Fix}(\boldsymbol{G}_\gamma) \subseteq \mathcal{P}^\star + \gamma \boldsymbol{D}\mathcal{D}^\star$, and the reverse inclusion in the second part.

Part 1. Our goal is to prove $\mathrm{Fix}(\boldsymbol{G}_\gamma) \subseteq \mathcal{P}^\star + \gamma\mathcal{D}^\star$. Let $\boldsymbol{y} \in \mathrm{Fix}(\boldsymbol{G}_\gamma)$ and $\boldsymbol{z}$, $\boldsymbol{u}$ be as defined above. From their definition we immediately have $\boldsymbol{z} + \gamma \boldsymbol{D}\boldsymbol{u} = \boldsymbol{y}$, and so we only need to prove that $\boldsymbol{z}, \boldsymbol{u}$ are minimizers of the primal and dual objective respectively. By definition of $\boldsymbol{z}$ we have the following subdifferential inclusions

$$\boldsymbol{z} = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}) \iff \frac{\boldsymbol{D}^{-1}}{\gamma}(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{u} \in \partial h(\boldsymbol{z}) \tag{56}$$

$$\iff \boldsymbol{z} \in \partial h^*(\boldsymbol{u}) \ , \tag{57}$$

where we have used Lemma 1 for the first equivalence and Lemma 2 for the second one. Adding together (56) with the first line of (54), and (57) minus the first line of (55) gives

$$0 \in \partial h(\boldsymbol{z}) + \partial g(\boldsymbol{z}) + \nabla f(\boldsymbol{z}) \tag{58}$$

$$\text{and} \quad 0 \in \partial h^*(\boldsymbol{u}) - \partial(f + g)^*(-\boldsymbol{u}) \ , \tag{59}$$

and so by the first-order optimality conditions $\boldsymbol{z}$ and $\boldsymbol{u}$ are minimizers of the primal and dual objectives respectively. We have proved $\mathrm{Fix}(\boldsymbol{G}_\gamma) \subseteq \mathcal{P}^\star + \gamma \boldsymbol{D}\mathcal{D}^\star$.

Part 2. Our goal now is to prove the inverse inclusion, $\mathcal{P}^\star + \gamma \boldsymbol{D}\mathcal{D}^\star \subseteq \mathrm{Fix}(\boldsymbol{G}_\gamma)$. Let $(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{P}^\star \times \mathcal{D}^\star$, we will prove that $\boldsymbol{y} \stackrel{\text{def}}{=} \boldsymbol{x} + \gamma \boldsymbol{D}\boldsymbol{u}$ is a fixed point of $\mathrm{Fix}(\boldsymbol{G}_\gamma)$.

We start by recalling the notion of *paramonotinicity*, which will play a key role in this part of the proof. This notion was introduced by Iusem (1998) and is key to characterizing the set of fixed points of related methods, such as the Douglas-Rachford splitting (Bauschke et al., 2012). An operator $\boldsymbol{C}$ is said to be paramonotonic if the following implication is verified

$$\left.\begin{array}{l} \boldsymbol{a}^\star \in \boldsymbol{C}\boldsymbol{a} \\ \boldsymbol{b}^\star \in \boldsymbol{C}\boldsymbol{b} \\ \langle \boldsymbol{a}^\star - \boldsymbol{b}^\star, \boldsymbol{a} - \boldsymbol{b} \rangle = 0 \end{array}\right\} \implies \boldsymbol{a}^\star \in \boldsymbol{C}\boldsymbol{b} \text{ and } \boldsymbol{b}^\star \in \boldsymbol{C}\boldsymbol{a} \ . \tag{60}$$

The usefulness of this notion in this case comes from the fact that the subdifferential of a convex proper lower semicontinuous function is paramonotonic (Iusem, 1998, Proposition 2.2). Hence we have that ∂h and $\partial(f+g)$ are paramonotonic.

By the first-order optimality conditions on the primal and dual loss we have that there exists elements $\boldsymbol{u}_{\boldsymbol{z}}$ and $\boldsymbol{z}_{\boldsymbol{u}}$ such that

$$\boldsymbol{u}_{\boldsymbol{z}} \in \partial h(\boldsymbol{z}) \cap (-\partial(f+g)(\boldsymbol{z})) \tag{61}$$
$$\boldsymbol{z}_{\boldsymbol{u}} \in \partial h^*(\boldsymbol{u}) \cap (\partial(f+g)^*(-\boldsymbol{u})) \ , \tag{62}$$

where the second inclusion can be written equivalently using the conjugate-inverse identity (Lemma 2) as

$$\boldsymbol{u} \in \partial h(\boldsymbol{z}_{\boldsymbol{u}}) \cap (-\partial(f+g)(\boldsymbol{z}_{\boldsymbol{u}})) \ . \tag{63}$$

Using Eq. (61) and (63) we have by monotony of ∂h and $\partial(f+g)$

$$\langle \boldsymbol{u}_{\boldsymbol{z}} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z}_{\boldsymbol{u}} \rangle \geq 0 \ \text{ and } \langle \boldsymbol{u}_{\boldsymbol{z}} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z}_{\boldsymbol{u}} \rangle \leq 0 \tag{64}$$

from where we necessarily have $\langle \boldsymbol{u}_{\boldsymbol{z}} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z}_{\boldsymbol{u}} \rangle = 0$. We hence have by paramonotonicity of ∂h

$$\left.\begin{aligned} &\boldsymbol{u}_{\boldsymbol{z}} \in \partial h(\boldsymbol{z}) \\ &\boldsymbol{u} \in \partial h(\boldsymbol{z}_{\boldsymbol{u}}) \\ &\langle \boldsymbol{u}_{\boldsymbol{z}} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z}_{\boldsymbol{u}} \rangle = 0 \end{aligned}\right\} \implies \boldsymbol{u} \in \partial h(\boldsymbol{z}) \tag{65}$$

Similarly, by paramonotonicity of $\partial(f+g)$ we have

$$\left.\begin{aligned} &-\boldsymbol{u}_{\boldsymbol{z}} \in \partial(f+g)(\boldsymbol{z}) \\ &-\boldsymbol{u} \in \partial(f+g)(\boldsymbol{z}_{\boldsymbol{u}}) \\ &\langle \boldsymbol{u}_{\boldsymbol{z}} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z}_{\boldsymbol{u}} \rangle = 0 \end{aligned}\right\} \implies -\boldsymbol{u} \in \partial(f+g)(\boldsymbol{z}) \tag{66}$$

Combining the last two equations we have by the definition of $\boldsymbol{y}$ the following inclusions

$$\begin{cases} -\boldsymbol{u} \in \partial(f+g)(\boldsymbol{z}) \\ \boldsymbol{u} \in \partial h(\boldsymbol{z}) \end{cases} \tag{67}$$

which by Eq. (54) implies that $\boldsymbol{y} \in \mathrm{Fix}(\boldsymbol{G}_\gamma)$ (note that these are all equivalences from (52) to (54)). This concludes the proof. □

Corollary 1 (Minimizer of our objective)**.** *Let $\boldsymbol{y} \in \mathrm{Fix}(\boldsymbol{G}_\gamma)$. Then we have that $\boldsymbol{z} = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y})$ is a minimizer of the primal objective $\mathcal{P}$, $\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}) = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z} - \boldsymbol{y} - \gamma \boldsymbol{D}\nabla f(\boldsymbol{z}))$, and $\boldsymbol{u} = \boldsymbol{D}^{-1}(\boldsymbol{y} - \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}))/\gamma$ is a minimizer of the dual objective.*

Proof. This follows from the first part of the proof of Theorem 4. In particular, Eq. (58) shows that $\boldsymbol{z} = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y})$ is a minimizer of the primal objective, while Eq. (59) shows that $\boldsymbol{u} = \boldsymbol{D}^{-1}(\boldsymbol{y} - \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}))/\gamma$ is a minimizer of the dual objective.

The identity $\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}) = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z} - \boldsymbol{y} - \gamma \boldsymbol{D}\nabla f(\boldsymbol{z}))$ comes from the definition of fixed point (52). □

Appendix C Iteration complexity analysis

In this section we provide a proof for the convergence rate analysis of the proposed methods of §4. We will start by with the proof of linear convergence (Theorem 3) and then prove the sublinear convergence rate (Theorem 1), as this last theorem reuses many elements from the first.

Unless explicitly stated (e.g., in Theorem 1), the results are only proven for the sparse variants. Since the dense variants are a special case of the sparse variants with $\boldsymbol{P}_i = \boldsymbol{I}$, $\boldsymbol{D} = \boldsymbol{I}$, the results for the dense variants follow as a special case.

Structure of this appendix.

- Appendix C.1 provides technical lemmas that will be used in later proofs.
- Appendix C.2 provides a proof for the linear convergence (under assumptions) of the proposed methods (Theorem 3).
- Appendix C.3 provides a sublinear convergence rate for the dense variants of the proposed methods (Theorem 1).
- Appendix C.4 provides a (weaker) sublinear convergence rate for the sparse variants of the proposed methods (Theorem 2).

Extra notation for this section.

- We define $f_i(\boldsymbol{x}) = \psi_i(\boldsymbol{x}) + \omega(\boldsymbol{x})$
- To provide a unified analysis of the dense and sparse algorithm, we define the following auxiliary function:

$$\xi_i(\boldsymbol{x}) \stackrel{\text{def}}{=} \psi_i(\boldsymbol{x}) + \sum_{B \in T_i} d_B \omega_B([\boldsymbol{x}]_B) \ . \tag{68}$$

  Note that $\frac{1}{n}\sum_{i=1}^n \xi_i = f$. Since ψ_i is L_ψ-smooth and ω is L_ω-smooth we have that ξ_i is L_f-smooth, with $L_f = L_\psi + d_{\max} L_\omega$ (as defined in §4).

- Contrary to full gradient algorithms, in stochastic variance reduced methods the objective function is not guaranteed to decrease at each iteration. To compensate for this, a common approach is to add a positive term that decreases throughout the iterations. The resulting function is often called a *Lyapunov* function. Throughout this paper, the positive term that we will add is the following:

$$H_t \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n m_{i,t} \tag{69}$$

  where $m_{i,t}$ are positive constants initialized as

$$m_{i,0} = \frac{1}{2L_f} \|\boldsymbol{\alpha}_{i,0} - \nabla \psi_i(\boldsymbol{x}^\star)\|^2 \tag{70}$$

  and updated at each iteration as

$$m_{i,t+1} = \begin{cases} B_{f_i}(\boldsymbol{z}_t, \boldsymbol{x}^\star) & \text{if } \boldsymbol{\alpha}_i \text{ has been updated} \\ m_{i,t} & \text{otherwise} \end{cases} , \tag{71}$$

  for all $i \in \{1, \dots, n\}$. This term is a hybrid between those used by Defazio et al. (2014) and Hofmann et al. (2015). Like Defazio et al. (2014), it will allow us to obtain a large $1/(3L_f)$ step size, contrary to the $< 1/4L_f$ step size of Hofmann et al. (2015). Like Hofmann et al. (2015) (and unlike Defazio et al. (2014)), it will allow to initialize $\boldsymbol{\alpha}_0$ arbitrarily.

- For convenience, we denote by $\langle \cdot, \cdot \rangle_{(i)}$ (resp. $\| \cdot \|_{(i)}$) the scalar product (resp. norm) restricted to blocks in the extended support, i.e., $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{(i)} \stackrel{\text{def}}{=} \langle \boldsymbol{x}, \boldsymbol{P}_i \boldsymbol{y} \rangle$ and $\|\boldsymbol{x}\|_{(i)} \stackrel{\text{def}}{=} \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{(i)}}$.
- We denote by $d_{\max}$ the maximum entry in the diagonal matrix $\boldsymbol{D}$, with $\boldsymbol{D}$ as defined in §2.1.

Appendix C.1 Preliminaries

In this subsection we state some key lemmas that are used in both the proof of linear and sublinear convergence.

Lemma 8 (Strong convexity inequality)**.** *Let ψ_i be μ_ψ-strongly convex. Let ω be μ_ω-strongly convex (where we allow $\mu_\psi = \mu_\omega = 0$). Then with $\mu = \mu_\psi + \mu_\omega$ we have the following inequality for arbitrary $\boldsymbol{x}$ and $\boldsymbol{y}$ in the domain:*

$$\begin{aligned} f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{1}{2n(L_f - \mu)} \sum_{i=1}^n \|\nabla \xi_i(\boldsymbol{x}) - \nabla \xi_i(\boldsymbol{y})\|^2 \\ + \frac{\mu L_f}{2d_{max}(L_f - \mu)} \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{\mu}{L_f - \mu} \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{y} - \boldsymbol{x} \rangle \end{aligned} \tag{72}$$

Proof. We start by proving that ξ_i is μ-strongly convex when restricted to his support. Let $\boldsymbol{a}, \boldsymbol{b}$ be arbitrary vectors in $\mathbb{R}^p$. Then we have the following sequence of inequalities:

$$\begin{aligned} &\langle \nabla \xi_i(\boldsymbol{a}) - \nabla \xi_i(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle && (73) \\ &= \langle \nabla \psi_i(\boldsymbol{a}) + \boldsymbol{P}_i \boldsymbol{D} \nabla \omega(\boldsymbol{a}) - \nabla \psi_i(\boldsymbol{b}) - \boldsymbol{P}_i \boldsymbol{D} \nabla \omega(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle && (74) \\ &\quad \text{(by definition of } \nabla \xi_i) \\ &= \langle \nabla \psi_i(\boldsymbol{a}) - \nabla \psi_i(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \langle \boldsymbol{P}_i \boldsymbol{D} \nabla \omega(\boldsymbol{a}) - \boldsymbol{P}_i \boldsymbol{D} \nabla \omega(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle && (75) \\ &\geq \mu_\psi \|\boldsymbol{a} - \boldsymbol{b}\|^2 + \langle \boldsymbol{P}_i \boldsymbol{D} \nabla \omega(\boldsymbol{a}) - \boldsymbol{P}_i \boldsymbol{D} \nabla \omega(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle && (76) \\ &\quad \text{(by strong convexity of } \nabla \psi_i) \\ &= \mu_\psi \|\boldsymbol{a} - \boldsymbol{b}\|^2 + \textstyle\sum_{B \in \mathcal{B}} d_B \langle \nabla \omega_B([\boldsymbol{a}]_B) - \nabla \omega_B([\boldsymbol{b}]_B), [\boldsymbol{a}]_B - [\boldsymbol{b}]_B \rangle && (77) \\ &\quad \text{(by block separability of } \omega \text{ and definition of } \boldsymbol{P}_i \boldsymbol{D}) \\ &\geq \mu_\psi \|\boldsymbol{a} - \boldsymbol{b}\| + \textstyle\sum_{B \in \mathcal{B}} d_B \mu_\omega \|[\boldsymbol{a}]_B - [\boldsymbol{b}]_B\|^2 && (78) \\ &\quad \text{(strong convexity of } \omega_B \text{, consequence of strong cvx of } \omega) \\ &\geq \mu_\psi \|\boldsymbol{a} - \boldsymbol{b}\|^2 + \mu_\omega \|\boldsymbol{a} - \boldsymbol{b}\|_{(i)}^2 && (79) \\ &\quad \text{(using } d_B \geq 1 \text{ by definition)} \\ &\geq \underbrace{(\mu_\psi + \mu_\omega)}_{=\mu} \|\boldsymbol{a} - \boldsymbol{b}\|_{(i)}^2 \ . && (80) \end{aligned}$$

We have proved that ξ_i is μ-strongly convex on the subspace generated by the extended support (i.e., with respect to the norm $\|\cdot\|_{(i)}$). Since it is also L_f-smooth by (68), we can apply (Defazio et al., 2014, Lemma 4) to obtain the following inequality, valid for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in its domain:

$$\begin{aligned} \xi_i(\boldsymbol{a}) \geq \xi_i(\boldsymbol{b}) + \langle \nabla \xi_i(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \frac{1}{2(L_f - \mu)} \|\nabla \xi_i(\boldsymbol{a}) - \nabla \xi_i(\boldsymbol{b})\|^2 \\ + \frac{\mu L_f}{2(L_f - \mu)} \|\boldsymbol{a} - \boldsymbol{b}\|_{(i)}^2 + \frac{\mu}{L_f - \mu} \langle \nabla \xi_i(\boldsymbol{a}) - \nabla \xi_i(\boldsymbol{b}), \boldsymbol{b} - \boldsymbol{a} \rangle \end{aligned} \tag{81}$$

We will apply the previous inequality at $\boldsymbol{a} = \boldsymbol{x}$, $\boldsymbol{b} = \boldsymbol{y}$ and average over all i. Note that $\frac{1}{n} \sum_{i=1}^n \xi_i(\boldsymbol{x}) = \psi(\boldsymbol{x}) + \omega(\boldsymbol{x}) = f(\boldsymbol{x})$ by definition of ξ_i and so we can write

$$\begin{aligned} f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \frac{1}{n} \sum_{i=1}^n \langle \nabla \xi_i(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{1}{2n(L_f - \mu)} \sum_{i=1}^n \|\nabla \xi_i(\boldsymbol{x}) - \nabla \xi_i(\boldsymbol{y})\|^2 \\ + \frac{\mu L_f}{2(L_f - \mu)} \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{y} - \boldsymbol{x}\|_{(i)}^2 + \frac{\mu}{L_f - \mu} \frac{1}{n} \sum_{i=1}^n \langle \nabla \xi_i(\boldsymbol{x}) - \nabla \xi_i(\boldsymbol{y}), \boldsymbol{y} - \boldsymbol{x} \rangle \end{aligned} \tag{82}$$

We can simplify the terms in this inequality as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\langle\nabla\xi_i(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle = \langle\frac{1}{n}\sum_{i=1}^{n}\nabla\xi_i(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle = \langle\nabla f(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle \tag{83}$$

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}-\boldsymbol{y}\|_{(i)}^2 = \|\boldsymbol{x}-\boldsymbol{y}\|_{\boldsymbol{D}^{-1}}^2 \geq \frac{1}{d_{\max}}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \tag{84}$$

$$\frac{1}{n}\sum_{i=1}^{n}\langle\nabla\xi_i(\boldsymbol{x})-\nabla\xi_i(\boldsymbol{y}),\boldsymbol{y}-\boldsymbol{x}\rangle = \langle\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y}),\boldsymbol{y}-\boldsymbol{x}\rangle \tag{85}$$

The second equality results by the definition of $\boldsymbol{D}$ which gives: $\mathbf{E}[\boldsymbol{P}_i] = \boldsymbol{D}^{-1}$.Using the previous identities (and inequality) into (82) we finally obtain the desired bound:

$$\begin{aligned} f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \langle\nabla f(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle + \frac{1}{2n(L_f-\mu)}\sum_{i=1}^{n}\|\nabla\xi_i(\boldsymbol{x})-\nabla\xi_i(\boldsymbol{y})\|^2 \\ + \frac{\mu L_f}{2d_{\max}(L_f-\mu)}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \frac{\mu}{L_f-\mu}\langle\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y}),\boldsymbol{y}-\boldsymbol{x}\rangle \end{aligned} \tag{86}$$

□

Lemma 9 (Bound on gradient estimate variance)**.** *Let* $\mathbf{E}$ *denote the conditional expectation with respect to the random index* i *selected at the* t*-th iteration. Then we have the following inequality:*

$$\begin{aligned} \mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2 \leq (1+\beta^{-1})2L_f H_t + (1+\beta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t)-\nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ - \beta\|\nabla f(\boldsymbol{z}_t)-\nabla f(\boldsymbol{x}^\star)\|^2 \ , \end{aligned} \tag{87}$$

valid for any $\beta > 0$.

Proof. Let $\psi \overset{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\psi_j$. Then we have the following sequence of inequalities:

$$\mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2 = \mathbf{E}\|\underbrace{\nabla\xi_i(\boldsymbol{z}_t) - \boldsymbol{\alpha}_{i,t} + \boldsymbol{D}_i\overline{\boldsymbol{\alpha}}_t - \boldsymbol{D}_i\nabla f(\boldsymbol{x}^\star)}_{=\boldsymbol{\zeta}_i}\|^2 \tag{88}$$

$$\begin{aligned} &= \mathbf{E}\|\overbrace{\big[-\boldsymbol{\alpha}_{i,t} + \boldsymbol{D}_i\overline{\boldsymbol{\alpha}}_t + \nabla\xi_i(\boldsymbol{x}^\star) - \boldsymbol{D}_i\nabla f(\boldsymbol{x}^\star)\big] + \big[\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\big]}^{\boldsymbol{\zeta}_i} - \overbrace{\big(\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\big)}^{\mathbf{E}\boldsymbol{\zeta}_i}\big]\|^2 \\ &\quad + \|\overbrace{\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)}^{\mathbf{E}\boldsymbol{\zeta}_i}\|^2 \end{aligned} \tag{89}$$

(by Lemma 3 with $\boldsymbol{Q}_i = \boldsymbol{I}$, $\boldsymbol{A} = \boldsymbol{I}$, and where we have also added and substracted $\nabla\xi_i(\boldsymbol{x}^\star)$)

$$\begin{aligned} &\leq (1+\beta^{-1})\mathbf{E}\|\boldsymbol{\alpha}_{i,t} - \psi_i(\boldsymbol{x}^\star) - \boldsymbol{D}_i\overline{\boldsymbol{\alpha}}_t + \boldsymbol{D}_i\nabla\psi(\boldsymbol{x}^\star)\|^2 \\ &\quad + (1+\beta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star) - \nabla f(\boldsymbol{z}_t) + \nabla f(\boldsymbol{x}^\star)\|^2 + \|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2 \end{aligned} \tag{90}$$

(by Young's inequality $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq (1+\beta^{-1})\|\boldsymbol{a}\|^2 + (1+\beta)\|\boldsymbol{b}\|^2$ and definition of ξ_i)

$$\begin{aligned} &= (1+\beta^{-1})\mathbf{E}\|\boldsymbol{\alpha}_{i,t} - \psi_i(\boldsymbol{x}^\star)\|^2 - (1+\beta^{-1})\mathbf{E}\|\overline{\boldsymbol{\alpha}}_t - \nabla\psi(\boldsymbol{x}^\star)\|_{\boldsymbol{D}}^2 \\ &\quad + (1+\beta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 - \beta\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2 \ , \end{aligned} \tag{91}$$

where in the last equivalence we have applied Lemma 3 both to the first term (with $\boldsymbol{Q}_i = \boldsymbol{P}_i$, $\boldsymbol{A} = \boldsymbol{D}$) and to the second term (this time with $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{Q}_i = \boldsymbol{I}$). In all, and dropping the negative second term we have the inequality

$$\begin{aligned} \mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2 \leq (1+\beta^{-1})\mathbf{E}\|\boldsymbol{\alpha}_{i,t} - \psi_i(\boldsymbol{x}^\star)\|^2 + (1+\beta)\mathbf{E}\|\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ - \beta\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2 \end{aligned} \tag{92}$$

We will now bound the first term of the above inequality. Let $\mathcal{J}$ denote the set of indices for which the memory terms have been updated at least once and $\mathcal{J}^c$ its complement. For $j \in \mathcal{J}$, we denote by $\boldsymbol{\phi}_{j,t}$ the iterate at which

$\boldsymbol{\alpha}_j$ was last updated, i.e., $\boldsymbol{\alpha}_{j,t} = \nabla\psi(\boldsymbol{\phi}_{j,t})$ for all j and t. Then we have

$$\mathbf{E}\|\boldsymbol{\alpha}_{i,t} - \nabla\psi_i(\boldsymbol{x}^\star)\|^2 = \frac{1}{n}\left(\sum_{j\in\mathcal{J}}\|\boldsymbol{\alpha}_{i,t} - \nabla\psi_i(\boldsymbol{x}^\star)\|^2 + 2L_f\sum_{j\in\mathcal{J}^c}\xi_{t,j}\right) \tag{93}$$

$$\leq \frac{2L_f}{n}\left(\sum_{j\in\mathcal{J}} B_{\psi_i}(\boldsymbol{\phi}_{i,k}, \boldsymbol{x}^\star) + \sum_{j\in\mathcal{J}^c}\xi_{t,j}\right) \tag{94}$$

(by Lemma 4 and also using $L_f \geq L_\psi$)

$$\leq \frac{2L_f}{n}\left(\sum_{j\in\mathcal{J}} B_{f_i}(\boldsymbol{\phi}_{i,k}, \boldsymbol{x}^\star) + \sum_{j\in\mathcal{J}^c}\xi_{t,j}\right) \tag{95}$$

(adding B_ω, which is positive by convexity of ω)

$$= 2L_f H_t\ . \tag{96}$$

Finally, plugging this bound back in (92) we obtain the desired inequality:

$$\begin{aligned}\mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2 \leq (1+\beta^{-1})2L_f H_t + (1+\beta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ - \beta\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2\end{aligned} \tag{97}$$

□

Lemma 10 (Evolution of H_t)**.** *Let* $\mathbf{E}$ *denote the conditional expectation with respect to the random index* i *selected at the* t*-th iteration. Then for every iteration* $t \geq 0$ *we have (with* $q = 1$ *for* Saga *variants):*

$$\mathbf{E}H_{t+1} = \frac{q}{n}B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) + \left(1 - \frac{q}{n}\right)H_t\ . \tag{98}$$

Proof. By definition of $m_{j,t+1}$ in Eq. (68), for a fixed index j we have:

$$\mathbf{E}[m_{j,t+1}] = \frac{q}{n}B_{f_j}(\boldsymbol{z}_t, \boldsymbol{x}^*) + (1 - \frac{q}{n})m_{j,t}\ . \tag{99}$$

Hence averaging over all indices we get

$$\begin{aligned}\mathbf{E}[H_{t+1}] &= \frac{1}{n}\sum_{j=1}^n \mathbf{E}[m_{j,t+1}] \\ &= \frac{1}{n}\sum_{j=1}^n\left(\frac{q}{n}B_{f_j}(\boldsymbol{z}_t, \boldsymbol{x}^*) + (1 - \frac{q}{n})m_{j,t}\right) \\ &= \frac{q}{n}B_f(\boldsymbol{z}_t, \boldsymbol{x}^*) + (1 - \frac{q}{n})H_t\end{aligned} \tag{100}$$

□

Appendix C.2 Linear convergence: proof of Theorem 3

The **proof is structured** as follows:

- We start by proving an inequality that relates $\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2$ with $\|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2$, where $\boldsymbol{y}^\star$ is a fixed point of $\boldsymbol{G}_\gamma$. This inequality will be central in both proofs of linear and sublinear convergence. We call this the "master recurrence inequality" (Lemma 11),
- As is often the case in variance reduced methods, a recurrence purely in terms of the iterates as the one in Lemma 11 does not provide the monotonic decrease required to prove a linear convergence rate. To overcome this, we will make use of an auxiliary function which is always larger than the suboptimality criterion and which *does* verify a monotonic decrease in expectation. This is often referred to as a *Lyapunov* function. The Lyapunov function that we will use is the following:

$$V_t \stackrel{\text{def}}{=} c\|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + H_t \quad , \tag{101}$$

  with H_t as defined in (69) and $\boldsymbol{y}^\star$ an arbitrary fixed point of $\boldsymbol{G}_\gamma$.
- Finally, in Theorem 3 we use the decrease of the Lyapunov function prove the desired rates of convergence.

Lemma 11 (Master recurrence inequality)**.** *Let $\{\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t\}$ be the iterates produced by any of the proposed algorithms, $\boldsymbol{y}^\star \in \operatorname{Fix}(\boldsymbol{G}_\gamma)$ and $\boldsymbol{x}^\star \stackrel{def}{=} \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}^\star)$ (with $\boldsymbol{D} = \boldsymbol{I}$ for the dense variants). Then we have the following inequality, valid for all $\beta > 0$ and $s > 0$:*

$$\begin{aligned}
\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 &+ (s-1)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 \\
&+ \frac{\gamma^2}{s}(1+\beta^{-1})2L_f H_t \\
&+ (\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f})\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\
&+ (-2\frac{\gamma^2\beta}{s}\mu - 2\gamma\frac{L_f - \mu}{L_f})B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) \\
&- \frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2
\end{aligned} \tag{102}$$

Proof. Developing the square we have

$$\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 = \mathbf{E}\|\boldsymbol{y}_t + (\boldsymbol{y}_{t+1} - \boldsymbol{y}_t) - \boldsymbol{y}^\star\|^2 \tag{103}$$

$$= \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + \mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 2\mathbf{E}\langle \boldsymbol{y}_{t+1} - \boldsymbol{y}_t, \boldsymbol{y}_t - \boldsymbol{y}^\star\rangle \ . \tag{104}$$

We will now work towards bounding the last term of this expression.

Let i denote the random index selected at iteration t. Note that by definition of $\boldsymbol{D}$ we have $\mathbf{E}[\boldsymbol{P}_i] = \boldsymbol{D}^{-1}$ and so we can write:

$$\mathbf{E}\langle \boldsymbol{z}_t - \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle = \langle \boldsymbol{z}_t - \boldsymbol{x}^\star, \boldsymbol{y}_t - \boldsymbol{y}^\star\rangle_{\boldsymbol{D}^{-1}} \tag{105}$$

$$\geq \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{D}^{-1}} \ , \tag{106}$$

where in the last inequality we have used Eq. (31) with $\boldsymbol{z}_t = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}_t)$ and $\boldsymbol{x}^\star = \mathbf{prox}_{\gamma g}^{\boldsymbol{D}}(\boldsymbol{y}^\star)$. Using once again the identity $\mathbf{E}[\boldsymbol{P}_i] = \boldsymbol{D}^{-1}$ and noting that $\boldsymbol{z}_t$ does not depend on i we have $\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{D}^{-1}} = \mathbf{E}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i}$ and so in all, we have

$$\mathbf{E}\langle \boldsymbol{z}_t - \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle \geq \mathbf{E}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} \ . \tag{107}$$

Furthermore, by the blockwise version of the firm non-expansiveness of the prox (Lemma 7), from the definition of $\boldsymbol{x}_t$ in Vr-Tos we also have the following inequality, with $\boldsymbol{x}_t = \mathbf{prox}_{\gamma g}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t)$ and $\boldsymbol{x}^\star = \mathbf{prox}^{\boldsymbol{D}^{-1}}(2\boldsymbol{x}^\star - \boldsymbol{y}^\star - \gamma\boldsymbol{D}\nabla f(\boldsymbol{x}^\star))$, where this last equality is a consequence of Cololary 1:

$$\langle 2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t - 2\boldsymbol{x}^\star + \boldsymbol{y}^\star + \gamma\boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{P}_i(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle - \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} \geq 0 \ , \tag{108}$$

which taking conditional expectation gives

$$\mathbf{E}\langle 2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t - 2\boldsymbol{x}^\star + \boldsymbol{y}^\star + \gamma\boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{P}_i(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle - \mathbf{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} \geq 0 \ . \tag{109}$$

We now have the following sequence of inequalities:

$$\begin{aligned}
&\mathbf{E}\langle \boldsymbol{y}_{t+1} - \boldsymbol{y}_t, \boldsymbol{y}_t - \boldsymbol{y}^\star\rangle && (110)\\
&\quad = \mathbf{E}\langle \boldsymbol{x}_t - \boldsymbol{z}_t, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle && (111)\\
&\qquad \text{(definition of } \boldsymbol{y}_{t+1}\text{)}\\
&\quad = \mathbf{E}\langle \boldsymbol{x}_t - \boldsymbol{z}_t - \boldsymbol{x}^\star + \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle && (112)\\
&\qquad \text{(adding and substracting } \boldsymbol{x}^\star\text{)}\\
&\quad = \langle \boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle - \mathbf{E}\langle \boldsymbol{z}_t - \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle && (113)\\
&\quad \leq \mathbf{E}\langle \boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{y}_t - \boldsymbol{y}^\star)\rangle - \mathbf{E}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} \qquad \text{(by Eq (107))} && (114)\\
&\quad \leq \mathbf{E}\langle 2\boldsymbol{z}_t - \gamma\boldsymbol{v}_t - 2\boldsymbol{x}^\star + \gamma\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{P}_i(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle - \mathbf{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} - \mathbf{E}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} && (115)\\
&\qquad \text{(adding Eq. (109))}\\
&\quad \leq \mathbf{E}\Big[\langle 2\boldsymbol{z}_t - \gamma\boldsymbol{v}_t - 2\boldsymbol{x}^\star + \gamma\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{P}_i(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle - \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} && (116)\\
&\qquad\qquad - \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i}\Big] \qquad \text{(by linearity of expectation)}\\
&\quad \leq \mathbf{E}\Big[- \big(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i} - 2\langle \boldsymbol{z}_t - \boldsymbol{x}^\star, \boldsymbol{P}_i(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle + \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2_{\boldsymbol{P}_i}\big) && (117)\\
&\qquad\qquad + \langle -\gamma\boldsymbol{v}_t + \gamma\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{P}_i(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle\Big] \qquad \text{(reordering terms)}\\
&\quad = \mathbf{E}\left[-\|\boldsymbol{z}_t - \boldsymbol{x}^\star - (\boldsymbol{x}_t - \boldsymbol{x}^\star)\|^2_{\boldsymbol{P}_i} - \gamma\langle \boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle\right] && (118)\\
&\qquad \text{(completing the square)}\\
&\quad = \mathbf{E}\Big[- \|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2_{\boldsymbol{P}_i} - \langle \gamma\boldsymbol{v}_t - \gamma\boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{x}_t - \boldsymbol{z}_t\rangle\\
&\qquad\qquad - \langle \gamma\boldsymbol{v}_t - \gamma\boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle\Big] \qquad \text{(adding and substracting } \boldsymbol{z}_t\text{)}\\
&\quad \leq \Big(\frac{s}{2} - 1\Big)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + \frac{\gamma^2}{2s}\mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2\\
&\qquad\qquad - \gamma\langle \nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle \ , && (119)
\end{aligned}$$

where in the last inequality we have used Young's inequality: $|\langle \boldsymbol{a}, \boldsymbol{b}]\rangle| \leq \frac{s}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2s}\|\boldsymbol{b}\|^2$ and definition of $\boldsymbol{y}_{t+1}$ for the first term and computed the expectation in the last term.

Replacing this last inequality into (104) we obtain

$$\begin{aligned}
\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + (s-1)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2\\
+ \frac{\gamma^2}{s}\mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2 - 2\gamma\langle \nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle
\end{aligned} \tag{120}$$

We will proceed to further bound the second and last terms using previous results. For the second term, we can use the bound $\mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{P}_i\boldsymbol{P}_i\boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2 \leq (1+\beta^{-1})2L_f H_t + (1+\beta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 - \beta\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2$ from Lemma 9, giving:

$$\begin{aligned}
\frac{\gamma^2}{2s}\mathbf{E}\|\boldsymbol{v}_t - \boldsymbol{D}\nabla f(\boldsymbol{x}^\star)\|^2_{(i)} \leq \frac{\gamma^2}{s}(1+\beta^{-1})L_f H_t + \frac{\gamma^2}{2s}(1+\beta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2\\
- \frac{\gamma^2\beta}{2s}\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2 \ ,
\end{aligned} \tag{121}$$

The third term can be bounded using the strong convexity inequality of Lemma 8 with $\boldsymbol{y} = \boldsymbol{z}_t, \boldsymbol{x} = \boldsymbol{x}^\star$, to obtain

$$-\gamma\langle\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle = \gamma\langle\nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle + \gamma\langle\nabla f(\boldsymbol{z}_t), \boldsymbol{x}^\star - \boldsymbol{z}_t\rangle \tag{122}$$

$$\leq \gamma\langle\nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle + \gamma(\frac{L_f - \mu}{L_f}(f(\boldsymbol{x}^\star) - f(\boldsymbol{z}_t)) \\ - \frac{1}{2L_f}\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 - \frac{\mu}{2d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 - \frac{\mu}{L_f}\langle\nabla f(\boldsymbol{x}^\star), \boldsymbol{z}_t - \boldsymbol{x}^\star\rangle) \tag{123}$$

$$\leq -\gamma\frac{L_f - \mu}{L_f}B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) + \gamma\left(-\frac{1}{2L_f}\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 - \frac{\mu}{2d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2\right) \tag{124}$$

Using the bound for these two terms in (119) we have

$$\begin{aligned}\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + (s-1)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 \\ (\boldsymbol{y}_t \text{ and } \boldsymbol{y}^\star \text{ do not depend on } i) \\ + \frac{\gamma^2}{s}(1 + \beta^{-1})2L_f H_t \\ + (\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L})\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ - \frac{\gamma^2\beta}{s}\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2 \\ - 2\gamma\frac{L-\mu}{L}B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) - \frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2\end{aligned} \tag{125}$$

We now use the bound $-\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}^\star)\|^2 \leq -2\mu B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star)$ (Nesterov, 2004, Theorem 2.1.10) to obtain:

$$\begin{aligned}\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + (s-1)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 \\ + \frac{\gamma^2}{s}(1 + \beta^{-1})2L_f H_t \\ + (\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L})\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ + (-2\frac{\gamma^2\beta}{s}\mu - 2\gamma\frac{L-\mu}{L})B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) \\ - \frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2\end{aligned} \tag{126}$$

□

Lemma 12 (Lyapunov inequality)**.** *Let $\{\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t\}$ be the iterates produced by any of the proposed algorithms for $t \geq 0$, $\boldsymbol{y}^\star \in \mathrm{Fix}(\boldsymbol{G}_\gamma)$ and $\boldsymbol{x}^\star \stackrel{def}{=} \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}^\star)$. Let the Lyapunov function V_t be as defined in (101). Then we have the following inequality:*

$$\begin{aligned}\mathbf{E}V_{t+1} \leq V_t + c(s-1)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + \left(\frac{2L_f c\gamma^2}{s}(1+\beta^{-1}) - \frac{q}{n}\right)H_t \\ + c\left(\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f}\right)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ + \left(-2\frac{c\gamma^2\beta}{s}\mu - 2c\gamma\frac{L_f - \mu}{L_f} + \frac{q}{n}\right)B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) - c\frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2\end{aligned} \tag{127}$$

Proof. We will first compute the conditional expectation of the Lyapunov term, $\mathbf{E}V_{t+1}$. For the first term we

can use the bound in Lemma 11 to obtain

$$\begin{aligned} c\mathbf{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}^\star\|^2 \leq\; & c\|\boldsymbol{y}_t-\boldsymbol{y}^\star\|^2 + c(s-1)\mathbf{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}_t\|^2 \\ & + \frac{\gamma^2}{s}(1+\beta^{-1})2cL_f H_t \\ & + c(\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f})\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ & + c(-2\frac{\gamma^2\beta}{s}\mu - 2\gamma\frac{L_f-\mu}{L_f})B_f(\boldsymbol{z}_t,\boldsymbol{x}^\star) \\ & - c\frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t-\boldsymbol{x}^\star\|^2 \end{aligned} \tag{128}$$

Using Lemma 10, the second term of the Lyapunov function gives :

$$\mathbf{E}\, H_{t+1} = (1-\frac{q}{n})H_t + \frac{q}{n}B_f(\boldsymbol{z}_t,\boldsymbol{x}^\star)\ . \tag{129}$$

and so adding both inequalities we have

$$\begin{aligned} \mathbf{E}V_{t+1} \leq\; & \overbrace{c\|\boldsymbol{y}_t-\boldsymbol{y}^\star\|^2 + H_t}^{V_t} + c(s-1)\mathbf{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}_t\|^2 \\ & + (\frac{2L_f c\gamma^2}{s}(1+\beta^{-1}) - \frac{q}{n})H_t \\ & + c(\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f})\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ & + (-2c\frac{\gamma^2\beta}{s}\mu - 2c\gamma\frac{L_f-\mu}{L_f} + \frac{q}{n})B_f(\boldsymbol{z}_t,\boldsymbol{x}^\star) \\ & - c\frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t-\boldsymbol{x}^\star\|^2 \end{aligned} \tag{130}$$

which completes the proof. □

Theorem 3. *Let ψ_i be μ_ψ-strongly convex and ω be μ_ω-strongly convex, where $\mu_\psi + \mu_\omega > 0$. Furthermore, let h be L_h-smooth. Then for any step size $\gamma \leq 1/(3L_f)$, all the proposed methods converge geometrically in expectation. For $\gamma = 1/(3L_f)$, we have the following bound for Algorithm 1 ($d_{\max} = 1$ in this case) and Algorithm 2:*

$$\mathbb{E}\|\boldsymbol{z}_{t+1}-\boldsymbol{x}^\star\|^2 \leq \left(1 - \min\left\{\frac{q}{4n}, \frac{1}{3d_{\max}^3\delta^2\kappa}\right\}\right)^t D_0 \quad , \tag{131}$$

with $D_0 \overset{def}{=} d_{\max}\left[\frac{q}{2\gamma(1-\gamma\mu)n}\|\boldsymbol{y}_0-\boldsymbol{y}^\star\|^2 + H_0\right]$, $\delta = (1 + L_h/(3L_f))$, $\kappa = L_f/\mu$ and $\boldsymbol{y}^\star \in \mathrm{Fix}(\boldsymbol{G}_\gamma)$.

Proof. From the Lyapunov inequality of Lemma 12 with $s = 1$ we have the following sequence of inequalities

$$\begin{aligned}
&\mathbf{E}V_{t+1} - (1-\rho)V_t \leq \\
&\quad \rho V_t + \left(2Lc\gamma^2(1+\beta^{-1}) - \frac{q}{n}\right) H_t + c\left(\gamma^2(1+\beta) - \frac{\gamma}{L_f}\right) \mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\
&\quad + \left(-2c\gamma^2\beta\mu - 2c\gamma\frac{L_f - \mu}{L_f} + \frac{q}{n}\right) B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) - c\frac{\gamma\mu}{d_{\max}}\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 \qquad (132) \\
&\leq \rho\left(c\|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + H_t\right) + \left(\frac{2L_f c\gamma^2}{s}(1+\beta^{-1}) - \frac{q}{n}\right) H_t \\
&\quad + c(\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f})\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 + (-2\frac{c\gamma^2\beta}{s}\mu - 2c\gamma\frac{L_f-\mu}{L_f} + \frac{q}{n})B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) \\
&\qquad - c\frac{\gamma\mu}{d_{\max}^3(1+\gamma L_h)^2}\|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 \qquad (133) \\
&\qquad\quad \text{(using Lemma 6 on the last term, where we have bounded } d_{\min} \geq 1) \\
&\leq c\left[\rho - \frac{\gamma\mu}{d_{\max}^3(1+\gamma L_h)^2}\right]\|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2 + \left[\rho + 2L_f c\gamma^2(1+\beta^{-1}) - \frac{q}{n}\right] H_t \\
&\quad + c\gamma\left[\gamma(1+\beta) - \frac{1}{L_f}\right]\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\
&\qquad + \left[-2c\gamma^2\beta\mu - 2c\gamma\frac{L_f-\mu}{L_f} + \frac{q}{n}\right] B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star) \qquad (134)
\end{aligned}$$

It is worth noting that Eq. (133) is the only part of the proof in which we use the smoothness of h.

Taking the coefficients

$$c = \frac{q}{2\gamma(1-\gamma\mu)n}\ , \quad \beta = 2\ , \quad \rho = \min\left\{\frac{q}{4n}, \frac{1}{3d_{\max}\delta^2\kappa}\right\}\ , \qquad (135)$$

With $\delta = d_{\max}(1 + \frac{L_h}{3L_f})$ and $\kappa = L_f/\mu$. One can verify that all square brackets are non-positive for $\gamma \leq 1/(3L_f)$ (the coefficients are the same, except for the first square bracket, than those that appear in (Defazio et al., 2014, Theorem 1). We hence have

$$\mathbf{E}V_{t+1} \leq (1-\rho)V_t\ , \qquad (136)$$

which chaining expectations gives

$$\mathbb{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq \mathbb{E}V_{t+1} \leq (1-\rho)^{t+1}V_0\ . \qquad (137)$$

This gives a geometric convergence on $\boldsymbol{y}_t$. However, we would like to have a convergence rate in terms of the primal iterate $\boldsymbol{x}_t$.

By Theorem 4 we have that $\boldsymbol{x}^\star = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}^\star)$, and in this case the minimizer is unique because of strong convexity. Then by firm nonexpansiveness of the prox (Lemma 6) we have

$$\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|_{\boldsymbol{D}}^2 \leq \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|_{\boldsymbol{D}}^2 \qquad (138)$$

which combined with Lemma 5 and bounding $d_{\min}$ by 1 (by definition all diagonal entries in $\boldsymbol{D}$ are ≥ 1) gives

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 &\leq \frac{1}{d_{\min}}\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|_{\boldsymbol{D}}^2 \leq \frac{1}{d_{\min}}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|_{\boldsymbol{D}}^2 \qquad (139) \\
&\leq \left(\frac{d_{\max}}{d_{\min}}\right)\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq d_{\max}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \qquad (140)
\end{aligned}$$

Combining this with (137) gives the following bound in $\boldsymbol{z}_{t+1}$:

$$\mathbb{E}\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 \leq d_{\max}\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq (1-\rho)^{t+1}d_{\max}V_0\ , \qquad (141)$$

and the claimed bound follows from definition of V_0.

□

Appendix C.3 Proof of sublinear convergence rate – dense algorithms

In this section we give a proof of convergence for the dense variants of the proposed algorithms (Algorithm 1). Because we will not be considering the sparse variants, we assume $\boldsymbol{D} = \boldsymbol{I}$ without explicit mention.

Lemma 13 (Bound on gradient estimate variance, Variant 2)**.**

$$\mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|^2 \leq (1+\eta)\mathbf{E}\|\nabla f_i(\boldsymbol{z}_t) - \nabla f_i(\boldsymbol{x}^\star)\|^2 + 2(1+\eta^{-1})L_f H_t \tag{142}$$

Proof.

$$\mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|^2 \tag{143}$$

$$= \mathbf{E}\|\nabla\psi_i(\boldsymbol{z}_t) - \boldsymbol{\alpha}_{i,t} + (\overline{\boldsymbol{\alpha}}_t + \nabla\omega(\boldsymbol{z}_t)) - \nabla f(\boldsymbol{z}_t)\|^2 \tag{144}$$

$$= \mathbf{E}\|\underbrace{\nabla f_i(\boldsymbol{z}_t) - \boldsymbol{\alpha}_{i,t}}_{\xi} + \underbrace{\overline{\boldsymbol{\alpha}}_t - \nabla f(\boldsymbol{z}_t)}_{-\mathbf{E}\xi}\|^2 \tag{145}$$

(By definition of f_i)

$$\leq \mathbf{E}\|\nabla f_i(\boldsymbol{\phi}_i^t) - \nabla f_i(\boldsymbol{z}_t)\|^2 \tag{146}$$

(Applying Lemma 3 and by definition of f_i)

$$\leq \mathbf{E}\|\nabla f_i(\boldsymbol{\phi}_i^t) - \nabla f_i(\boldsymbol{x}^\star) + \nabla f_i(\boldsymbol{x}^\star) - \nabla f_i(\boldsymbol{z}_t)\|^2 \tag{147}$$

(Adding and substracting $\nabla f_i(\boldsymbol{x}^\star)$)

$$\leq (1+\eta^{-1})\mathbf{E}\|\nabla f_i(\boldsymbol{\phi}_i^t) - \nabla f_i(\boldsymbol{x}^\star)\|^2 + (1+\eta)\mathbf{E}\|\nabla f_i(\boldsymbol{z}_t) - \nabla f_i(\boldsymbol{x}^\star)\|^2 \tag{148}$$

(Applying Young's inequality)

$$\leq 2L_f(1+\eta^{-1})H_t + (1+\eta)\mathbf{E}\|\nabla f_i(\boldsymbol{z}_t) - \nabla f_i(\boldsymbol{x}^\star)\|^2 \tag{149}$$

(Applying Lemma 6 from (Defazio et al., 2014) on the first term)

□

Lemma 14 (Saddle point recursive inequality)**.** *Let $\gamma \leq 1/L$ and $\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{u}_t$ be the iterates generated by either* VR-TOS *(Algorithm 1). Then we have the following inequality for any $(\boldsymbol{x}, \boldsymbol{u}) \in \mathrm{dom}(g)\times\mathrm{dom}(h)$, with $\boldsymbol{y} = \boldsymbol{x}+\gamma\boldsymbol{u}$:*

$$\begin{aligned} &2\gamma\mathbf{E}(\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t)) + \mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}\|^2 \\ &\quad\leq \mathbf{E}\|\boldsymbol{y}_t - \boldsymbol{y}\|^2 + 2\gamma^2(1+\eta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 + 4\gamma^2(1+\eta^{-1})L_f H_t \end{aligned} \tag{150}$$

Proof. By the convexity and the L-smoothness inequality, f verifies the following inequalities for an arbitrary $\boldsymbol{x}$:

$$f(\boldsymbol{z}_t) - f(\boldsymbol{x}) \leq \langle\nabla f(\boldsymbol{z}_t), \boldsymbol{z}_t - \boldsymbol{x}\rangle \tag{151}$$

$$f(\boldsymbol{x}_t) - f(\boldsymbol{z}_t) \leq \langle\nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{z}_t\rangle + \frac{L}{2}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2 \tag{152}$$

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \leq \langle\nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2 \ , \tag{153}$$

where the last equation is derived from adding the previous two. We now derive inequalities for g and h^*. From the subdifferential characterization of the proximal operator (Lemma 1), the update $\boldsymbol{z}_t = \mathbf{prox}_{\gamma h}(\boldsymbol{y}_t)$ implies the inclusion

$$\boldsymbol{u}_t = \frac{1}{\gamma}(\boldsymbol{y}_t - \boldsymbol{z}_t) \in \partial h(\boldsymbol{z}_t) \iff \boldsymbol{z}_t \in \partial h^*(\boldsymbol{u}_t) \tag{154}$$

where the implication is a consequence of the Fenchel-Young inequality, see e.g. (Bauschke and Combettes, 2017, Proposition 16.10) or (Rockafellar and Wets, 1998, Proposition 11.3). Similarly, the update $\boldsymbol{x}_t = \mathbf{prox}_{\gamma g}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t)$ in its turn gives the inclusion

$$\frac{1}{\gamma}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t - \boldsymbol{x}_t) \in \partial g(\boldsymbol{x}_t) \tag{155}$$

By convexity of g and h^* we then have the inequalities

$$h^*(\boldsymbol{u}_t) - h^*(\boldsymbol{u}) \leq \langle\boldsymbol{z}_t, \boldsymbol{u}_t - \boldsymbol{u}\rangle \ . \tag{156}$$

$$g(\boldsymbol{x}_t) - g(\boldsymbol{x}) \leq \frac{1}{\gamma}\langle\boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \langle\boldsymbol{u}_t + \boldsymbol{v}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \tag{157}$$

Adding (153) and (157) we obtain

$$f(\boldsymbol{x}_t) + g(\boldsymbol{x}_t) - f(\boldsymbol{x}) - g(\boldsymbol{x}) \leq \frac{1}{\gamma}\langle \boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2 - \langle \boldsymbol{u}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \tag{158}$$

Using these, we can now write the following sequence of inequalities for the Lagrangian suboptimality

$$\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}_t) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t) = f(\boldsymbol{x}_t) - f(\boldsymbol{x}) + g(\boldsymbol{x}_t) - g(\boldsymbol{x}) + \langle \boldsymbol{x}_t - \boldsymbol{x}, \boldsymbol{u}_t\rangle \tag{159}$$

$$\overset{(158)}{\leq} \frac{1}{\gamma}\langle \boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2 + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \tag{160}$$

$$\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}_t) = h^*(\boldsymbol{u}_t) - h^*(\boldsymbol{u}) + \langle \boldsymbol{x}_t, \boldsymbol{u} - \boldsymbol{u}_t\rangle \overset{(156)}{\leq} \langle \boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{u}_t - \boldsymbol{u}\rangle \tag{161}$$

Adding these two last equations we have

$$\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t) \leq \frac{1}{\gamma}\langle \boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2 + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle + \langle \boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{u}_t - \boldsymbol{u}\rangle \tag{162}$$

$$= \frac{1}{\gamma}\langle \boldsymbol{z}_t - \boldsymbol{x}_t, \boldsymbol{x}_t + \gamma\boldsymbol{u}_t - \boldsymbol{x} - \gamma\boldsymbol{u}\rangle + \frac{L}{2}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^2 + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \tag{163}$$

$$= \frac{1}{\gamma}\langle \boldsymbol{y}_t - \boldsymbol{y}_{t+1}, \boldsymbol{y}_{t+1} - \boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t+1}\|^2 + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \tag{164}$$

(using $\boldsymbol{y}_{t+1} - \boldsymbol{y}_t = \boldsymbol{x}_t - \boldsymbol{z}_t$)

$$= \frac{1}{2\gamma}\|\boldsymbol{y}_t - \boldsymbol{y}\|^2 + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 - \frac{1}{2\gamma}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}\|^2 + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \quad , \tag{165}$$

$$\leq \frac{1}{2\gamma}\|\boldsymbol{y}_t - \boldsymbol{y}\|^2 - \frac{1}{2\gamma}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}\|^2 + \langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle \quad , \tag{166}$$

where the second equality comes from the definition of $\boldsymbol{u}_t, \boldsymbol{y}_{t+1}$ and $\boldsymbol{y}$ and in the last equality we have applied the identity $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \|\boldsymbol{a} + \boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2$. In the last inequality we have used the assumption $\gamma \leq 1/L$.

We will now upper bound the last term. For this, we introduce the variable $\widetilde{\boldsymbol{x}}$, which represents the step in $\boldsymbol{x}$ that would be taken if we used the full gradient rather than the SAGA gradient approximation:

$$\widetilde{\boldsymbol{x}}_t \overset{\text{def}}{=} \mathbf{prox}_{\gamma g}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\nabla f(\boldsymbol{z}_t)) \quad . \tag{167}$$

Taking expectations on this last quantity we have

$$\mathbf{E}\langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle = \mathbf{E}\langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_t - \widetilde{\boldsymbol{x}}_t\rangle + \mathbf{E}\langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \widetilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\rangle \tag{168}$$

$$\leq \mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|\|\boldsymbol{x}_t - \widetilde{\boldsymbol{x}}_t\| + \mathbf{E}\langle \boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t), \widetilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\rangle \tag{169}$$

(Cauchy-Schwarz)

$$= \mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|\|\boldsymbol{x}_t - \widetilde{\boldsymbol{x}}_t\| \tag{170}$$

(since $\widetilde{\boldsymbol{x}}_t$ does not depend on i and $\mathbf{E}\boldsymbol{v}_t = \nabla f(\boldsymbol{z}_t)$)

$$= \mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|\|\, \mathbf{prox}_{\gamma g}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t) - \mathbf{prox}_{\gamma g}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\nabla f(\boldsymbol{z}_t))\| \tag{171}$$

$$\leq \gamma\mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|^2 \tag{172}$$

(nonexpansiveness of **prox**)

$$= \gamma\mathbf{E}\|\boldsymbol{v}_t - \nabla f(\boldsymbol{z}_t)\|^2 \tag{173}$$

$$\leq \gamma(1+\eta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 + 2\gamma(1+\eta^{-1})L_f H_t \ , \tag{174}$$

where the last inequality follows by Lemma 13 for dense update variants. Taking conditional expectations in (166), plugging this bound, multiplying everything by 2γ and rearranging we obtain

$$\begin{aligned} &2\gamma\mathbf{E}(\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{u})-\mathcal{L}(\boldsymbol{x},\boldsymbol{u}_t))+\mathbf{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}\|^2 \\ &\quad\leq \mathbf{E}\|\boldsymbol{y}_t-\boldsymbol{y}\|^2+2\gamma^2(1+\eta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t)-\nabla\xi_i(\boldsymbol{x}^\star)\|^2+4\gamma^2(1+\eta^{-1})L_fH_t \end{aligned} \tag{175}$$

which is the desired bound. □

Theorem 1. *Let $\overline{\boldsymbol{x}}_t$ denote the averaged (also known as ergodic) iterate, i.e., $\overline{\boldsymbol{x}}_t=(\sum_{k=0}^t\boldsymbol{x}_k)/(t+1)$ and $\overline{\boldsymbol{u}}_t=(\sum_{k=0}^t\boldsymbol{u}_k)/(t+1)$. Then* VR-TOS *(Algorithm 1) methods converge for any step size $\gamma\leq 1/(3L_f)$, and for $\gamma=1/(3L_f), t\geq 0$ we have the following bound for all $(\boldsymbol{x},\boldsymbol{u})\in\operatorname{dom}g\times\operatorname{dom}h^*$:*

$$\mathbb{E}\left[\mathcal{L}(\overline{\boldsymbol{x}}_t,\boldsymbol{u})-\mathcal{L}(\boldsymbol{x},\overline{\boldsymbol{u}}_t)\right]\leq\frac{10n}{q(t+1)}\left[\frac{3L_fq}{20n}\|\boldsymbol{y}_0-\boldsymbol{y}\|^2+\frac{3L_fq}{2n}\|\boldsymbol{y}_0-\boldsymbol{y}^\star\|^2+H_0\right], \tag{176}$$

with $\boldsymbol{y}=\boldsymbol{x}+\gamma\boldsymbol{u}$, $\boldsymbol{y}^\star\in\operatorname{Fix}(\boldsymbol{G}_\gamma)$.

Furthermore, if h is β_h-Lipschitz we have the following rate in terms of the primal objective:

$$\mathcal{P}(\overline{\boldsymbol{x}}_t)-\mathcal{P}(\boldsymbol{x}^\star)\leq\frac{10n}{q(t+1)}\left[\frac{6L_fq}{20n}\|\boldsymbol{z}_0-\boldsymbol{x}^\star\|^2+\frac{3L_fq}{2n}\|\boldsymbol{y}_0-\boldsymbol{y}^\star\|^2+\frac{q}{15nL_f}\beta_h^2+H_0\right]. \tag{177}$$

Proof. We define the following Lyapunov function:

$$W_t(\boldsymbol{x},\boldsymbol{u})\overset{\text{def}}{=}V_t+\lambda\|\boldsymbol{y}_t-\boldsymbol{y}\|^2\quad\text{with }\boldsymbol{y}=\boldsymbol{x}+\gamma\boldsymbol{u}\,. \tag{178}$$

We will now aim to bound $\mathbf{E}W_{t+1}-W_t$ by using Lemma 12 with $\mu=0$ and $s=1$, we have for $\mathbf{E}V_{t+1}$ that

$$\begin{aligned} \mathbf{E}V_{t+1}\leq V_t&+\left(2L_fc\gamma^2(1+\beta^{-1})-\frac{q}{n}\right)H_t \\ &+c\left(\gamma^2(1+\beta)-\frac{\gamma}{L}\right)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t)-\nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ &+\left(-2c\gamma+\frac{q}{n}\right)B_f(\boldsymbol{z}_t,\boldsymbol{x}^\star) \end{aligned} \tag{179}$$

while for the last term from Lemma 14 we have

$$\begin{aligned} \mathbf{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}\|^2\leq\|\boldsymbol{y}_t-\boldsymbol{y}\|^2&-2\gamma\mathbf{E}(\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{u})-\mathcal{L}(\boldsymbol{x},\boldsymbol{u}_t)) \\ &+2\gamma^2(1+\eta)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t)-\nabla\xi_i(\boldsymbol{x}^\star)\|^2+4\gamma^2(1+\eta^{-1})L_fH_t \end{aligned} \tag{180}$$

Adding (179) and (180) times λ we have

$$\begin{aligned} \mathbf{E}W_{t+1}(\boldsymbol{x},\boldsymbol{u})-W_t(\boldsymbol{x},\boldsymbol{u})\leq&-2\lambda\gamma\mathbf{E}(\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{u})-\mathcal{L}(\boldsymbol{x},\boldsymbol{u}_t)) \\ &+\left[4\gamma^2\lambda(1+\eta^{-1})L_f+2L_fc\gamma^2(1+\beta^{-1})-\frac{q}{n}\right]H_t \\ &+\gamma\left[2\lambda(1+\eta)\gamma+\gamma(1+\beta)c-\frac{c}{L_f}\right]\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t)-\nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ &+\left[-2c\gamma+\frac{q}{n}\right]B_f(\boldsymbol{z}_t,\boldsymbol{x}^\star) \end{aligned} \tag{181}$$

We can now verify that with the coefficients

$$\gamma=\frac{1}{3L_f},\quad c=\frac{3L_fq}{2n},\quad\beta=\eta=\frac{3}{2},\quad\lambda=\frac{3L_fq}{20n}\,, \tag{182}$$

all the square brackets are negative and so we have

$$\mathbf{E}W_{t+1}(\boldsymbol{x},\boldsymbol{u})-W_t(\boldsymbol{x},\boldsymbol{u})\leq-\frac{q}{10n}\mathbf{E}(\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{u})-\mathcal{L}(\boldsymbol{x},\boldsymbol{u}_t)) \tag{183}$$

These expectations are conditional on information from step t. Taking full expectations (with respect to all randomness) we have

$$\mathbb{E}W_{t+1}(\boldsymbol{x}, \boldsymbol{u}) - \mathbb{E}W_t(\boldsymbol{x}, \boldsymbol{u}) \leq -\frac{q}{10n}\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t)\right] \ , \tag{184}$$

where all expectations are unconditional. Adding the previous inequality from 0 to t, the terms in W_t cancel each other and we have

$$\frac{q}{10n}\mathbb{E}\left[\sum_{k=0}^{t} \mathcal{L}(\boldsymbol{x}_k, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_k)\right] \leq W_0(\boldsymbol{x}, \boldsymbol{u}) - \mathbb{E}W_{t+1}(\boldsymbol{x}, \boldsymbol{u}) \ . \tag{185}$$

We can drop the last term since it is always negative. Note that the function $\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t)$ is convex in $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$ and so we can apply Jensen's inequality. This gives

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(\overline{\boldsymbol{x}}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \overline{\boldsymbol{u}}_k)\right] &\leq \frac{10n}{q(t+1)} W_0(\boldsymbol{x}, \boldsymbol{u}) \\ &= \frac{10n}{q(t+1)}\left[\frac{3L_f q}{20n}\|\boldsymbol{y}_0 - \boldsymbol{y}\|^2 + \frac{3L_f q}{2n}\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + H_0\right] \ , \end{aligned} \tag{186}$$

which proves the first result of the theorem.

For the second result, let $\widehat{\boldsymbol{u}} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{u}} \mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \boldsymbol{u})$ and $(\boldsymbol{x}^\star, \boldsymbol{u}^\star)$ be a saddle point of $\mathcal{L}$. Then $\mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \widehat{\boldsymbol{u}}) = P(\overline{\boldsymbol{x}}_{t+1})$ and $\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{u}^\star) = P(\boldsymbol{x}^\star)$ by definition of Fenchel dual.

At the same time, by the β_h-Lipschitz assumption on h implies that the norm of every element in $\operatorname{dom} h^*$ is bounded by β_h (see e.g., (Rockafellar, 1997, Corollary 13.3.3)). This way we bound

$$\|\boldsymbol{y}_0 - \boldsymbol{y}\|^2 = \|\boldsymbol{z}_0 + \gamma \boldsymbol{u}_0 - \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + 2\gamma^2\|\boldsymbol{u}_0 - \boldsymbol{u}\|^2 \tag{187}$$

$$\leq 2\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + 4\gamma^2\beta_h^2 \tag{188}$$

Plugging this bound into the last inequality with $\boldsymbol{x} = \boldsymbol{x}^\star$

$$P(\overline{\boldsymbol{x}}_{t+1}) - P(\boldsymbol{x}^*) \leq \frac{10n}{q(t+1)}\left[\frac{6L_f q}{20n}\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{3L_f q}{2n}\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + \frac{q}{15nL_f}\beta_h^2 + H_0\right] \ . \tag{189}$$

□

Appendix C.4 Sublinear convergence – sparse algorithms

Theorem 2. *Sparse* VR-TOS *(Algorithm 2) converges for every step size $\gamma \leq 1/(3L_f)$. In particular, for $\gamma = 1/(3L_f)$ and $\boldsymbol{y}_t$ obtained after $t \geq 1$ updates we have the bound*

$$\min_{k=0,\ldots,t} \{\mathbb{E}\|\boldsymbol{y}_k - \boldsymbol{G}_\gamma(\boldsymbol{y}_k)\|\} \leq \sqrt{\frac{C_0}{Lq(t+1)}} = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right), \tag{190}$$

with $C_0 = \frac{5d_{\max}n}{Lq(t+1)}\left[(2Lq/n)\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|^2 + H_0\right]$.

Proof. Using the Lyapunov inequality of Lemma 12 for non-strongly convex functions, i.e., with $\mu = 0$ we have

$$\begin{aligned}\mathbf{E}V_{t+1} \leq V_t + c(s-1)\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + \left(\frac{2L_f c\gamma^2}{s}(1+\beta^{-1}) - \frac{q}{n}\right) H_t \\ + c\left(\frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f}\right)\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ + \left(-2c\gamma + \frac{q}{n}\right) B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star)\end{aligned} \tag{191}$$

where V_t and H_t are as defined in Eq. (101). For notational convenience, we define $\boldsymbol{R}$ as the operator residual $\boldsymbol{R}(\boldsymbol{y}) = \boldsymbol{G}_\gamma(\boldsymbol{y}) - \boldsymbol{y}$, and denote by i the random index selected at the t-th iteration. The term $\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2$ can be bounded in term of the gradient mapping using the following inequality, where $\widetilde{\boldsymbol{x}} = \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \boldsymbol{D}\nabla f(\boldsymbol{z}_t))$ is the value of $\boldsymbol{x}_t$ had we used the full gradient instead of the stochastic approximation:

$$\begin{aligned}\|\boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t)\|^2 &= \|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t + \boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t) - \boldsymbol{y}_{t+1} + \boldsymbol{y}_t\|^2 && (192)\\ &\leq 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 2\|\boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t) - \boldsymbol{y}_{t+1} + \boldsymbol{y}_t\|^2 && (193)\\ &= 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 2\|\boldsymbol{R}(\boldsymbol{y}_t) - \boldsymbol{y}_{t+1} + \boldsymbol{y}_t\|_{(i)}^2 && (194)\\ &\quad \text{(since both } \boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t) \text{ and } \boldsymbol{y}_{t+1} + \boldsymbol{y}_t \text{ have support in } T_i) \\ &= 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 2\|\boldsymbol{G}_\gamma(\boldsymbol{y}_t) - \boldsymbol{y}_{t+1}\|_{(i)}^2 && (195)\\ &\quad \text{(by definition of } \boldsymbol{R}) \\ &= 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 2\|\boldsymbol{y}_t - \boldsymbol{z}_t + \widetilde{\boldsymbol{x}}_t - (\boldsymbol{y}_t - \boldsymbol{z}_t + \boldsymbol{x}_t)\|_{(i)}^2 && (196)\\ &\quad \text{(by definition of } \boldsymbol{G}_\gamma \text{ and } \boldsymbol{y}_{t+1}) \\ &= 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 2\|\widetilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\|_{(i)}^2 \,. && (197)\end{aligned}$$

For the last term, we further have

$$\begin{aligned}\|\boldsymbol{x}_t - \widetilde{\boldsymbol{x}}_t\|_{(i)}^2 &= \|\,\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t) - \mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(2\boldsymbol{z}_t - \boldsymbol{y} - \gamma\boldsymbol{D}\nabla f(\boldsymbol{z}_t))\|_{(i)}^2 && (198)\\ &\leq \gamma^2\|\boldsymbol{v}_t - \boldsymbol{D}\nabla f(\boldsymbol{z}_t)\|_{(i)}^2 \qquad \text{(by Lemma 7)} && (199)\\ &\leq 2\gamma^2\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 + 4\gamma^2 L_f H_t \quad \text{(by Lemma 13)} && (200)\end{aligned}$$

Combining this into Eq. (192) and tacking expectation, we have:

$$\begin{aligned}&\mathbf{E}\|\boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t)\|^2 \leq 2\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 + 4\gamma^2\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 + 8\gamma^2 L_f H_t && (201)\\ \iff\; &-\mathbf{E}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|^2 \leq -\frac{1}{2}\mathbf{E}\|\boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t)\|^2 + 2\gamma^2\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 + 4\gamma^2 L_f H_t && (202)\end{aligned}$$

Plugging this last inequality in Eq. (191) gives

$$\begin{aligned}\mathbf{E}V_{t+1} \leq V_t + \frac{c(s-1)}{2}\mathbf{E}\|\boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t)\|^2 + \left[2c(s-1)\gamma^2 L_f + \frac{2L_f c\gamma^2}{s}(1+\beta^{-1}) - \frac{q}{n}\right] H_t \\ + c\left[(s-1)\gamma^2 + \frac{\gamma^2}{s}(1+\beta) - \frac{\gamma}{L_f}\right]\mathbf{E}\|\nabla\xi_i(\boldsymbol{z}_t) - \nabla\xi_i(\boldsymbol{x}^\star)\|^2 \\ + \left[-2c\gamma + \frac{q}{n}\right] B_f(\boldsymbol{z}_t, \boldsymbol{x}^\star)\end{aligned} \tag{203}$$

We can verify that with the following values

$$\gamma = \frac{1}{3L_f} \ , \quad \beta = 3/2 \ , \quad s = 8/10 \ , \quad c = \frac{2Lq}{n} \ , \tag{204}$$

all the square brackets in the previous expression are non-positive and so we have

$$\begin{aligned}
\mathbf{E}V_{t+1} - V_t &\leq -\frac{Lq}{5n}\mathbf{E}\|\boldsymbol{P}_i\boldsymbol{R}(\boldsymbol{y}_t)\|^2 && (205)\\
&= -\frac{Lq}{5n}\mathbf{E}\|\boldsymbol{R}(\boldsymbol{y}_t)\|^2_{(i)} = -\frac{Lq}{5n}\|\boldsymbol{R}(\boldsymbol{y}_t)\|^2_{\boldsymbol{D}^{-1}} && (206)\\
&\leq -\frac{Lq}{5d_{\max}n}\|\boldsymbol{R}(\boldsymbol{y}_t)\|^2 && (207)\\
\iff V_t - \mathbf{E}V_{t+1} &\geq \frac{Lq}{5d_{\max}n}\|\boldsymbol{R}(\boldsymbol{y}_t)\|^2 && (208)
\end{aligned}$$

Summing from 0 to t and chaining expectations have

$$\mathbb{E}V_0 - \mathbb{E}V_{t+1} \geq \frac{Lq}{5d_{\max}n}\sum_{k=0}^{k}\|\boldsymbol{R}(\boldsymbol{y}_k)\|^2 \geq \frac{Lq}{5d_{\max}n}\sum_{k=0}^{t}\|\boldsymbol{R}(\boldsymbol{y}_k)\|^2 \geq \frac{Lq(t+1)}{5d_{\max}n}\min_{k=0,\dots,t}\|\boldsymbol{R}(\boldsymbol{y}_t)\|^2$$

Dropping $\mathbb{E}V_{t+1}$ (since it is positive) and taking the square root we have

$$\min_{k=0,\dots,t}\|\boldsymbol{R}(\boldsymbol{y}_t)\|^2 \leq \frac{5d_{\max}n}{Lq(t+1)}V_0 \ , \tag{209}$$

The final results follows then by definition of $\boldsymbol{R}$. □

Appendix D Learning with multiple penalties

In this section we review some cases in which we can compute the scaled proximal operator $\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}$ for some diagonal matrix $\boldsymbol{D}$. We refer to (Pedregosa and Gidel, 2018) for a discussion on how common penalties such as ℓ_1 trend filtering, multidimensional total variation, overlapping group lasso, etc. can be split as a sum of proximal terms.

Appendix D.1 ℓ_1 norm

We consider the case in which g is the ℓ_1 or Lasso penalty, $g(\boldsymbol{x}) \stackrel{\text{def}}{=} \|\boldsymbol{x}\|_1$. Since this function is fully separable, its resolvent can be computed component-wise. Hence, the reweighting matrix $\boldsymbol{D}$ can be associated with the step size γ and using the known prox for the Lasso penalty we obtain

$$\left[(\mathrm{Id} + \gamma \boldsymbol{D} \partial g)^{-1} \boldsymbol{x}\right]_j = \left(1 - \frac{[\boldsymbol{D}]_{j,j}\gamma}{|\boldsymbol{x}_j|}\right)_+ \boldsymbol{x}_j$$

Appendix D.2 Fused lasso

The fused lasso penalty, also known as 1-dimensional total variation, is defined as the ℓ_1 norm of the differences between consecutive coefficients. Although in this case direct methods have been developed to compute its proximal operator (Condat, 2013b; Johnson, 2013), there still exist advantages in splitting the penalty. In particular, existing direct approaches involve dense updates due to the non-separability of the penalty. However, by splitting the penalty into constituents that are block-separable, it is possible to optimize with this penalty while only performing sparse updates. The split is the following:

$$\|\boldsymbol{x}\|_{\text{FL}} \stackrel{\text{def}}{=} \sum_{i=1}^{p-1} |\boldsymbol{x}_i - \boldsymbol{x}_{i+1}| = \underbrace{\sum_{i=1}^{r} |\boldsymbol{x}_{2i-1} - \boldsymbol{x}_{2i}|}_{\stackrel{\text{def}}{=} g(\boldsymbol{x})} + \underbrace{\sum_{i=1}^{s} |\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i+1}|}_{\stackrel{\text{def}}{=} h(\boldsymbol{x})}, \tag{210}$$

with $r = \lfloor (p-1)/2 \rfloor$ and $s = \lfloor p/2 \rfloor$. We note that both g and h are block-separable with blocks of size 2. Furthermore, it is possible to compute the scaled proximal operator of $\mathbf{prox}_{\gamma g}^{Q}(\boldsymbol{x})$ in closed form. The advantages of Vr-Tos with this formulation on large and sparse problems is demonstrated experimentally in §5.

Both functions g and h are block-separable with blocks of size two. Hence it is sufficient to specify the proximal operator on a vector of size two. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{q}_1, \boldsymbol{q}_2)$. Then we have

$$\mathbf{prox}_{\gamma g}^{\boldsymbol{D}^{-1}}(\boldsymbol{x}) = \begin{cases} \left(\boldsymbol{x}_1 - \gamma/\boldsymbol{q}_1, \boldsymbol{x}_2 + \gamma/\boldsymbol{q}_2\right) & \text{if } \boldsymbol{x}_1 - \gamma/\boldsymbol{q}_1 \geq \boldsymbol{x}_2 + \gamma \boldsymbol{q}_2 \\ \left(\boldsymbol{x}_1 + \gamma/\boldsymbol{q}_1, \boldsymbol{x}_2 - \gamma/\boldsymbol{q}_2\right) & \text{if } \boldsymbol{x}_1 + \gamma/\boldsymbol{q}_1 \leq \boldsymbol{x}_2 - \gamma \boldsymbol{q}_2 \\ \left(\frac{\boldsymbol{q}_1 \boldsymbol{x}_1 + \boldsymbol{q}_2 \boldsymbol{x}_2}{\boldsymbol{q}_1 + \boldsymbol{q}_2}, \frac{\boldsymbol{q}_1 \boldsymbol{x}_1 + \boldsymbol{q}_2 \boldsymbol{x}_2}{\boldsymbol{q}_1 + \boldsymbol{q}_2}\right) & \text{otherwise} \quad . \end{cases} \tag{211}$$

Proof. Let $(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathbf{prox}_{\gamma g}^{\boldsymbol{D}^{-1}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. The first order optimality conditions applied to this problem give

$$\frac{\boldsymbol{D}^{-1}}{\gamma}((\boldsymbol{x}_1, \boldsymbol{x}_2) - (\boldsymbol{z}_1, \boldsymbol{z}_2)) \in \partial |\boldsymbol{z}_1 - \boldsymbol{z}_2|$$

We now perform a dichotomy of cases. Suppose first $\boldsymbol{z}_1 - \boldsymbol{z}_2 > 0$. Then the above becomes

$$\frac{\boldsymbol{D}^{-1}}{\gamma}((\boldsymbol{x}_1, \boldsymbol{x}_2) - (\boldsymbol{z}_1, \boldsymbol{z}_2)) = (1, -1)$$

from where the solution is given by $(\boldsymbol{x}_1 - \gamma/\boldsymbol{q}_1, \boldsymbol{x}_2 + \gamma/\boldsymbol{q}_2)$, but only if $\boldsymbol{x}_2 - \gamma/\boldsymbol{q}_1 \geq \boldsymbol{x}_2 + \gamma/\boldsymbol{q}_2$, otherwise the assumption $\boldsymbol{z}_1 - \boldsymbol{z}_2 < 0$ would be violated.

Repeating this for $\boldsymbol{z}_1 - \boldsymbol{z}_2 < 0$ and $\boldsymbol{z}_1 - \boldsymbol{z}_2 = 0$ yields the above rule. □

Appendix E Pseudocode for the extension to k proximal terms

The extension of the proposed method to k proximal terms consists in running Algorithm 1 or 2 on particular values of g and h. Some tricks can help to reduce the memory usage of this algorithm, reducing the storage of vectors $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{v}_t$ from $k \times p$ to p. In this subsection we provide the pseudocode for runnning Sparse VR-TOS on its k-proximal terms extension.

As in §2.2 we consider an optimization problem of the form

$$\begin{aligned} &\underset{\boldsymbol{X} \in \mathbb{R}^{k \times p}}{\text{minimize}} \; f(\overline{\boldsymbol{X}}) + \textstyle\sum_{j=1}^{k} g_j(\boldsymbol{X}_j) + h(X)\,, \\ &\text{with } f(\boldsymbol{x}) = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \psi_i(\boldsymbol{x}) + \omega(\boldsymbol{x})\;, \end{aligned} \tag{OPT-k}$$

where $h(X) = \imath\{\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_k\}$. We will first detail how the scaled proximal operator of h can be computed

Lemma 15. *Let $h(X) = \imath\{\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_k\}$. Then we have that*

$$\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{x}) = \boldsymbol{z}\mathbf{1}_k{}^T \text{ for } \boldsymbol{z} \in \mathbb{R}^p \text{ defined as} \tag{212}$$

$$\boldsymbol{z}_j = \left(\sum_{i=1}^{k} a_{i,j} \boldsymbol{X}_{i,j}\right) / \left(\sum_{i=1}^{k} a_{i,j}\right) \text{ with } a_{i,j} = \boldsymbol{D}^{-1}_{ip+j,ip+j}. \tag{213}$$

Proof. Let $\mathcal{S}$ denote the domain of h, i.e., $\mathcal{S} \stackrel{\text{def}}{=} \{\boldsymbol{X} \in \mathbb{R}^{k \times p} | \boldsymbol{X}_1 = \boldsymbol{X}_2 = \cdots = \boldsymbol{X}_k\}$. Computing this proximal operator consists by definition of scaled proximal operator in solving the following optimization problem

$$\underset{\boldsymbol{Z} \in S}{\arg\min} \, \|\mathbf{vec}(\boldsymbol{Z}) - \mathbf{vec}(\boldsymbol{X})\|^2_{\boldsymbol{D}^{-1}} = \underset{\boldsymbol{z} \in \mathbb{R}^p}{\arg\min} \, \|\mathbf{vec}(\boldsymbol{z}\mathbf{1}_k^T) - \mathbf{vec}(\boldsymbol{X})\|^2_{\boldsymbol{D}^{-1}} \tag{214}$$

The problem is then separable along the components of $\boldsymbol{z}$, and the j-th component is the solution to the problem

$$\underset{\boldsymbol{z}_j \in \mathbb{R}}{\arg\min} \sum_{i=1}^{k} a_{i,j} (\boldsymbol{z}_j - \boldsymbol{X}_{j,i})^2 \text{ with } a_{i,j} = \boldsymbol{D}^{-1}_{ip+j,ip+j}\;, \tag{215}$$

and whose solution is

$$\boldsymbol{z}_j = \left(\sum_{i=1}^{k} a_{i,j} \boldsymbol{X}_{i,j}\right) / \left(\sum_{i=1}^{k} a_{i,j}\right) \tag{216}$$

□

Before introducing the algorithm, we make the following definitions:

- Let $\mathcal{B}_j$ denote the blocks of g_j, that is, g_j can be decomposed block coordinate-wise as $g_j(\boldsymbol{x}) = \sum_{B \in \mathcal{B}_j} g_{j,B}([\boldsymbol{x}]_B)$.
- Let $T_{i,j}$ denote the extended support of $\nabla \psi_i$ in $\mathcal{B}_j$, that is, $T_{i,j} \stackrel{\text{def}}{=} \{B : \text{supp}(\nabla f_i) \cap B \neq \varnothing,\, B \in \mathcal{B}_j\}$.
- Let S_i be the set of coordinates that are at least in one block of one of the extended supports, that is, $S_i \stackrel{\text{def}}{=} \{c : c \in B \text{ for any } B \in T_{i,j} \text{ and any } j = 1, \ldots, k\}$.

With respect to Algorith 2, compute the $\boldsymbol{z}$ update at the end of the algorithm instead of the beginning to efficiently use the extended support.

Algorithm 3: Sparse VR-TOS for k proximal terms

Input: $\boldsymbol{Y}_0 \in \mathbb{R}^{k\times p}$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^{n\times p}$, $\gamma > 0$

Temporary storage: $\boldsymbol{z}_t$, $\boldsymbol{v}_t$ and $\boldsymbol{x}_t$, all in $\mathbb{R}^p$

Result: approximate solution to (OPT-k)

for $t = 0, 1, \ldots$ **do**

  Sample $i \in \{1, \ldots, n\}$ uniformly at random

  Compute $\nabla\psi_i(\boldsymbol{z}_t)$

  for $j = 1, \ldots, k$ **do**

    $[\boldsymbol{v}_t]_{T_{i,j}} = \frac{1}{k}[\nabla\psi_i(\boldsymbol{z}_t) - \boldsymbol{\alpha}_{i,t} + \boldsymbol{D}^{(j)}(\overline{\boldsymbol{\alpha}}_t + \nabla\omega(\boldsymbol{z}_t))]_{T_i}$

    $[\boldsymbol{x}_t]_{T_{i,j}} = [\mathbf{prox}_{\gamma\varphi_{i,j}}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma\boldsymbol{v}_t)]_{T_i}$

    $[\boldsymbol{Y}_{j,t+1}]_{T_{i,j}} = [\boldsymbol{Y}_{j,t} + \boldsymbol{x}_t - \boldsymbol{z}_t]_{T_i}$

    for $b \in T_i$ **do**

      $\boldsymbol{z}_{t+1,b} = \left(\sum_{l=1}^n \boldsymbol{D}^{(l)}_{b,b}\boldsymbol{Y}_{l,b}\right) / \left(\sum_{l=1}^n \boldsymbol{D}^{(l)}_{b,b}\right)$

  update $\boldsymbol{\alpha}_{t+1}$ according to (1)

return $\mathbf{prox}_{\gamma h}^{\boldsymbol{D}^{-1}}(\boldsymbol{y}_t)$

Appendix F Experiments

Appendix F.1 Implementation aspects

We review some implementation details for the proposed algorithms

Update of memory terms. In a practical implementation of the SAGA variants, the vector $\overline{\boldsymbol{\alpha}}_t = (1/n)\sum_{i=1}^n \boldsymbol{\alpha}_{i,t}$ is also stored in memory and updated incrementally as $\overline{\boldsymbol{\alpha}}_{t+1} = \overline{\boldsymbol{\alpha}}_t + (\boldsymbol{\alpha}_{i,t+1} - \boldsymbol{\alpha}_{i,t})/n$.

Compressed memory storage. Like other SAGA variants, VR-TOS with the SAGA-like update of memory terms requires to store a table of partial gradients. In the general case, this requires a matrix of size $n \times p$. However, for linearly-parametrized loss functions this can be compressed into a matrix of size n. Linearly-parametrized functions are of the form $\psi_i(\boldsymbol{x}) = l_i(\boldsymbol{a}_i^T\boldsymbol{x})$ for some input dataset $\{\boldsymbol{a}_i\}_{i=1}^n$ and some real functions $\{l_i\}_{i=1}^n$. Deriving with respect to $\boldsymbol{x}$ one obtains $\nabla\psi_i(\boldsymbol{x}) = \boldsymbol{a}_i l_i'(\boldsymbol{a}_i^T\boldsymbol{x})$. In this expression only the factor $l_i'(\boldsymbol{a}_i^T\boldsymbol{x})$ depends on the iterate $\boldsymbol{x}$, and it is a scalar. Hence, we only need to store this scalar and we can construct the partial gradient at run time by multiplying by the vector $\boldsymbol{a}_i$. The memory cost is hence reduced to a list of n scalars.

Initialization of $\boldsymbol{\alpha}_0$. The original SAGA algorithm of (Defazio et al., 2014) required to initialize the memory terms as $\boldsymbol{\alpha}_{i,t} = \nabla\psi_i(\boldsymbol{z}_0)$. This is no longer required in our algorithm, in which these memory terms can be initialized arbitrarily. In fact, we recommend to initialize them to zero. This is convenient and makes the gradient estimate $\boldsymbol{v}_t$ close to the SGD estimate during the first iterations.

Initialization of $\boldsymbol{y}_0$. An "initial guess" $\boldsymbol{y}_0$ must also be provided. From Theorem 1 and Appendix B, we have that $\boldsymbol{y}_t$ converges towards $\boldsymbol{x}^\star + \gamma\boldsymbol{u}^\star$, where $\boldsymbol{u}^\star$ is a minimizer of the dual objective $\mathcal{D}$. Hence, the ideal initialization for this vector is $\boldsymbol{x}_0 + \gamma\boldsymbol{u}_0$, where $\boldsymbol{x}_0$ is an initial guess for (OPT) and $\boldsymbol{u}_0$ is an initial guess for the dual problem. However, we rarely have an initial guess for the dual problem, in which case one can set $\boldsymbol{u}_0 = \boldsymbol{0}$.

SGD-TOS. Following (Yurtsever et al., 2016), we used a step size of the form γ/t in this case, where t is the number of iterations.

Software. All methods are implemented in Python. Numba was used to speed up the inner loops of stochastic methods (VR-TOS, SAGA, ProxSVRG and STOS). For the Adaptive Three Operator splitting method we used the implementation provided by the authors[5].

Appendix F.2 Overlapping Group Lasso Benchmarks

In this subsection we giver some details on the benchmarks reported in §5 that were omitted from the main text.

The associated objective function that we consider is

$$\frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-b_i\boldsymbol{a}_i^\intercal\boldsymbol{x})\right) + \frac{\lambda_1}{2}\|\boldsymbol{x}\|^2 + \lambda_2\|\boldsymbol{x}\|_{\text{OGL}},$$

where $\boldsymbol{a}_i \in \mathbb{R}^p$ and $b_i \in \{-1, +1\}$ are the data samples.

The overlapping group lasso penalty $\|\cdot\|_{\text{OGL}}$ is defined as the sum over the group norms. Given a collection of (potentially overlapping) groups $\mathcal{G}$, the overlapping group penalty is given by

$$\|\boldsymbol{x}\|_{\text{OGL}} = \sum_{g\in\mathcal{G}} \|[\boldsymbol{x}]_g\|_2 \ . \tag{217}$$

In our comparison the groups are chosen to have 10 variables with 2 variables of overlap between two successive groups: $\{\{1, \ldots, 10\}, \{8, \ldots, 18\}, \{16, \ldots, 26\}, \ldots\}$.

Although this penalty can be expressed as a sum of only two proximal terms, we instead use the formulation in §2.2 in order to avoid computing the scaled proximal operator and to better leverage the sparsity in the dataset.

[5] http://openopt.github.io/copt/

Extra experiments. We also run the same benchmark on the KDD12 dataset (149,639,105 samples and 54,686,452 features) but was not shown in the main paper due to lack of space. The results are displayed below and are consistent with the rest of the experiments.

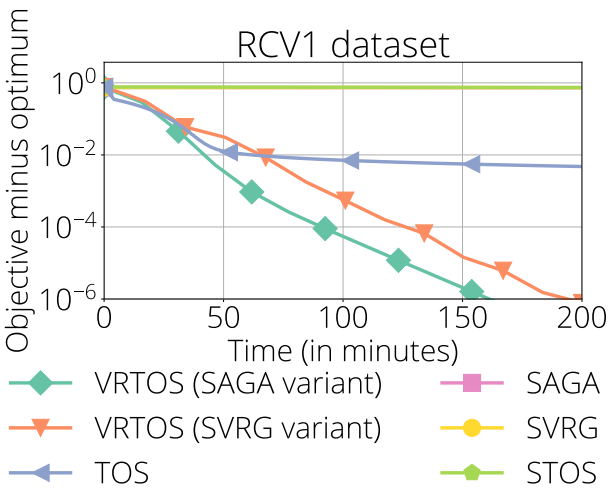

Figure 2: Benchmarks on the KDD12 dataset.