Adaptive Three Operator Splitting

Fabian Pedregosa [1 2] **Gauthier Gidel** [3]

Abstract

We propose and analyze an adaptive step-size variant of the Davis-Yin three operator splitting. This method can solve optimization problems composed of a sum of a smooth term for which we have access to its gradient and an arbitrary number of potentially non-smooth terms for which we have access to their proximal operator. The proposed method sets the step-size based on local information of the objective –hence allowing for larger step-sizes–, only requires two extra function evaluations per iteration and does not depend on any step-size hyperparameter besides an initial estimate. We provide an iteration complexity analysis that matches the best known results for the non-adaptive variant: sublinear convergence for general convex functions and linear convergence under strong convexity of the smooth term and smoothness of one of the proximal terms. Finally, an empirical comparison with related methods on 6 different problems illustrates the computational advantage of the proposed method.

1 Introduction

Minimizing the sum of a smooth and a non-smooth term is at the core of many optimization problems that arise in machine learning and signal processing (Rudin et al., 1992; Candès et al., 2006; Chambolle & Pock, 2016). In a few but important cases, such as ℓ_1 or group lasso regularization, the non-smooth term is simple enough so that its proximal operator is available in closed form or at least fast to compute. In this case, highly scalable methods such as proximal gradient descent (Beck & Teboulle, 2009; Nesterov et al., 2013) or proximal coordinate descent (Richtárik & Takáč, 2014) have shown state of the art performance. However,

the desire to model increasingly complex phenomena has led to the development of a flurry of penalties with costly to compute proximal operator. Examples are the overlapping group lasso (Jacob et al., 2009), multidimensional total variation (Barbero & Sra, 2014) or trend filtering (Kim et al., 2009), to name a few.

A key observation is that, despite the difficulty in computing its proximal operator, many of these penalties can be decomposed as a sum of terms for which we have access to their proximal operator. Proximal splitting methods like the three operator splitting (Davis & Yin, 2017) offer a principled way to incorporate these penalties into the optimizer. In this work we will describe a method to solve optimization problems of the form

$$\underset{\boldsymbol{x} \in \mathbb{R}^p}{\text{minimize}}\; f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x})\,, \qquad \text{(OPT)}$$

where f is convex and L_f-smooth (i.e., differentiable with L_f-Lipschitz gradient) and g, h are both convex but potentially non-smooth. We further assume g and h are *proximal*, i.e., we have access to the proximal operator.

This formulation allows to express a broad range of problems arising in machine learning and signal processing: the smooth term includes the least squares or logistic loss functions; the two proximal terms can be extended to an arbitrary number via a product space formulation and as we will see in §4.1 include many important penalties such as the group lasso with overlap, total variation, ℓ_1 trend filtering, etc. Furthermore, the penalties can be extended-valued, thus allowing an intersection for convex constraints through the use of the indicator function.

The three operator splitting (TOS) method (Davis & Yin, 2017) is a recently proposed method for problems of the form (OPT). At each iteration, it only requires to evaluate once the gradient of f and the proximal operator of g and h. It also relies on one step-size parameter, and while it can be set based on the Lipschitz constant of the gradient of f, this is not entirely satisfactory for two reasons. First, this constant is often costly to compute. Second, this constant is a global upper bound on the Lipschitz constant, while locally the Lipschitz constant might be smaller, allowing for larger step-sizes.

[1]University of California at Berkeley, USA [2]Department of Computer Science, ETH Zurich, Switzerland [3]Mila - Université de Montréal, Canada. Correspondence to: Fabian Pedregosa <f@bianp.net>.

Adaptive step-size methods, also known as inexact and backtracking line search, instead choose the step-size by verifying a sufficient decrease condition at each iteration. This allows to take larger step-sizes and has proven to be an important ingredient in the practical implementation of first and second-order methods (Nocedal & Wright, 2006).

Outline and main contributions. Our main contribution is the development and analysis of an adaptive variant of the TOS algorithm. The proposed algorithm does not depend on any step-size hyperparameter (besides an initial estimate) and enjoys similar convergence guarantees as the non adaptive variant. The paper is organized as follows:

- *Methods*. §2 describes the proposed algorithm, extended in §2.1 to an arbitrary number of proximal terms.

- *Analysis*. §3 provides a convergence analysis based on an interpretation of the algorithm as a saddle-point optimization method. This significantly departs from the analysis of Davis & Yin (2017) for the non adaptive variant and results in improved and more general rates.

- *Applications*. §4 discusses the application to different penalties and presents an empirical comparison on 6 different problems and 5 different penalties.

Notation. We denote vectors with boldface lower case letters (i.e., $\boldsymbol{x}$), and matrices and vector-valued functions in boldface upper case (i.e., $\boldsymbol{X}, \boldsymbol{T}(\cdot)$). $\|\cdot\|$ denotes the euclidean vector norm. Given a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, we denote by $\overline{\boldsymbol{X}}$ the average along rows, that is, $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$. We make extensive use of the proximal operator, defined for a convex function φ and $\gamma > 0$ as

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

The domain of a function $f : \mathbb{R}^p \to]-\infty, \infty]$ is $\operatorname{dom} f \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \mathbb{R}^p \mid f(\boldsymbol{x}) < \infty\}$. The indicator function is denoted $\imath\{\text{condition}\}$, which is 0 if condition is verified and $+\infty$ otherwise. Basic properties and definitions of convex functions are provided for convenience in Appendix A.

1.1 Related work

Proximal splitting methods that can solve problems involving a sum of terms by accessing the proximal operators of their constituents can be traced back to the 1970s in the works of Glowinski & Marroco (1975); Gabay & Mercier (1976); Lions & Mercier (1979). There has been a surge in interest in these methods in the last years due to their applicability in machine learning (Parikh & Boyd, 2013), signal processing (Combettes & Pesquet, 2011) and parallel optimization (Boyd et al., 2011).

Algorithms to solve problems 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 generalized forward-backward splitting (Raguet et al., 2013), the three operator splitting (TOS) (Davis & Yin, 2017), the primal-dual hybrid gradient (PDHG) method, proposed in (Condat, 2013b; Vũ, 2013) and analyzed by Chambolle & Pock (2015) and the very recent primal-dual three operator splitting (Yan, 2018). We note that the last two methods can optimize a more general objective function in which $h(\boldsymbol{x})$ is replaced with $h(\boldsymbol{K}\boldsymbol{x})$ for an arbitrary matrix $\boldsymbol{K}$. The original formulation of these methods requires to set the step-size based on criteria such as the Lipschitz constant of the gradient of the smooth term, but variants with adaptive step-size have recently emerged.

An adaptive step-size variant of the PDHG algorithm has recently been proposed by Malitsky & Pock (2018, §5). Compared to the proposed method, it requires one less function evaluation per iteration but since the original algorithm has two step-sizes, it still relies on one step-size hyperparameter. Convergence rates are not derived.

A different adaptive step-size strategy was proposed by Giselsson et al. (2016) as a general scheme for averaged operators. TOS is averaged for step-sizes $< 2/L_f$, and we denote the combination of both methods LSAO-TOS. An $\mathcal{O}(1/\sqrt{t})$ convergence rate in terms of the operator residual norm is derived. Unfortunately, this quantity is difficult to relate to the more common objective function suboptimality used in the other contributions.

Another adaptive step-size variant of TOS was proposed without proof in the technical report Davis & Yin (2015, Algorithm 3). It uses the same sufficient decrease inequality as our method, although the iterates are defined differently. We found the algorithm sometimes non-convergent and did not consider it further.

In contrast, we provide a convergence analysis for our method that achieves a $\mathcal{O}(1/t)$ convergence rate for the ergodic (i.e., averaged) iterate, and linear convergence under stronger assumptions, matching and in some cases even improving the best known rates of the non adaptive variant.

Method	Adaptive	Sublinear rate	Linear rate
Adaptive TOS (*this work*)	✓	✓	✓
TOS (Davis & Yin, 2017)	✗	✓	✓
LSAO-TOS (Giselsson et al., 2016)	✓	✓[1]	✗
PDHG (Condat, 2013b; Vũ, 2013)	✗	✓	✗
PDHG-LS (Malitsky & Pock, 2018)	✓	✗	✗

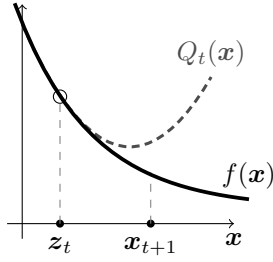[1]Convergence rate in terms of operator residuals.

2 Methods

In this section we present our main contribution, a three operator splitting method with adaptive step-size. The method is detailed in Algorithm 1 and requires at each iteration to evaluate once the gradient of f and the proximal operators of g and h, and perform two function evaluations of f. At iteration t the candidate step-size γ_t is chosen as to verify the following *sufficient decrease* condition between the iterates z_t and x_{t+1} (Line 4):

$$f(x_{t+1}) \leq Q_t(x_{t+1}, \gamma_t), \text{ with } Q_t \text{ defined as}$$

$$Q_t(x, \gamma) \stackrel{\text{def}}{=} f(z_t) + \langle \nabla f(z_t), x - z_t \rangle + \frac{1}{2\gamma} \|x - z_t\|^2.$$

This inequality can be interpreted as a quadratic upper bound condition on f at x_{t+1}: the right-hand side is a quadratic Q_t which is tangent to f at z_t with amplitude $(2\gamma_t)^{-1}$, and both sides are evaluated at x_{t+1}, defined in Line 3. The underlying principle of choosing the step-size based on the minimization of a quadratic upper bound has already been successful for the proximal-gradient method, where it is also referred to as backtracking (Beck & Teboulle, 2009) or full relaxation (Nesterov et al., 2013). In fact, the proposed method coincides with the aforementioned when one of the proximal terms is constant.

By the properties of L_f-smooth functions, the sufficient decrease condition is verified for any $\gamma_t \leq 1/L_f$. Hence the step-size search loop always has a finite termination and the step-size is lower bounded by $\gamma_t \geq \min\{\tau/L_f, \gamma_0\}$. The practical advantage of this strategy is that it allows to consider a step-size potentially larger than $1/L_f$ and verify whether the above is verified at each iteration. If it is, then the algorithm uses the current step-size, and if not, it decreases the step-size by a factor which we denote τ.

Growing step-size strategies. We consider two different strategies to initialize next iterate step-size. The first strategy (*Variant 1*) is the simplest and consists in initializing the next step-size with the current one (Line 12). In this variant, the step-size is only allowed to decrease.

The second strategy (*Variant 2*) allows the step-size to increase but in exchange requires the proximal term h to be Lipschitz continuous (note, not smooth as f but only Lipschitz). This is the case of most penalties (i.e., ℓ_1, group lasso, total variation, etc.) but not of indicator functions and so is less general than the first variant. As we will see in the applications section, the ability to grow the step-size has an important effect on its empirical performance.

Algorithm 1: Adaptive Three Operator Splitting

Input: $z_0 \in \mathbb{R}^p$, $u_0 \in \mathbb{R}^p$, $\gamma_0 > 0$, $\tau \in (0,1)$

1 **for** $t = 0, 1, 2, \ldots$ **do**
2   **repeat** ▷ step-size search loop
3          $x_{t+1} = \mathbf{prox}_{\gamma_t g}(z_t - \gamma_t u_t - \gamma_t \nabla f(z_t))$
4   **if** $f(x_{t+1}) \leq Q_t(x_{t+1}, \gamma_t)$ **then**
5   **break** ▷ sufficient decrease verified
6   **else**
7              $\gamma_t = \tau \gamma_t$ ▷ decrease step-size
8      $z_{t+1} = \mathbf{prox}_{\gamma_t h}(x_{t+1} + \gamma_t u_t)$
9      $u_{t+1} = u_t + (x_{t+1} - z_{t+1})/\gamma_t$
10   ▷ choose step-size for next iteration, two variants
11   **Variant 1**
12         $\gamma_{t+1} = \gamma_t$
13   **Variant 2** ▷ only if h is β_h-Lipschitz
14         $\delta_t = Q_t(x_{t+1}, \gamma_t) - f(x_{t+1})$
15   Choose any $\gamma_{t+1} \in [\gamma_t, \sqrt{\gamma_t^2 + \gamma_t \delta_t (2\beta_h)^{-2}}]$
16 **return** x_{t+1}, u_{t+1}

Initial and default values. The proposed method takes as input 4 parameters, which we briefly discuss, together with a growing step-size heuristic for Variant 2:

- *Initial guess z_0 and u_0.* z_0 is an initial guess of the primal problem (OPT), while u_0 is an initial guess for a minimizer of a (yet to be defined) dual function (11). In practice, we initialize both variables to zero.

- *Initial step-size γ_0.* To estimate a starting value for the step-size, we start with $\varepsilon = 10^{-3}$, $\tilde{z} = z_0 - \varepsilon \nabla f(z_0)$ and divide ε by 10 until $f(\tilde{z}) \leq f(z_0)$. Then we solve $f(\tilde{z}) = Q_0(\tilde{z})$ for γ_0 and double that estimate, giving

$$\gamma_0 = 4(f(z_0) - f(\tilde{z}))\|\nabla f(z_0)\|^{-2}. \tag{2}$$

- The *line search decrease parameter τ* regulates the factor by which the step-size is decreased each time the line search condition is unsuccessful. This is a parameter that is common to all line search methods and can be set to any value $\tau \in (0,1)$. Following (Malitsky & Pock, 2018) we set it to $\tau = 0.7$.

- *step-size growth.* Variant 2 allows the step-size to grow by an amount that depends on β_h^{-2}. This quantity can be arbitrarily large (e.g., vanishing regularization), and so choosing the largest admissible step-size might result in too many decrease corrections. This can be avoided e.g. by limiting its growth to double every 20 iterations. Line 15 then becomes:

$$\gamma_{t+1} = \min\{\gamma_t 2^{0.05}, \sqrt{\gamma_t^2 + \gamma_t \delta_t (2\beta_h)^{-2}}\}. \tag{3}$$

Upon termination, the algorithm returns two vectors. The first vector is an approximate solution to (OPT), while the second vector is an approximate minimizer of a dual objective which we will detail in §3.

Special cases and related methods. We mention two notable special cases of this algorithm. First, for any step-size $\gamma_t \leq 1/L_f$, the line search condition will always succeed by the properties of L_f-smooth functions and so the step-size in Variant 1 is constant. Defining $\boldsymbol{y}_t = \boldsymbol{x}_t + \gamma_t \boldsymbol{u}_{t-1}$, it is easy to verify that Algorithm 1 (Variant 1) can be written with a constant step-size $\gamma = \gamma_t$ as an iteration of the form

$$\begin{aligned} \boldsymbol{z}_t &= \mathbf{prox}_{\gamma h}(\boldsymbol{y}_t) \\ \boldsymbol{x}_{t+1} &= \mathbf{prox}_{\gamma g}(2\boldsymbol{z}_t - \boldsymbol{y}_t - \gamma \nabla f(\boldsymbol{z}_t)) \quad (4) \\ \boldsymbol{y}_{t+1} &= \boldsymbol{y}_t - \boldsymbol{z}_t + \boldsymbol{x}_{t+1} \,, \end{aligned}$$

which is the standard (non-overrelaxed) form of the three operator splitting (Davis & Yin, 2017, Algorithm 1). The adaptive variant requires two extra function evaluations $f(\boldsymbol{z}_t)$ and $f(\boldsymbol{x}_{t+1})$ for the line search condition in Line 4, but as we will see in the experimental section, most often the ability to take larger step outweighs this extra cost.

Second, for $h = 0$, we have from lines 8 and 9 that $\boldsymbol{u}_t = 0$ and in this case (ignoring growing step-size strategies), this algorithm simplifies to the proximal gradient descent with line search of (Beck & Teboulle, 2009).

Algorithm 1 can be written equivalently in a way that highlights similarities and differences with the PDHG method. Using Moreau's decomposition $\mathbf{prox}_{\gamma h}(\boldsymbol{x}) = \boldsymbol{x} - \gamma \mathbf{prox}_{\gamma h^*}(\boldsymbol{x}/\gamma)$ yields the following recurrence

$$\begin{aligned} \boldsymbol{u}_{t+1} &= \mathbf{prox}_{h^*/\gamma}(\boldsymbol{u}_t + \boldsymbol{x}_t/\gamma) \,, \quad (5) \\ \boldsymbol{x}_{t+2} &= \mathbf{prox}_{\gamma g}(\boldsymbol{x}_{t+1} - \gamma(\nabla f(\boldsymbol{z}_{t+1}) + 2\boldsymbol{u}_{t+1} - \boldsymbol{u}_t)) \,. \end{aligned}$$

This form is almost identical to Algorithm 3.2 in (Condat, 2013b), but with a different step-size and the gradient evaluated at the extrapolated $\boldsymbol{z}_{t+1} = \boldsymbol{x}_{t+1} - \gamma(\boldsymbol{u}_{t+1} - \boldsymbol{u}_t)$ instead of the previous iterate $\boldsymbol{x}_{t+1}$ in PDHG.

2.1 Extension to multiple proximal terms

We now consider the problem of minimizing an objective of the form:

$$\underset{\boldsymbol{x} \in \mathbb{R}^p}{\text{minimize}} \ \varphi(\boldsymbol{x}) + \sum_{j=1}^k h_j(\boldsymbol{x}) \,, \quad \text{(OPT-}k\text{)}$$

where φ is L_φ-smooth and each h_j is proximal. The adaptive three operator splitting can be used to solve problems of this form by reducing them to a problem of the form (OPT) in an enlarged space. Consider consider the

following problem in $\mathbb{R}^{k \times p}$,

$$\underset{\boldsymbol{X} \in \mathbb{R}^{k \times p}}{\text{minimize}} \ \underbrace{\varphi(\overline{\boldsymbol{X}})}_{=f(\boldsymbol{X})} + \underbrace{\sum_{j=1}^k h_j(\boldsymbol{X}_j)}_{=h(\boldsymbol{X})} + \underbrace{\iota\{\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_k\}}_{=g(\boldsymbol{X})} \,.$$

It is easy to see that this problem shares the same set of solutions as (OPT-k) with the correspondence $\boldsymbol{x} = \overline{\boldsymbol{X}}$, as the last term forces all the $\boldsymbol{X}_i$ terms to be equal. In this formulation the first term is smooth, the second term is proximal (variables in h_i are separated) and the proximal operator of the last term is given by $\overline{\boldsymbol{X}}\mathbf{1}^T$. Hence Algorithm 1 can be applied to this problem. Deriving the complete algorithm is now merely a matter of replacing f, g, h by its appropriate values in Algorithm 1 and is specified in Appendix B. The resulting adaptive algorithm seems to be new also in this extended formulation.

It is also possible to swap the definitions of g and h, which results in a different algorithm that can be seen as an adaptive variant of the Generalized Forward-Backward splitting of Raguet et al. (2013). However, this formulation is less convenient for our purpose, since in this case the h term is always an indicator function and so it would not be possible to apply variant 2 of our algorithm.

3 Analysis

In this section we provide a convergence rate analysis of the proposed method. We start by a characterization the set of fixed points of the algorithm, followed by a discussion on the gap function used to measure suboptimality. Finally, we present convergence rates for two different function classes. All proofs can be found in Appendix C.

Assumption 1: Regularity. We assume that f is convex and L_f-smooth in $\mathbb{R}^p$ and that g and h are proper (i.e., have nonempty domain), lower semicontinuous (i.e., its sublevel sets are closed) convex functions. We note that lower semicontinuity is a weak form of continuity that allows extended-valued functions (such as the indicator function) over a closed domain.

Assumption 2: Qualification conditions. We assume the relative interior of $\text{dom}\, g$ and $\text{dom}\, h$ have a non-empty intersection. This is a weak and standard assumption that, together with the regularity assumption, guarantees strong (also known as total) duality (Bertsekas, 2015, Prop. 5.3.8).

Using the definition of Fenchel conjugate, we can can reformulate (OPT) as a saddle-point problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x}) \quad (6)$$

$$= \min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) + g(\boldsymbol{x}) + \max_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \langle \boldsymbol{x}, \boldsymbol{u} \rangle - h^*(\boldsymbol{u}) \right\} \quad (7)$$

$$= \min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{u} \in \mathbb{R}^d} \underbrace{f(\boldsymbol{x}) + g(\boldsymbol{x}) + \langle \boldsymbol{x}, \boldsymbol{u} \rangle - h^*(\boldsymbol{u})}_{\stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u})} \,. \quad (8)$$

We recall that a saddle point of $\mathcal{L}$ is a pair $(\boldsymbol{x}^\star, \boldsymbol{u}^\star)$ such that the following is verified for any $(\boldsymbol{x}, \boldsymbol{u})$ in the domain (Hiriart-Urruty & Lemaréchal, 1993, §4.1):

$$\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{u}) \leq \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}^\star) . \tag{9}$$

A consequence of strong duality is the equivalence between the saddle points of $\mathcal{L}$ and the minimizers of the primal and dual objectives. Let P and D denote these primal and dual objectives:

$$P(\boldsymbol{x}) \overset{\text{def}}{=} f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x}) \tag{10}$$

$$D(\boldsymbol{u}) \overset{\text{def}}{=} (f + g)^*(-\boldsymbol{u}) + h^*(\boldsymbol{u}). \tag{11}$$

Then if $(\boldsymbol{x}^\star, \boldsymbol{u}^\star)$ is a saddle point of $\mathcal{L}$, $\boldsymbol{x}^\star$ is a minimizer of P and $\boldsymbol{u}^\star$ is a minimizer of D. Likewise, a pair of minimizers of P and D form a saddle point of $\mathcal{L}$.

3.1 Fixed point characterization

A common first step in the analysis of optimization methods is the study of its set of fixed or stationary points. While this does not necessarily imply convergence, knowing which elements will be left invariant by the method improves our understanding and is a stepping stone for further analysis. We will show that the set of fixed points of the algorithm has a particularly simple and elegant structure: the Cartesian product of primal and dual solutions.

For the purpose of analysis it will be useful to express Algorithm 1 as an iteration of the form, $(\boldsymbol{z}_{t+1}, \boldsymbol{u}_{t+1}) = \boldsymbol{T}_{\gamma_t}(\boldsymbol{z}_t, \boldsymbol{u}_t)$, where the operator $\boldsymbol{T}_\gamma$ is defined as

$$\boldsymbol{T}_\gamma(\boldsymbol{z}, \boldsymbol{u}) \overset{\text{def}}{=} (\boldsymbol{z}^+, \boldsymbol{u}^+), \text{ with} \tag{12}$$

$$\begin{cases} \boldsymbol{z}^+ = \mathbf{prox}_{\gamma h}(\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) + \gamma \boldsymbol{u}) \\ \boldsymbol{u}^+ = \boldsymbol{u} + (\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) - \boldsymbol{z}^+)/\gamma \\ \boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) = \mathbf{prox}_{\gamma g}(\boldsymbol{z} - \gamma(\boldsymbol{u} + \nabla f(\boldsymbol{z}))) . \end{cases}$$

The following theorem characterizes the set of fixed points of this operator, denoted $\mathrm{Fix}(\boldsymbol{T}_\gamma)$.

Theorem 1. *Let $\mathcal{P}^\star$ denote the set of minimizers of the primal objective P and $\mathcal{D}^\star$ the set of minimizers of the dual objective D. Then the fixed points of $\boldsymbol{T}_\gamma$ are given by*

$$\mathrm{Fix}(\boldsymbol{T}_\gamma) = \mathcal{P}^\star \times \mathcal{D}^\star . \tag{13}$$

3.2 Gap function

The progress of optimization methods is commonly measured in terms of a gap or merit function that is zero at optimum and nonzero otherwise. An appropriate gap function for many first-order methods is the suboptimality of the objective function. However, this is not an appropriate suboptimality measure for this algorithm, as the objective

function might be $+\infty$ at an iterate, for example when the two proximal terms are an indicator function.

Davis & Yin (2015) avoid the issue by either evaluating h and g at different iterates (Davis & Yin, 2015, Corollary D.5.1) or assuming Lipschitz continuity of one of the proximal terms (Davis & Yin, 2015, Corollary D.5.2).

In this work we take an alternative approach, and instead use the following *saddle point suboptimality* criterion to measure the progress of our algorithm:

$$\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1}) . \tag{14}$$

From the definition of saddle point in Eq. (9), this criterion is non-positive for all $(\boldsymbol{x}, \boldsymbol{u})$ if and only if $(\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1})$ is a saddle point, and is so an appropriate suboptimality criterion. Furthermore, contrary to the primal objective function, this is defined for all iterates without further assumptions. Finally, we mention that this criteria has been previously used in the analysis of primal-dual methods, see e.g., Chambolle & Pock (2016; 2015) and Gidel et al. (2017) for a discussion of saddle point gap functions.

This suboptimality criteria can also be related to the primal and dual gap, as minimizing (14) over $\boldsymbol{x}$ and maximizing over $\boldsymbol{u}$ one recovers the primal-dual gap $P(\boldsymbol{x}_t) - D(\boldsymbol{u}_t)$ by definition of Fenchel conjugate.

3.3 Sublinear convergence

The following theorem gives a sublinear convergence rate for Algorithm 1. This convergence will be given in terms of the weighted ergodic (i.e., averaged) sequence. Denoting by s_t the sum of all step-sizes up to iteration t, i.e., $s_t \overset{\text{def}}{=} \sum_{i=0}^{t-1} \gamma_t$, the ergodic iterates $\overline{\boldsymbol{x}}_t$ and $\overline{\boldsymbol{u}}_t$ are defined as

$$\overline{\boldsymbol{x}}_t \overset{\text{def}}{=} \left(\sum_{i=0}^{t-1} \gamma_i \boldsymbol{x}_{i+1}\right)/s_t , \quad \overline{\boldsymbol{u}}_t \overset{\text{def}}{=} \left(\sum_{i=0}^{t-1} \gamma_i \boldsymbol{u}_{i+1}\right)/s_t . \tag{15}$$

While results in this subsection will be stated in terms of this ergodic sequence, in practice the last iterate gives most often a better empirical convergence, see e.g., (Chambolle & Pock, 2015, §7.2.1) for a discussion of this phenomenon. For a more theoretically-sound algorithm, one can compare the objective at the ergodic and last iterate, and return the one with smallest objective.

Theorem 2 (sublinear convergence rate). *For every $t \geq 0$ and any $(\boldsymbol{x}, \boldsymbol{u})$ in the domain of $\mathcal{L}$ we have the following convergence rate for Algorithm 1 (both variants):*

$$\mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \overline{\boldsymbol{u}}_{t+1}) \leq \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \gamma_0^2 \|\boldsymbol{u}_0 - \boldsymbol{u}\|^2}{2s_t} .$$

Convergence in terms of function value suboptimality. The previous result gives an $\mathcal{O}(1/t)$ convergence rate for

arbitrary convex functions in terms of the saddle point sub-optimality. As we have discussed previously, it is not possible to obtain similar rates in terms of the function suboptimality without further assumptions. We will now show that it is sufficient to assume Lipschitz continuity on h to derive from the previous theorem a convergence rate in terms of the primal function suboptimality.

The following Corollary can be obtained by optimizing with respect to $\boldsymbol{u}$ the bound in the previous theorem and using the Lipschitz continuity to bound $\|\boldsymbol{u}_0 - \boldsymbol{u}\|^2$. This gives an $\mathcal{O}(1/t)$ convergence rate for the primal function suboptimality, roughly matching that of Davis & Yin (2015, Corollary D.5.2) for the non adaptive variant:

Corollary 1. *Let h be β_h-Lipschitz. Then, we have the following rate for the weighted ergodic iterate on the objective* $P(\boldsymbol{x}) \overset{def}{=} f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x})$:

$$P(\overline{\boldsymbol{x}}_{t+1}) - P(\boldsymbol{x}^\star) \leq \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + 2\gamma_0^2(\|\boldsymbol{u}_0\|^2 + \beta_h^2)}{2s_t} \, .$$

3.4 Linear convergence

In this subsection we assume that f is μ_f-strongly convex and h is L_h-smooth (with $0 < \mu_f, 0 < L_h < +\infty$). We denote by $\boldsymbol{x}^\star$ the minimizer of the primal loss (unique by strong convexity of P) and by $\boldsymbol{u}^\star$ the minimizer of the dual loss (also unique by strong convexity of D, consequence of the L_h-smoothness of h).

The convergence rates will be given in terms of the following quantities

$$\rho \overset{def}{=} \mu_f \min\{\gamma_0, \tau/L_f\} \, , \quad \sigma \overset{def}{=} 1/(1 + \gamma_0 L_h) \\ \xi \overset{def}{=} \mu_f/(\mu_f + L_h) \, . \quad (16)$$

All these belong to the interval $(0, 1)$. Assuming $\gamma_0 \geq \tau/L_f$, then ρ is the inverse of f's condition number, while σ depends the smoothness of h. ξ is only used by variant 2 and is a less tight bound that σ that depends on both the strong convexity of f and smoothness of h. Note that by strong convexity, $\gamma_0 < 1/\mu_f$ as otherwise the sufficient decrease condition would not succeed and so $\sigma \geq \xi$.

Theorem 3 (linear convergence rate). *Let $\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1}$ be the iterates produced by Algorithm 1 after t iterations. Then we have the following linear convergence for Variant 1 (V1) and Variant 2 (V2):*

$$V1 : \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \min\{\rho, \sigma\}\right)^{t+1} D_0 \quad (17)$$

$$V2 : \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \min\{\rho, \xi, \tfrac{1}{2}\}\right)^{t+1} E_0 \, , \quad (18)$$

with $D_0 \overset{def}{=} 6\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{6}{1-\sigma}\|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u}^\star)\|^2$ *and* $E_0 \overset{def}{=} 6\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{6}{1-\xi}\|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u}^\star)\|^2$.

Discussion. For $\gamma_t = 1/L_f$, the sufficient decrease condition is always verified and the algorithm can be run with $\tau = 1$. In this case, Variant 1 of Algorithm 1 defaults to TOS, and we can compare the obtained rates with those in (Davis & Yin, 2015).

While the sublinear convergence rate obtained in Corollary 1 roughly matches the rate obtained in (Davis & Yin (2015, Corollary D.5.2, see our Appendix C.5)), the linear convergence rates are instead significantly different. The linear convergence rate obtained in (Davis & Yin, 2015, Theorem D.6.6), after optimizing for all parameters, yields a rate of $\rho\sigma^2$, which is *strictly worse* than the $\min\{\rho, \sigma\}$ rate that we obtained. This difference can be quite large, e.g., for $\rho = \sigma$ this becomes ρ versus ρ^3.

Finally, we note that the number of evaluations of the sufficient decrease condition can be bounded as in (Nesterov et al., 2013, Lemma 4).

4 Applications

4.1 Learning with Multiple Penalties

In this subsection we discuss how some penalties with costly to compute proximal operator can be decomposed as a sum of proximal terms and so fall within the current framework. The exact expression of the proximal operators is given in Appendix D.

Group lasso with overlap. Jacob et al. (2009) generalized group ℓ_1 norm by allowing each variable to belong to more than one group, thereby introducing overlaps among groups and allowing for more complex prior knowledge on the structure. For a set of subindices $\mathcal{G}$ which we will call groups, this penalty is defined as $\|\boldsymbol{x}\|_{\mathcal{G}} = \sum_{G \in \mathcal{G}} \|[\boldsymbol{x}]_G\|_2$. If each coefficient is at most in s groups, then $\mathcal{G}$ can be decomposed as $\mathcal{G} = \mathcal{G}_1 \cup \ldots \cup \mathcal{G}_s$, where the $\mathcal{G}_i$ are disjoint. This allows to express the group lasso with overlap as a sum of s non-overlapping group lasso penalties, for which the proximal operator has a closed form expression.

Multidimensional total variation. For the task of image restoration and denoising it is common to consider a regularization term in the form of a total variation regularizer. For an image $\boldsymbol{x}$ of size $p \times q$, the 2-dimensional total variation norm $\|\boldsymbol{X}\|_{\mathrm{TV}}$ is defined as

$$\underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q-1}|\boldsymbol{X}_{i,j+1} - \boldsymbol{X}_{i,j}|}_{=g(\boldsymbol{X})} + \underbrace{\sum_{j=1}^{q}\sum_{ji=1}^{p-1}|\boldsymbol{X}_{i+1,j} - \boldsymbol{X}_{i,j}|}_{=h(\boldsymbol{X})} \, .$$

From here we recognize that g and h are fused lasso (also known as 1D-total variation) penalties acting on the columns and rows of $\boldsymbol{X}$ respectively. Efficient methods to

evaluate the proximal operator of the fused lasso penalty have been developed by Condat (2013a); Johnson (2013).

Isotonic and nearly isotonic penalties. In some applications there exists a natural ordering between variables: $\boldsymbol{x}_1 \leq \boldsymbol{x}_2 \leq \cdots \leq \boldsymbol{x}_p$. This can be enforced through constraints, and the projection onto these is known as isotonic regression (Best & Chakravarti, 1990). The indicator function over the set of constraints can also be split into a sum of two proximal terms (see Appendix D.2) as

$$\imath\{\boldsymbol{x}_1 \leq \boldsymbol{x}_2 \leq \boldsymbol{x}_3 \leq \boldsymbol{x}_4 \leq \cdots\} \qquad (19)$$
$$= \underbrace{\imath\{\boldsymbol{x}_1 \leq \boldsymbol{x}_2; \boldsymbol{x}_3 \leq \boldsymbol{x}_4; \cdots\}}_{=g(\boldsymbol{x})} + \underbrace{\imath\{\boldsymbol{x}_2 \leq \boldsymbol{x}_3; \boldsymbol{x}_4 \leq \boldsymbol{x}_5; \cdots\}}_{=h(\boldsymbol{x})} .$$

In cases in which the variables are only "mostly" non-decreasing, the constraint can be relaxed via a nearly-isotonic penalty (Tibshirani et al., 2011) of the form $\sum_{i=1}^{p-1} \max\{\boldsymbol{x}_i - \boldsymbol{x}_{i+1}, 0\}$, in which only the non-increasing coefficients are penalized. This penalty can be split the same way as the isotonic constraints above.

ℓ_1 trend filtering. This penalty is defined by the absolute value of the second order differences and promotes piecewise-linear coefficients (Kim et al., 2009). It is defined as $\|\boldsymbol{x}\|_{\text{TF}} \overset{\text{def}}{=} \sum_{i=1}^{p-2} |\boldsymbol{x}_i - 2\boldsymbol{x}_{i+1} + \boldsymbol{x}_{i+2}|$. We can split this sum into 3 proximal terms such that the resulting terms: the j-th term contains the factors for which i is congruent to 3 modulo j.

Constraints over doubly stochastic matrices. Optimization problems with constraints on the set of doubly stochastic matrices appear in many convex relaxations of combinatorial problems such as seriation (Fogel et al., 2013), quadratic assignment (Lawler, 1963) and graph matching (Conte et al., 2004; Aflalo et al., 2015). The set of double stochastic matrices is composed of square matrices with nonnegative entries, each of whose rows and columns sum to 1, i.e., $\{\boldsymbol{X}^T\boldsymbol{1} = \boldsymbol{1}, \boldsymbol{X}\boldsymbol{1} = \boldsymbol{1}, \boldsymbol{X} \geq \boldsymbol{0}\}$. The indicator function over this set can be split as

$$\underbrace{\imath\{\boldsymbol{X}^T\boldsymbol{1} = \boldsymbol{1}, \boldsymbol{X}\boldsymbol{1} = \boldsymbol{1}\}}_{=g(\boldsymbol{X})} + \underbrace{\imath\{\boldsymbol{X} \geq \boldsymbol{0}\}}_{=h(\boldsymbol{X})} \quad , \qquad (20)$$

and the projection onto both sets is available in closed form (Lu et al., 2016, §4.3).

Dispersive sparsity. In some applications it is desirable to encourage dispersion of the sparse coefficients. This happens for example in the modeling of neural spiking, as the spikes are assumed to be spaced across time (Hegde et al., 2009). El Halabi & Cevher (2015) showed that this behavior can be promoted by considering a penalty of the form $\|\boldsymbol{x}\|_1 + \imath\{\boldsymbol{B}|\boldsymbol{x}| \leq \boldsymbol{c}\}$ for a matrix $\boldsymbol{B}$ and some predefined constant c, where $|\boldsymbol{x}|$ denotes the component-wise

absolute value. This penalty can be split into three proximal terms by the introduction of a dummy variable $\boldsymbol{z}$, resulting in $\|\boldsymbol{x}\|_1 + \imath\{\boldsymbol{B}\boldsymbol{z} \leq \boldsymbol{c}\} + \imath\{\boldsymbol{z} = |\boldsymbol{x}|\}$.

Combination by addition. A popular method to promote the joint behavior of different penalties is by adding them. This has been used to successfully learn models with sparse and nonnegative coefficients (Yuan & Lin, 2007), sparse and low rank matrices (Richard et al., 2012), sparse and piecewise constant (Gramfort et al., 2013), to name a few.

4.2 Benchmarks

In this subsection we provide an empirical evaluation of the proposed method. Due to space constraints we only give here a high level overview, deferring details as well as an extended set of experiments to Appendix E. We consider the following methods:

- The proposed Adaptive TOS method (Algorithm 1), in its both variants.

- The TOS method of Davis & Yin (2015), with step-sizes $1/L_f$ and $1.99/L_f$ (the method is convergent for step-sizes $< 2/L_f$).

- The PDHG or Condat-Vũ algorithm (Condat, 2013b; Vũ, 2013), with step-sizes τ and β/τ, where β was chosen as the one giving the best overall performance over the grid $\beta = 0.9, 0.5, 0.1$ (giving it a slight advantage).

- The adaptive PDHG of Malitsky & Pock (2018), with step-size hyperparameter β chosen by the same technique as for PDHG.

- The averaged operator line search method of Giselsson et al. (2016) combined with TOS, named TOS-AOLS.

We compared these methods on 4 different problems and show the results in Figure 1. In the first row we show the benchmarks on a logistic regression problem with overlapping group lasso penalty that we apply to two text datasets (RCV1 and real-sim). Subfigures A and C were run with the regularization parameter chosen to give 50% of sparsity, while B, E are run with higher levels of sparsity, chosen to give 5% of sparsity.

In the second and third row we considered a battery of inverse problems with different penalties on synthetic datasets. These consists of a least squares (G, H, I, J) or logistic regression (rest) smooth term and 4 different penalties specified in the title of each plot (overlapping group lasso, total variation, trace norm ℓ_1 and nearly isotonic, see Appendix E for a precise formulation). For each problem, we show 2 different benchmarks, corresponding to the low and high regularization regimes (denoted low reg and high reg). We comment on a few trends from Fig. 1:
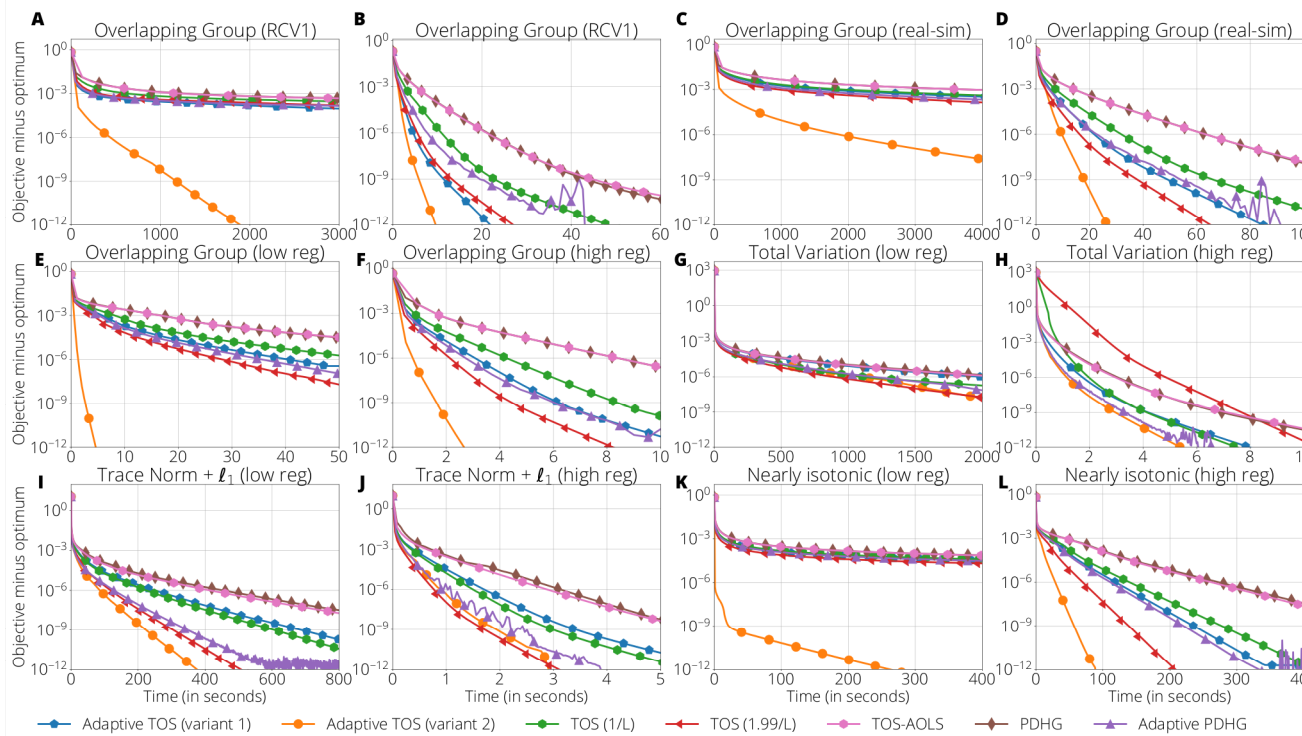
Figure 1: **Comparison of different proximal splitting methods**. The *top row* gives result for two real datasets with an overlapping group lasso penalty. The *second and third row* show results on synthetic datasets for 4 different penalties: overlapping group lasso (E, F), 2-dimensional total variation (G, H), trace norm + ℓ_1 (I, J) and nearly isotonic (K, L). The Adaptive TOS (Variant 2, i.e., with growing step-sizes) is the best performing method on 10 out of 12 experiments, and roughly equivalent to the best performing method in the other 2 cases (G, J).

- **Best performing method**. On 10 out of 12 experiments, the adaptive TOS algorithm (Variant 2) is the best performing method, and in the other cases (E, H) its performance is roughly the same as that of the best performing method. In contrast, on 3 instances (A, I, K) it is an order of magnitude faster than the next method.

- **Low vs high regularization regime.** The advantage of the adaptive method is highly correlated with the amount of regularization: in the low regularization regime, on 3 out of 6 the adaptive TOS is an order of magnitude faster then the fixed step-size method, while in the high regularization regime the difference shrinks and in the same problems is never more than a factor 2.6.

- **Uniform curvature**. The problems (G, H, I, J) in Fig. 1 use as smooth term a quadratic loss (i.e., constant Hessian), while the other methods use a logistic loss (non-constant Hessian). This suggests that the use of the adaptive step-size strategy (and in particular Variant 2 with its growing step-size) is more beneficial for smooth terms with non-uniform curvature.

5   Conclusion and Future Work

We have presented and analyzed an adaptive step-size method to solve optimization problems consisting in a sum of a smooth term accessed through its gradient and two or more potentially non-smooth terms accessed through their proximal operator. The method does not rely on any step-size hyperparameter (except for an initial estimate) and extensive empirical evaluation has showed computational gains on a variety of problems. We mention two possible extensions of this work.

First, existing convergence results fail to fully explain their surprisingly good empirical convergence. To the best of our knowledge, no work so far has derived linear convergence rates in absence of strong convexity and smoothness of one of the proximal terms for these methods (as is however empirically observed, see e.g. Figure 1).

Second, it is an open question whether this or other adaptive step-size methods can be accelerated, as is the case of proximal gradient descent, which admits the adaptive FISTA variant (Beck & Teboulle, 2009).

Acknowledgements

References

Aflalo, Y., Bronstein, A., and Kimmel, R. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 2015.

Agarwal, A., Negahban, S., and Wainwright, M. J. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, 2010.

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

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

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

Beck, A. and Teboulle, M. Gradient-based algorithms with applications to signal recovery. *Convex Optimization in Signal Processing and Communications*, 2009.

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

Best, M. J. and Chakravarti, N. Active set algorithms for isotonic regression; A unifying framework. *Mathematical Programming*, 1990.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011.

Cai, J.-F., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.

Candès, E. J., Romberg, J., and Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 2006.

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

Chambolle, A. and Pock, T. An introduction to continuous optimization for imaging. *Acta Numerica*, 2016.

Combettes, P. L. and Pesquet, J.-C. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011.

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

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

Conte, D., Foggia, P., Sansone, C., and Vento, M. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 2004.

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

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

El Halabi, M. and Cevher, V. A totally unimodular view of structured sparsity. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Fogel, F., Jenatton, R., Bach, F., and d'Aspremont, A. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, 2013.

Gabay, D. and Mercier, B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 1976.

Gidel, G., Jebara, T., and Lacoste-Julien, S. Frank-Wolfe Algorithms for Saddle Point Problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Giselsson, P., Fält, M., and Boyd, S. Line search for averaged operator iteration. In *IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016.

Glowinski, R. and Marroco, A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle*, 1975.

Gramfort, A., Thirion, B., and Varoquaux, G. Identifying predictive regions from fMRI with TV-L1 prior. In *International Workshop on Pattern Recognition in Neuroimaging*. IEEE, 2013.

Hegde, C., Duarte, M. F., and Cevher, V. Compressive sensing recovery of spike trains using a structured sparsity model. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms I: Fundamentals*. Springer science & business media, 1993.

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

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

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

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. ℓ_1 trend filtering. *SIAM review*, 2009.

Lawler, E. L. The quadratic assignment problem. *Management science*, 1963.

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

Lions, P.-L. and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 1979.

Lu, Y., Huang, K., and Liu, C.-L. A fast projected fixed-point algorithm for large graph matching. *Pattern Recognition*, 2016.

Malitsky, Y. and Pock, T. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 2018.

Nesterov, Y. et al. Gradient methods for minimizing composite objective function. *Mathematical Programming*, 2013.

Nocedal, J. and Wright, S. *Numerical optimization*. Springer Science & Business Media, 2006.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in optimization*, 2013.

Pedregosa, F. C-OPT: composite optimization in Python. 2018. doi: 10.5281/zenodo.1283339. URL http://openopt.github.io/copt/.

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

Richard, E., Savalle, P.-a., and Vayatis, N. Estimation of Simultaneously Sparse and Low Rank Matrices. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 2014.

Rockafellar, R. T. Convex analysis, 1997.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*. Springer Science & Business Media, 1998.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 1992.

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. Nearly-isotonic regression. *Technometrics*, 2011.

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

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

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006.

Yuan, M. and Lin, Y. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007.

Adaptive Three Operator Splitting

Supplementary material

Appendix A Basic definitions and properties

Definition 1 (proper function). *A function $f : \mathcal{X} \subseteq \mathbb{R}^p \to\ ]-\infty, \infty]$ is said to be proper if its domain not empty.*

Definition 2 (Fenchel conjugate). *The Fenchel conjugate of a function $f : \mathcal{X} \subseteq \mathbb{R}^p \to\ ]-\infty, \infty]$ is defined as*

$$f^*(\boldsymbol{x}^\star) = \sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{x}^\star, \boldsymbol{x} \rangle - f(\boldsymbol{x}) \, . \tag{21}$$

Definition 3 (lower semicontinuity). *We say that a proper convex function f is lower-semicontinuous if all of its levelsets $\{\boldsymbol{x} \in \mathrm{dom}(f) \,|\, f(\boldsymbol{x}) \leq \alpha\}$ are closed.*

Definition 4 (strong convexity). *A function f is said μ-strongly convex if it verifies the following inequality for all $\boldsymbol{x}, \boldsymbol{y}$ in the domain and any $\boldsymbol{u} \in \partial f(\boldsymbol{x})$*

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + \langle \boldsymbol{u}, \boldsymbol{x} - \boldsymbol{y} \rangle - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{22}$$

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

$$\mathrm{relint}(C) = \{\boldsymbol{x} \in C : \forall \boldsymbol{y} \in C \,\exists \lambda > 1 : \lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in C\} \tag{23}$$

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

$$\boldsymbol{z} = \mathbf{prox}_{\gamma g}(\boldsymbol{x}) \iff \frac{1}{\gamma}(\boldsymbol{x} - \boldsymbol{z}) \in \partial g(\boldsymbol{z}) \tag{24}$$

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

$$\boldsymbol{z} \in \underset{\boldsymbol{z}'}{\arg\min}\, g(\boldsymbol{z}') + \frac{1}{2\gamma}\|\boldsymbol{z}' - \boldsymbol{x}\|^2 \tag{25}$$

$$\iff 0 \in \partial g(\boldsymbol{z}) + \frac{1}{\gamma}(\boldsymbol{z} - \boldsymbol{x}) \tag{26}$$

$$\iff \frac{1}{\gamma}(\boldsymbol{x} - \boldsymbol{z}) \in \partial g(\boldsymbol{z}) \tag{27}$$

where the first equivalence is a consequence of the first order optimality conditions. $\qquad\square$

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

$$\boldsymbol{u} \in \partial h(\boldsymbol{z}) \iff \boldsymbol{z} \in \partial h^*(\boldsymbol{u}) \, . \tag{28}$$

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

Proof. See e.g. (Bauschke & Combettes, 2017, Corollary 16.30) or (Rockafellar & Wets, 1998, Proposition 11.3). $\qquad\square$

Lemma 3 (Improved Young's inequality)**.** *For all $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ we have,*

$$\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|^2 \leq 3(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + \|\boldsymbol{c}\|^2) \tag{29}$$

Proof. We develop the squared norm to get,

$$\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|^2 = \|\boldsymbol{a}\|^2 + 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle + \|\boldsymbol{b}\|^2 + 2\langle \boldsymbol{b}, \boldsymbol{c} \rangle + \|\boldsymbol{c}\|^2 + 2\langle \boldsymbol{a}, \boldsymbol{c} \rangle \tag{30}$$

$$\leq \|\boldsymbol{a}\|^2 + \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + \|\boldsymbol{b}\|^2 + \|\boldsymbol{b}\|^2 + \|\boldsymbol{c}\|^2 + \|\boldsymbol{c}\|^2 + \|\boldsymbol{a}\|^2 + \|\boldsymbol{c}\|^2 \tag{31}$$

$$\text{(Young's inequality on } 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle, 2\langle \boldsymbol{b}, \boldsymbol{c} \rangle \text{ and } 2\langle \boldsymbol{a}, \boldsymbol{c} \rangle)$$

$$= 3(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + \|\boldsymbol{c}\|^2) \tag{32}$$

$\square$

Appendix B   Adaptive (k+1) operator splitting

For completeness we write below the full adaptive $(k + 1)$ operator splitting, which is merely Algorithm 1 applied to the reformulation of (OPT-k) in §2.1. It hence solves an optimization problem of the form

$$\underset{\boldsymbol{x}\in\mathbb{R}^p}{\text{minimize}}\ \varphi(\boldsymbol{x}) + \sum_{j=1}^{k} h_j(\boldsymbol{x})\ , \tag{33}$$

where φ is L_f-smooth and each h_j is proximal.

In this case we name the quadratic used in the sufficient decrease condition $\widetilde{Q}_t$, and is defined as

$$\widetilde{Q}_t(\boldsymbol{x}) \overset{\text{def}}{=} \varphi(\overline{\boldsymbol{Z}}_t) + \langle \nabla\varphi(\overline{\boldsymbol{Z}}_t), \boldsymbol{x} - \overline{\boldsymbol{Z}}_t \rangle + \frac{1}{2\gamma_t}\|\boldsymbol{x}\mathbf{1}_k^T - \boldsymbol{Z}_t\|_F^2\ , \tag{34}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{1}_k$ is the k-dimensional vector of ones.

Algorithm 3: Adaptive $(k + 1)$-Operator Splitting

Input: $\boldsymbol{z}_0 \in \mathbb{R}^p, \boldsymbol{u}_0 \in \mathbb{R}^p, \gamma_0 > 0, \tau \in (0, 1)$

1   $\boldsymbol{U}_0 = \boldsymbol{u}_0 \mathbf{1}_k^T$
2   $\boldsymbol{Z}_0 = \boldsymbol{z}_0 \mathbf{1}_k^T$
3   **for** $t = 1, \ldots, T$ **do**
4   **repeat**   ▷ step-size search loop
5        $\boldsymbol{r}_t = \frac{1}{k}\nabla\varphi(\boldsymbol{z}_t)$
6        $\boldsymbol{x}_{t+1} = \overline{\boldsymbol{Z}}_t - \gamma_t\overline{\boldsymbol{U}}_t - \gamma_t\boldsymbol{r}_t$
7        $\delta_t = \widetilde{Q}_t(\boldsymbol{x}_{t+1}) - \varphi(\boldsymbol{x}_{t+1})$
8   **if** $\delta_t \geq 0$ **then**
9   **break**   ▷ sufficient decrease verified
10   **else**
11           $\gamma_t = \tau\gamma_t$ ▷ decrease step-size
12   **for** $j = 1, \ldots, k$ **do**
13        $\boldsymbol{Z}_{j,t+1} = \textbf{prox}_{\gamma h_j}(\boldsymbol{x}_{t+1} + \gamma_t\boldsymbol{U}_{j,t})$
14        $\boldsymbol{U}_{j,t+1} = \boldsymbol{U}_{j,t} + (\boldsymbol{x}_{t+1} - \boldsymbol{z}_{t+1})/\gamma_t$
15   **Variant 1**
16        $\gamma_{t+1} = \gamma_t$
17   **Variant 2**   ▷ only if each h_j is β_h-Lipschitz
18   Choose any $\gamma_{t+1} \in [\gamma_t, \sqrt{\gamma_t^2 + \gamma_t\delta_t(2\beta_h)^{-2}}]$
19 **return** $\boldsymbol{x}_{t+1}, \boldsymbol{U}_{t+1}$

Discussion. The above algorithm requires the storage of a vector of size p and two matrices of size $k \times p$. Hence, for three operators ($k = 2$) this requires to store $2p$ more elements than the three previous formulation, which only requires to store three arrays of size p. Hence, for $k = 2$, the formulation of Algorithm 1 should be preferred. Of course, for $k > 2$, it is necessary to use this formulation.

Appendix C   Analysis

In this section we provide the proofs for the theorems stated in the Analysis section without proof. The appendix is organized as follows:

- In Appendix C.1 we prove the fixed point characterization of our three operator splitting variant. Its proof is mostly independent of the other results.

- In Appendix C.2 we prove an inequality that relates the saddle point suboptimality to the current and previous iterates. This inequality forms the core of both linear and sublinear convergence proofs. Because of its importance we name it "Key recursive inequality".

- Finally, in Appendix C.3 and Appendix C.4 we make use of the previous "key recursive inequality" to prove the sublinear (Theorem 2 and Corollary 1) and linear (Theorem 3) convergence results.

Throughout this section we assume that Assumptions 1 and 2 of §3 are verified without explicit mention.

Appendix C.1   Fixed point characterization

In this subsection we provide a proof for Theorem 1. Davis & Yin (2017, Lemma 2.2) proved a weaker result that characterized only the first coordinate: in our notation they proved that if $(\boldsymbol{x}, \boldsymbol{u}) \in \mathrm{Fix}(\boldsymbol{T}_\gamma)$, then $\boldsymbol{x}$ is a primal solution. Our theorem extends this results into a full characterization of the operator.

Theorem 1. *Let $\mathcal{P}^\star$ denote the set of minimizers of the primal objective* (10) *and $\mathcal{D}^\star$ the set of minimizers of the dual objective* (11). *Then the set of fixed points of $\boldsymbol{T}_\gamma$ is given by*

$$\mathrm{Fix}(\boldsymbol{T}_\gamma) = \mathcal{P}^\star \times \mathcal{D}^\star . \tag{35}$$

Proof. We will find it useful to first characterize the fixed points of $\boldsymbol{T}_\gamma$ by a subdifferential inclusion. Consider the following set of equivalences

$$(\boldsymbol{z}, \boldsymbol{u}) \in \mathrm{Fix}(\boldsymbol{T}_\gamma) \iff (\boldsymbol{z}, \boldsymbol{u}) = \boldsymbol{T}_\gamma(\boldsymbol{z}, \boldsymbol{u}) \tag{36}$$

$$\iff \begin{cases} \boldsymbol{z} = \mathbf{prox}_{\gamma h}(\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) + \gamma \boldsymbol{u}) \\ \boldsymbol{u} = \boldsymbol{u} + (\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) - \boldsymbol{z})/\gamma \qquad\qquad \text{(by definition of } \boldsymbol{T}_\gamma) \\ \text{with } \boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) = \mathbf{prox}_{\gamma g}(\boldsymbol{z} - \gamma(\boldsymbol{u} + \nabla f(\boldsymbol{z}))) \end{cases} \tag{37}$$

$$\iff \begin{cases} \boldsymbol{z} = \mathbf{prox}_{\gamma g}(\boldsymbol{z} - \gamma(\boldsymbol{u} + \nabla f(\boldsymbol{z}))) \\ \boldsymbol{z} = \mathbf{prox}_{\gamma h}(\boldsymbol{z} + \gamma \boldsymbol{u}) \end{cases} \qquad (\boldsymbol{z} = \boldsymbol{x}(\boldsymbol{z}, \boldsymbol{u}) \text{ by second equation)} \tag{38}$$

$$\iff \begin{cases} -\boldsymbol{u} - \nabla f(\boldsymbol{z}) \in \partial g(\boldsymbol{z}) \\ \boldsymbol{u} \in \partial h(\boldsymbol{z}) \end{cases} \qquad\qquad\qquad\qquad \text{(by Lemma 1)} \tag{39}$$

$$\iff \begin{cases} -\boldsymbol{u} \in \partial(f + g)(\boldsymbol{z}) \\ \boldsymbol{u} \in \partial h(\boldsymbol{z}) \end{cases} \tag{40}$$

$$\iff \begin{cases} \boldsymbol{z} \in \partial(f + g)^*(-\boldsymbol{u}) \\ \boldsymbol{z} \in \partial h^*(\boldsymbol{u}) \end{cases} \qquad\qquad\qquad\qquad \text{(by Lemma 2)} \tag{41}$$

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

Part 1. Our goal is to prove $\mathrm{Fix}(\boldsymbol{T}_\gamma) \subseteq \mathcal{P}^\star \times \mathcal{D}^\star$. Let $(\boldsymbol{z}, \boldsymbol{u}) \in \mathrm{Fix}(\boldsymbol{T}_\gamma)$. Adding together the equations in (40) one obtains

$$0 \in \partial h(\boldsymbol{z}) + \partial g(\boldsymbol{z}) + \nabla f(\boldsymbol{z}) , \tag{42}$$

and so $\boldsymbol{z}$ is a minimizer of the primal objective. Subtracting the equations in (41) we also have

$$0 \in \partial h^*(\boldsymbol{u}) - \partial (f + g)^*(-\boldsymbol{u}) , \tag{43}$$

and so $\boldsymbol{u}$ is a minimizer of the dual objective. We have proved $\mathrm{Fix}(\boldsymbol{T}_\gamma) \subseteq \mathcal{P}^\star \times \mathcal{D}^\star$.

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

We start by recalling the notion of *paramonotonicity*, which will play a key role in this 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

$$\left.\begin{aligned} \widetilde{\boldsymbol{a}} &\in \boldsymbol{C}\boldsymbol{a} \\ \widetilde{\boldsymbol{b}} &\in \boldsymbol{C}\boldsymbol{b} \\ \langle \widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}, \boldsymbol{a} - \boldsymbol{b}\rangle &= 0 \end{aligned}\right\} \implies \widetilde{\boldsymbol{a}} \in \boldsymbol{C}\boldsymbol{b} \text{ and } \widetilde{\boldsymbol{b}} \in \boldsymbol{C}\boldsymbol{a} . \tag{44}$$

It is known 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 $\boldsymbol{u_z}$ and $\boldsymbol{z_u}$ such that

$$\boldsymbol{u_z} \in \partial h(\boldsymbol{z}) \cap (-\partial (f + g)(\boldsymbol{z})) \tag{45}$$
$$\boldsymbol{z_u} \in \partial h^*(\boldsymbol{u}) \cap (\partial (f + g)^*(-\boldsymbol{u})) , \tag{46}$$

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

$$\boldsymbol{u} \in \partial h(\boldsymbol{z_u}) \cap (-\partial (f + g)(\boldsymbol{z_u})) . \tag{47}$$

Using Eq. (45) and (47) we have by monotony of ∂h and $\partial (f + g)$

$$\langle \boldsymbol{u_z} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z_u}\rangle \geq 0 \text{ and } \langle \boldsymbol{u_z} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z_u}\rangle \leq 0 \tag{48}$$

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

$$\left.\begin{aligned} \boldsymbol{u_z} &\in \partial h(\boldsymbol{z}) \\ \boldsymbol{u} &\in \partial h(\boldsymbol{z_u}) \\ \langle \boldsymbol{u_z} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z_u}\rangle &= 0 \end{aligned}\right\} \implies \boldsymbol{u} \in \partial h(\boldsymbol{z}) \tag{49}$$

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

$$\left.\begin{aligned} -\boldsymbol{u_z} &\in \partial (f + g)(\boldsymbol{z}) \\ -\boldsymbol{u} &\in \partial (f + g)(\boldsymbol{z_u}) \\ \langle \boldsymbol{u_z} - \boldsymbol{u}, \boldsymbol{z} - \boldsymbol{z_u}\rangle &= 0 \end{aligned}\right\} \implies -\boldsymbol{u} \in \partial (f + g)(\boldsymbol{z}) \tag{50}$$

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

$$\begin{cases} -\boldsymbol{u} \in \partial (f + g)(\boldsymbol{z}) \\ \boldsymbol{u} \in h(\boldsymbol{z}) . \end{cases} \tag{51}$$

These are the same subdifferential inclusions of (40), which are equivalent to $(\boldsymbol{z}, \boldsymbol{u}) \in \mathrm{Fix}(\boldsymbol{T}_\gamma)$ (Eq. (36)). This concludes the proof. $\qquad \square$

Appendix C.2 Key recursive inequality

In this subsection we present a lemma that will be key for further proofs, as it relates the saddle point suboptimality to the current and previous iterates.

Lemma 4 (Key recursive inequality). *Let f be μ_f-strongly convex and h^* be μ_h-strongly convex (where we allow $\mu_f = 0$ and/or $\mu_h = 0$). Then after t iterations of Algorithm 1 we have the following inequality for all $(\boldsymbol{x}, \boldsymbol{u})$ in the domain of $\mathcal{L}$, with $\gamma_{-1} = \gamma_0$, $\delta_{-1} = 0$.*

For Variant 1:

$$2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) + \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 + (1-\sigma)^{-1}\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 \leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 \,. \tag{52}$$

For Variant 2:

$$\begin{aligned}
2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) + \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 + (1-\xi)^{-1}\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 & \\
\leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 + \gamma_{t-1}\delta_{t-1} - 2\gamma_t\delta_t \,. &
\end{aligned} \tag{53}$$

Proof. The proof is structured in two parts. In the first part, we will bound the "primal" suboptimality $\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})$ and in the second part we will bound the "dual" suboptimality $\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1})$. Finally, adding both will yield the desired result.

Part 1. By μ_f-strong convexity f verifies the following inequality for an arbitrary $\boldsymbol{x}$

$$f(\boldsymbol{z}_t) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{z}_t), \boldsymbol{z}_t - \boldsymbol{x} \rangle - \frac{\mu_f}{2}\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 \,. \tag{54}$$

By the line search condition and the definition of δ_t (we recall $\delta_t = Q_t(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_{t+1})$) we have at each iteration $f(\boldsymbol{x}_{t+1}) = Q_t(\boldsymbol{x}_{t+1}) - \delta_t$. This gives by definition of Q_t:

$$f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{z}_t) + \langle \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_{t+1} - \boldsymbol{z}_t \rangle + \frac{1}{2\gamma_t}\|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\|^2 - \delta_t \tag{55}$$

Adding the previous two equations gives the following inequality, which we will make use of later on

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle + \frac{1}{2\gamma_t}\|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\|^2 - \frac{\mu_f}{2}\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \delta_t \,. \tag{56}$$

From the subdifferential characterization of the proximal operator (Lemma 1), the update $\boldsymbol{x}_{t+1} = \mathbf{prox}_{\gamma_t g}(\boldsymbol{z}_t - \gamma_t(\boldsymbol{u}_t + \nabla f(\boldsymbol{z}_t)))$ of Line 3 implies the following subdifferential inclusion

$$\frac{1}{\gamma_t}(\boldsymbol{z}_t - \gamma_t(\boldsymbol{u}_t + \nabla f(\boldsymbol{z}_t)) - \boldsymbol{x}_{t+1}) \in \partial g(\boldsymbol{x}_{t+1}) \,. \tag{57}$$

By of g we then have the following inequality

$$g(\boldsymbol{x}_{t+1}) - g(\boldsymbol{x}) \leq \frac{1}{\gamma_t}\langle \boldsymbol{z}_t - \gamma_t(\boldsymbol{u}_t + \nabla f(\boldsymbol{z}_t)) - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle \tag{58}$$

$$= \frac{1}{\gamma_t}\langle \boldsymbol{z}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle - \langle \boldsymbol{u}_t + \nabla f(\boldsymbol{z}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle \,. \tag{59}$$

Adding together (56) and (59) we obtain

$$\begin{aligned}
f(\boldsymbol{x}_{t+1}) + g(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) - g(\boldsymbol{x}) \leq & \frac{1}{\gamma_t}\langle \boldsymbol{z}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle + \frac{1}{2\gamma_t}\|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\|^2 \\
& - \langle \boldsymbol{u}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle - \frac{\mu_f}{2}\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \delta_t \,.
\end{aligned} \tag{60}$$

We will now use this last inequality to bound $\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}_t) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t)$. For this, we will make extensive of the cosine identity $2\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \|\boldsymbol{a} + \boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2$ and the primal-dual relationship $\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}_t) = \boldsymbol{x}_{t+1} - \boldsymbol{z}_t$, which is an immediate consequence of the definition of $\boldsymbol{u}_{t+1}$ in Line 9.

$$\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1}) = f(\boldsymbol{x}_{t+1}) + g(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) - g(\boldsymbol{x}) + \langle \boldsymbol{x}_{t+1} - \boldsymbol{x}, \boldsymbol{u}_{t+1} \rangle \tag{61}$$

$$\overset{(60)}{\leq} \frac{1}{\gamma_t}\langle \boldsymbol{z}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x} \rangle + \frac{1}{2\gamma_t}\|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\|^2 + \langle \boldsymbol{x}_{t+1} - \boldsymbol{x}, \boldsymbol{u}_{t+1} - \boldsymbol{u}_t \rangle - \frac{\mu_f}{2}\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \delta_t \tag{62}$$

$$= \left(\frac{1 - \gamma_t\mu_f}{2\gamma_t}\right)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \frac{1}{2\gamma_t}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 - \frac{1}{\gamma_t}\langle \boldsymbol{x}_{t+1} - \boldsymbol{x}, \boldsymbol{z}_{t+1} - \boldsymbol{x}_{t+1} \rangle - \delta_t \tag{63}$$

(cosine identity on first term and primal-dual relationship on second-last term)

$$= \left(\frac{1 - \gamma_t\mu_f}{2\gamma_t}\right)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \frac{1}{2\gamma_t}\|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 + \frac{1}{2\gamma_t}\|\boldsymbol{z}_{t+1} - \boldsymbol{x}_{t+1}\|^2 - \delta_t \tag{64}$$

(cosine identity on the second-last term)

Part 2. From the subdifferential characterization of the proximal operator (Lemma 1), the update $\boldsymbol{z}_{t+1} = \mathbf{prox}_{\gamma_t h}(\boldsymbol{x}_{t+1} + \gamma_t \boldsymbol{u}_t)$ of Line 8 we have the following subdifferential inclusion

$$\boldsymbol{u}_{t+1} = \frac{1}{\gamma_t}(\boldsymbol{x}_{t+1} - \boldsymbol{z}_{t+1}) + \boldsymbol{u}_t \in \partial h(\boldsymbol{z}_{t+1}) \implies \boldsymbol{z}_{t+1} \in \partial h^*(\boldsymbol{u}_{t+1}), \tag{65}$$

where the implication is a consequence of the conjugate-inverse identity (Lemma 2). As we did before, this inclusion can be used to obtain an inequality in terms of the function values. By μ_h-strong convexity of h^* we have

$$h^*(\boldsymbol{u}_{t+1}) - h^*(\boldsymbol{u}) \leq \langle \boldsymbol{z}_{t+1}, \boldsymbol{u}_{t+1} - \boldsymbol{u} \rangle - \frac{\mu_h}{2}\|\boldsymbol{u}_{t+1} - \boldsymbol{u}\|^2. \tag{66}$$

We can now use this inequality to to bound $\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}_{t+1})$ as follows:

$$\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1}) = h^*(\boldsymbol{u}_{t+1}) - h^*(\boldsymbol{u}) + \langle \boldsymbol{x}_{t+1}, \boldsymbol{u} - \boldsymbol{u}_{t+1} \rangle \tag{67}$$

$$\overset{(66)}{\leq} \langle \boldsymbol{z}_{t+1} - \boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1} - \boldsymbol{u} \rangle - \frac{\mu_h}{2}\|\boldsymbol{u}_{t+1} - \boldsymbol{u}\|^2 \tag{68}$$

$$= \frac{1}{\gamma_t}\langle \boldsymbol{z}_{t+1} - \boldsymbol{x}_{t+1}, \gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}) \rangle - \frac{\mu_h}{2}\|\boldsymbol{u}_{t+1} - \boldsymbol{u}\|^2 \tag{69}$$

$$= \frac{1}{\gamma_t}\langle \boldsymbol{z}_{t+1} - \boldsymbol{x}_{t+1}, \gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}) \rangle - \frac{\gamma_t^{-1}\mu_h}{2\gamma_t}\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 \tag{70}$$

$$= \frac{1}{2\gamma_t}\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - \frac{1}{2\gamma_t}\|\boldsymbol{z}_{t+1} - \boldsymbol{x}_{t+1}\|^2 - \left(\frac{1 + \gamma_t^{-1}\mu_h}{2\gamma_t}\right)\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2, \tag{71}$$

where in the second line we have used again the primal-dual relationship $\boldsymbol{x}_{t+1} - \boldsymbol{z}_{t+1} = \gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}_t)$.

Third part: putting it all together. Adding the inequalities from Eq. (64) and Eq. (71) and multiplying everything by by $2\gamma_t$ we obtain

$$2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) \leq (1 - \gamma_t\mu_f)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2$$
$$+ \|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - (1 + \gamma_t^{-1}\mu_h)\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 - 2\gamma_t\delta_t. \tag{72}$$

We will now prove that γ_t is lower bounded by $\min\{\tau L_f^{-1}, \gamma_0\}$ through a distinction of cases. If $\gamma_0 \geq \tau/L_f$, then by the properties of L_f-smooth functions, the sufficient decrease condition is verified for all step-size smaller than $1/L_f$. The resulting step-size can still be smaller than this quantity if the sufficient decrease condition fails for some step-size $\gamma > 1/L_f$ and τ is small enough such that $\tau\gamma \leq 1/L_f$. Even in this worst case scenario we have $\gamma_t \geq \tau/L_f$, which proves the bound for the case $\gamma_0 \geq \tau/L_f$. If $\gamma_0 \leq \tau/L_f$, then the sufficient decrease condition is verified at this first iterate, and the step-size can only increase.

We can hence bound $\gamma_t \mu_f$ by ρ, defined in Eq. (16), to obtain

$$
\begin{aligned}
2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) \leq {}& (1 - \rho)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 \\
& + \|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - (1 + \gamma_t^{-1}\mu_h)\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 - 2\gamma_t\delta_t .
\end{aligned}
\tag{73}
$$

For Variant 1, we can just drop the non-positive term $-2\gamma_t\delta_t$ and use the non-increasing step to bound the following terms as:

$$
\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 \leq \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 \quad \text{(with } \gamma_{-1} = \gamma_0 \text{ by definition)} \tag{74}
$$

$$
-(1 + \gamma_t^{-1}\mu_h) \leq -(1 + \gamma_0^{-1}\mu_h) \tag{75}
$$

$$
= -\frac{\gamma_0 + \mu_h}{\gamma_0} = -\left(\frac{\gamma_0 + \mu_h - \mu_h}{\gamma_0 + \mu_h}\right)^{-1} \tag{76}
$$

$$
= -\left(1 - \frac{\mu_h}{\gamma_0 + \mu_h}\right)^{-1} = -(1 - \sigma)^{-1} \tag{77}
$$

Replacing in (73) gives

$$
\begin{aligned}
2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) \leq {}& (1 - \rho)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 \\
& + \|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - (1 - \sigma)^{-1}\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 ,
\end{aligned}
\tag{78}
$$

from where we obtain the desired bound (52) by reordering the terms.

We will now derive a similar bound for Variant 2. In this case, 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 $\|\boldsymbol{u}_t - \boldsymbol{u}\|^2 \leq 2\|\boldsymbol{u}_t\|^2 + 2\|\boldsymbol{u}\|^2 \leq 4\beta_h$.

Assuming first $t > 0$, we have the following sequence of inequalities for any $\gamma_t \geq \gamma_{t-1}$:

$$
\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - \gamma_{t-1}\delta_{t-1} = (\gamma_t^2 - \gamma_{t-1}^2)\|\boldsymbol{u}_t - \boldsymbol{u}\|^2 - \gamma_{t-1}\delta_{t-1} \tag{79}
$$

$$
\leq (\gamma_t^2 - \gamma_{t-1}^2)4\beta_h^2 - \gamma_{t-1}\delta_{t-1} \tag{80}
$$

$$
\text{(using } \gamma_t \geq \gamma_{t-1} \text{ and that } h \text{ is } \beta_h\text{-Lipschitz)}
$$

$$
\leq (\gamma_{t-1}^2 + \gamma_{t-1}\delta_{t-1}\beta_h^{-2}/4 - \gamma_{t-1}^2)4\beta_h^2 - \gamma_{t-1}\delta_{t-1} \tag{81}
$$

$$
\text{(by the choice of } \gamma_{t+1} \text{ in Line 15)}
$$

$$
= \gamma_{t-1}^2\beta_h^2 + \gamma_{t-1}\delta_{t-1} - \gamma_{t-1}^2\beta_h^2 - \gamma_{t-1}\delta_{t-1} \tag{82}
$$

$$
= 0 , \tag{83}
$$

which reordering gives

$$
\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u})\|^2 \leq \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 + \gamma_{t-1}\delta_{t-1} . \tag{84}
$$

This inequality is also trivially true when $\gamma_t \leq \gamma_{t-1}$ because of the non-negativity of δ_t. If $t = 0$, then we have by definition $\gamma_0 = \gamma_{-1}, \delta_{-1} = 0$ and so the above inequality is also trivially verified.

Finally, plugging this lasts bound in (73) gives

$$
\begin{aligned}
2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) \leq {}& (1 - \rho)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 \\
& + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 - (1 + \gamma_t^{-1}\mu_h)\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 - 2\gamma_t\delta_t + \gamma_{t-1}\delta_{t-1} .
\end{aligned}
$$

Now we will use the following bound on γ_t^{-1}: $\mu_f \leq \gamma_t^{-1}$. This bound can be deduced from the the strong convexity inequality (Definition 4), as otherwise the sufficient decrease condition $\delta_t \geq 0$ would not hold. In all, we have the desired bound

$$
\begin{aligned}
2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{t+1})) \leq {}& (1 - \rho)\|\boldsymbol{z}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u})\|^2 \\
& - (1 + \mu_f\mu_h)\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 + \gamma_{t-1}\delta_{t-1} - 2\gamma_t\delta_t .
\end{aligned}
$$

Finally, by trivial algebraic manipulations we have

$$(1 - \xi)^{-1} = (1 - \frac{\mu_f \mu_h}{1 + \mu_f \mu_h})^{-1} = (\frac{1}{1 + \mu_f \mu_h})^{-1} = 1 + \mu_f \mu_h \tag{85}$$

which is the term that multiplies $\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2$ in the previous equation. Replacing with $(1 - \xi)^{-1}$ in that equation we have the desired bound.

$\square$

Appendix C.3 Sublinear convergence

Theorem 2 (sublinear convergence rate). *For every iteration $t \geq 0$ and any $(\boldsymbol{x}, \boldsymbol{u})$ in the domain of $\mathcal{L}$ we have the following convergence rate for Algorithm 1 (both variants):*

$$\mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \overline{\boldsymbol{u}}_{t+1}) \leq \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \gamma_0^2 \|\boldsymbol{u}_0 - \boldsymbol{u}\|^2}{2 s_t} . \tag{86}$$

Proof. Adding the Equation of Lemma 4 with $\mu_f = \mu_h = 0$ (which implies $\rho = \sigma = \xi = 0$) from 0 to t and dropping positive terms in the left hand side we get for both Variant 1:

$$\sum_{i=0}^{t} \gamma_i \left(\mathcal{L}(\boldsymbol{x}_{i+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{i+1}) \right) \leq \frac{1}{2} \|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \frac{1}{2} \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2$$

$$- \|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2 - \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 \tag{87}$$

$$\leq \frac{1}{2} \|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \frac{1}{2} \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2 . \tag{88}$$

Similarly, for Variant 2 we have

$$\sum_{i=0}^{t} \gamma_i \left(\mathcal{L}(\boldsymbol{x}_{i+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{i+1}) \right) \leq \frac{1}{2} \|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \frac{1}{2} \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2 - \|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u})\|^2$$

$$- \|\boldsymbol{z}_{t+1} - \boldsymbol{x}\|^2 + \gamma_{-1}\delta_{-1} - \sum_{i=0}^{t} \gamma_t \delta_t \tag{89}$$

$$\leq \frac{1}{2} \|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \frac{1}{2} \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2 + \gamma_{-1}\delta_{-1} - \sum_{i=0}^{t} \gamma_t \delta_t \tag{90}$$

$$\leq \frac{1}{2} \|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \frac{1}{2} \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2 , \tag{91}$$

where the last line follows from the definition $\delta_{-1} = 0$ and the non-negativity of δ_t.

In all, we have for both variants

$$\sum_{i=0}^{t} \gamma_i \left(\mathcal{L}(\boldsymbol{x}_{i+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{i+1}) \right) \leq \frac{1}{2} \|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \frac{1}{2} \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2 . \tag{92}$$

For a fixed $(\boldsymbol{x}, \boldsymbol{u})$, the saddle point suboptimality $\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_t)$ is convex in $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$. By Jensens inequality we can then bound the left hand side of Eq. (92) as

$$\frac{1}{s_t} \sum_{i=0}^{t} \gamma_i (\mathcal{L}(\boldsymbol{x}_{i+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}_{i+1})) \geq \mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \overline{\boldsymbol{u}}_{t+1}) . \tag{93}$$

Combining this last inequality with Eq. (92) we obtain

$$\mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \boldsymbol{u}) - \mathcal{L}(\boldsymbol{x}, \overline{\boldsymbol{u}}_{t+1}) \leq \frac{1}{2s_t} \left(\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u})\|^2 \right) , \tag{94}$$

which is the desired result. $\qquad\square$

Corollary 1. *Let h be β_h-Lipschitz. Then, we have the following rate for the weighted ergodic iterate*

$$P(\overline{\boldsymbol{x}}_{t+1}) - P(\boldsymbol{x}^\star) \leq \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + 2\gamma_0^2(\|\boldsymbol{u}_0\|^2 + \beta_h^2)}{2s_t} \tag{95}$$

Proof. 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.

Using this and the previous theorem we can write the following set of inequalities

$$\begin{align}
P(\overline{\boldsymbol{x}}_{t+1}) - P(\boldsymbol{x}^\star) &= \mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \widehat{\boldsymbol{u}}) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{u}^\star) \tag{96} \\
&\leq \mathcal{L}(\overline{\boldsymbol{x}}_{t+1}, \widehat{\boldsymbol{u}}) - \mathcal{L}(\boldsymbol{x}^\star, \widehat{\boldsymbol{u}}) \tag{97} \\
&\qquad \text{(definition of saddle point, Eq. (9) with } \boldsymbol{x} = \boldsymbol{x}^\star) \\
&\leq \frac{1}{2s_t} \left(\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \|\gamma_0(\boldsymbol{u}_0 - \widehat{\boldsymbol{u}})\|^2 \right) \tag{98} \\
&\qquad \text{(Theorem 2 with } \boldsymbol{x} = \boldsymbol{x}^\star, \boldsymbol{u} = \widehat{\boldsymbol{u}})
\end{align}$$

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 $\|\gamma_0(\boldsymbol{u}_0 - \widehat{\boldsymbol{u}})\|^2 \leq 2\gamma_0^2 \|\boldsymbol{u}_0\|^2 + 2\gamma_0^2 \|\widehat{\boldsymbol{u}}\|^2 \leq 2\gamma_0^2 (\|\boldsymbol{u}_0\|^2 + \beta_h^2)$. Plugging this bound into the last inequality we have the desired bound

$$P(\overline{\boldsymbol{x}}_{t+1}) - P(\boldsymbol{x}^\star) \leq \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \gamma_0^2(\|\boldsymbol{u}_0\|^2 + \beta_h^2)}{2s_t} . \tag{99}$$

$\qquad\square$

Appendix C.4   Linear convergence

In this subsection we assume that f is μ_f-strongly convex and h is L_h-smooth (with $\mu_f > 0, 0 < L_h < +\infty$). We denote by $\boldsymbol{x}^\star$ the minimizer of the primal loss (unique by strong convexity of P) and by $\boldsymbol{u}^\star$ the minimizer of the dual loss (also unique by strong convexity of D, which is a consequence of the duality between L-smoothness and strong convexity).

Theorem 3. *Let $\boldsymbol{x}_{t+1}, \boldsymbol{u}_{t+1}$ be the iterates produced by Algorithm 1 after t iterations. Then we have the following linear convergence for Variant 1 (V1) and Variant 2 (V2):*

$$V1 : \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \min\{\rho, \sigma\}\right)^{t+1} D_0 \tag{100}$$

$$V2 : \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1} E_0 , \tag{101}$$

with $D_0 \stackrel{\text{def}}{=} 6\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{6}{1-\sigma}\|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u}^\star)\|^2$ and $E_0 \stackrel{\text{def}}{=} 6\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{6}{1-\xi}\|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u}^\star)\|^2$.

Proof. **Variant 1:** By the duality between Lipschitz gradient and strong convexity, h being L_h-smooth implies that h^* is L_h^{-1}-strongly convex. Applying Lemma 4 with $\boldsymbol{x} = \boldsymbol{x}^\star$, $\boldsymbol{u} = \boldsymbol{u}^\star$ we obtain the following inequality for Variant 1:

$$2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{u}_{t+1})) + \|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 + (1-\sigma)^{-1}\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}^\star)\|^2 \tag{102}$$

$$\leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 , \tag{103}$$

where we note that the assumption $0 < L_h$ implies $\sigma < 1$ and so $(1-\sigma)^{-1}$ is well defined. For convenience we introduce the notation $\alpha_t \overset{\text{def}}{=} \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + (1-\sigma)^{-1}\|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2$. With this, the previous inequality can be simplified to

$$2\gamma_t(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{u}_{t+1})) + \alpha_{t+1} \leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 . \tag{104}$$

By the definition of saddle point we have $\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{u}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{u}_{t+1}) \geq 0$. Dropping this non-negative term gives

$$\alpha_{t+1} \leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 . \tag{105}$$

We now make a distinction of cases based on the relative magnitude of ρ and σ.

- If $\rho \leq \sigma$ then $(1-\rho) \geq (1-\sigma) \implies 1 \leq (1-\rho)(1-\sigma)^{-1}$ and so from the previous equation we have

$$\alpha_{t+1} \leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 \tag{106}$$

$$\leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + (1-\rho)(1-\sigma)^{-1}\|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 \tag{107}$$

$$= (1-\rho)\alpha_t \tag{108}$$

- Otherwise, if $\sigma < \rho$, then $(1-\rho) < (1-\sigma)$ and we have

$$\alpha_{t+1} \leq (1-\rho)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 \tag{109}$$

$$\leq (1-\sigma)\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^2 + \|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 \tag{110}$$

$$= (1-\sigma)\alpha_t \tag{111}$$

Combining the two previous equations we have

$$\alpha_{t+1} \leq \left(1 - \min\{\rho, \sigma\}\right)\alpha_t , \tag{112}$$

which leads by recurrence to

$$\alpha_{t+1} \leq \left(1 - \min\{\rho, \sigma\}\right)^{t+1}\alpha_0 . \tag{113}$$

By definition of α_{t+1} the previous inequality gives

$$\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 + (1-\sigma)^{-1}\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}^\star)\|^2 \leq \left(1 - \min\{\rho, \sigma\}\right)^{t+1}\alpha_0. \tag{114}$$

From where we have the following geometric bounds for the primal and dual variables

$$\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \min\{\rho, \sigma\}\right)^{t+1}\alpha_0 \tag{115}$$

$$\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}^\star)\|^2 \leq (1-\sigma)\left(1 - \min\{\rho, \sigma\}\right)^{t+1}\alpha_0 \tag{116}$$

$$\leq \left(1 - \min\{\rho, \sigma\}\right)^{t+2}\alpha_0 \tag{117}$$

This would be sufficient to derive a convergence rate in terms of $\|z_{t+1} - x^\star\|^2$. However, since our sublinear convergence result was in terms of x_{t+1}, we would like to state this result in terms of x_{t+1} too. This is possible with a small amount of work and loosing a constant factor 6.

Using the primal-dual relationship $x_{t+1} = z_{t+1} + \gamma_t(u_{t+1} - u_t)$ and an improved version of Young's inequality (Lemma 3) $\|a + b + c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ we have

$$\|x_{t+1} - x^\star\|^2 = \|z_{t+1} + \gamma_t(u_{t+1} - u_*) + \gamma_t(u_* - u_t) - x^\star\|^2 \quad \text{(adding and substracting } u^\star) \tag{118}$$

$$\leq 3\|z_{t+1} - x^\star\|^2 + 3\|\gamma_t(u_{t+1} - u_*)\|^2 + 3\|\gamma_t(u_* - u_t)\|^2 \quad \text{(Young's inequality)} \tag{119}$$

$$\leq 3\|z_{t+1} - x^\star\|^2 + 3\|\gamma_t(u_{t+1} - u^\star)\|^2 + 3\|\gamma_{t-1}(u_t - u^\star)\|^2 \tag{120}$$

$$(\gamma_t \leq \gamma_{t-1})$$

$$\leq 3\|z_{t+1} - x^\star\|^2 + 3(1 - \sigma)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + 3\|\gamma_{t-1}(u_t - u^\star)\|^2 \tag{121}$$

$$(1 \leq (1 - \sigma)^{-1})$$

$$\leq 3\Big(1 - \min\{\rho, \sigma\}\Big)^{t+1}\alpha_0 + 3\|\gamma_{t-1}(u_t - u^\star)\|^2 \tag{122}$$

$$\text{(by Eq.(113))}$$

$$\leq 3\Big(1 - \min\{\rho, \sigma\}\Big)^{t+1}\alpha_0 + 3\Big(1 - \min\{\rho, \sigma\}\Big)^{t+1}\alpha_0 \quad \text{(by Eq. (117))} \tag{123}$$

$$\leq \Big(1 - \min\{\rho, \sigma\}\Big)^{t+1} 6\alpha_0 . \tag{124}$$

The claimed rate then follows by definition of α_0.

Variant 2: As in Variant 1, by the duality between Lipschitz gradient and strong convexity, h being L_h-smooth implies that h^* is L_h^{-1}-strongly convex. Using Lemma 4 with $x = x^\star$, $u = u^\star$ we obtain the following inequality for Variant 2:

$$2\gamma_t(\mathcal{L}(x_{t+1}, u^\star) - \mathcal{L}(x^\star, u_{t+1})) + \|z_{t+1} - x^\star\|^2 + (1 - \xi)^{-1}\|\gamma_{t-1}(u_t - u^\star)\|^2 + 2\gamma_t\delta_t \tag{125}$$

$$\leq (1 - \rho)\|z_t - x^\star\|^2 + \|\gamma_{t-1}(u_t - u^\star)\|^2 + \gamma_{t-1}\delta_{t-1} ., \tag{126}$$

where we note that the assumption $0 < L_h$ implies $\xi < 1$ and so $(1 - \xi)^{-1}$ well defined. Once again by the definition of saddle point we can drop the first non-negative term to get

$$\widetilde{\alpha}_{t+1} \leq (1 - \rho)\|z_t - x^\star\|^2 + \|\gamma_{t-1}(u_t - u^\star)\|^2 + \gamma_{t-1}\delta_{t-1} , \tag{127}$$

where we noted $\widetilde{\alpha}_t \overset{\text{def}}{=} \|z_t - x^\star\|^2 + (1 - \xi)^{-1}\|\gamma_t(u_{t-1} - u^\star)\|^2 + 2\gamma_{t-1}\delta_{t-1}$. Similarly as for the first variant, we can make a distinction of cases to prove the contraction:

- If $\rho < \xi$, then $(1 - \rho) \implies (1 - \rho)(1 - \xi) \geq 1$ and we have

$$\widetilde{\alpha}_{t+1} \leq (1 - \rho)\|z_t - x^\star\|^2 + (1 - \rho)(1 - \xi)^{-1}\|\gamma_{t-1}(u_t - u^\star)\|^2 + \frac{1}{2}(2\gamma_{t-1}\delta_{t-1}) \tag{128}$$

$$= \Big(1 - \min\{\rho, \tfrac{1}{2}\}\Big)\widetilde{\alpha}_t \tag{129}$$

- If $\xi < \rho$, then $(1 - \rho) < (1 - \xi)$ and we have

$$\widetilde{\alpha}_{t+1} \leq (1 - \xi)\|z_t - x^\star\|^2 + \|\gamma_{t-1}(u_t - u^\star)\|^2 + \frac{1}{2}(2\gamma_{t-1}\delta_{t-1}) \tag{130}$$

$$= \Big(1 - \min\{\xi, \tfrac{1}{2}\}\Big)\widetilde{\alpha}_t \tag{131}$$

Combining the previous two bounds, we have the following geometric decrease on the sequence $(\widetilde{\alpha}_t)$,

$$\widetilde{\alpha}_{t+1} \leq \left(1 - \min\left\{\rho, \tfrac{1}{2}, \xi\right\}\right)\widetilde{\alpha}_t . \tag{132}$$

Then by definition of $\widetilde{\alpha}_{t+1}$ we have

$$\|z_{t+1} - x^\star\|^2 + (1-\xi)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + 2\gamma_t\delta_t \leq \left(1 - \min\left\{\rho, \sigma, \tfrac{1}{2}\right\}\right)^{t+1}\widetilde{\alpha}_0. \tag{133}$$

From where we can derive the following geometric bounds for the primal and dual variables:

$$\|z_{t+1} - x^\star\|^2 \leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1}\widetilde{\alpha}_0 \tag{134}$$

$$(1-\xi)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + 2\gamma_t\delta_t \leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1}\widetilde{\alpha}_0 \tag{135}$$

We will now make a distinction of cases on ξ to derive a more convenient bound for this last inequality. If $\xi \leq \tfrac{1}{2} \implies (1-\xi)^{-1} \leq \tfrac{1}{2}$ and so we have

$$(1-\xi)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + (1-\xi)^{-1}\gamma_t\delta_t \leq (1-\xi)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + 4(1-\xi)^{-1}\gamma_t\delta_t \tag{136}$$
$$\text{(by non-negativity of } \delta_t)$$
$$\leq (1-\xi)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + 2\gamma_t\delta_t \tag{137}$$
$$\leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1}\widetilde{\alpha}_0 \tag{138}$$

Multiplying both sides by $(1-\xi)$ and using $(1-\xi) \leq (1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\})$ we have

$$\|\gamma_t(u_{t+1} - u^\star)\|^2 + \gamma_t\delta_t \leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+2}\widetilde{\alpha}_0 . \tag{139}$$

If on the other hand, $\xi \geq \tfrac{1}{2} \implies (1 - \tfrac{1}{2})(1-\xi)^{-1} \leq 1$ and so we have

$$\|\gamma_t(u_{t+1} - u^\star)\|^2 + \gamma_t\delta_t \leq (1 - \tfrac{1}{2})\left((1-\xi)^{-1}\|\gamma_t(u_{t+1} - u^\star)\|^2 + 2\gamma_t\delta_t\right) \tag{140}$$
$$\leq (1 - \tfrac{1}{2})\left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1}\widetilde{\alpha}_0 \tag{141}$$
$$\leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+2}\widetilde{\alpha}_0 \tag{142}$$

By (138) and (142), we see that in both cases we have the following bound:

$$\|\gamma_t(u_{t+1} - u^\star)\|^2 + \gamma_t\delta_t \leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+2}\widetilde{\alpha}_0 . \tag{143}$$

To obtain a convergence rate in terms of $\|x_{t+1} - x^\star\|^2$, we will use the primal-dual relationship $x_{t+1} = z_{t+1} + \gamma_t(u_{t+1} - u_t)$ and an improved version of Young's inequality (Lemma 3) $\|a + b + c\|^2 \leq$

$3(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + \|\boldsymbol{c}\|^2)$. We have the following sequence of inequalities:

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 = \|\boldsymbol{z}_{t+1} + \gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}_*) + \gamma_t(\boldsymbol{u}_* - \boldsymbol{u}_t)) - \boldsymbol{x}^\star\|^2 \tag{144}$$
$$\text{(adding and substracting } \boldsymbol{u}^\star)$$

$$\leq 3\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 + 3\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}_*)\|^2 + 3\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u}_*)\|^2 \tag{145}$$
$$\text{(Young's inequality)}$$

$$\leq 3\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 + 3\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}^\star)\|^2 + 3\|\gamma_{t-1}(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 + 3(\gamma_{t-1}\delta_{t-1}(2\beta_h)^{-2})\|\boldsymbol{u}_t - \boldsymbol{u}^\star\|^2 \tag{146}$$
$$\text{(by the maximum step-size increase } \gamma_t^2 \leq \gamma_{t-1}^2 + \gamma_{t-1}\delta_{t-1}(2\beta_h)^{-2})$$

$$\leq 3\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^2 + 3\|\gamma_t(\boldsymbol{u}_{t+1} - \boldsymbol{u}^\star)\|^2 + 3\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 + 3\gamma_{t-1}\delta_{t-1} \tag{147}$$
$$\text{(by the Lipschitz assumption on } h \text{ have } \|\boldsymbol{u}_t - \boldsymbol{u}^\star\|^2 \leq 4\beta_h^2)$$

$$\leq 3\left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1} \widetilde{\alpha}_0 + 3\|\gamma_t(\boldsymbol{u}_t - \boldsymbol{u}^\star)\|^2 + 3\gamma_{t-1}\delta_{t-1} \tag{149}$$
$$\text{(by Eq.(133) and using the bound } (1 - \xi)^{-1} > 1)$$

$$\leq 3\left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1} \widetilde{\alpha}_0 + 3\left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1} \widetilde{\alpha}_0 \tag{150}$$
$$\text{(by Eq. (143))}$$

$$\leq \left(1 - \min\left\{\rho, \xi, \tfrac{1}{2}\right\}\right)^{t+1} 6\widetilde{\alpha}_0. \tag{151}$$

The desired rate is then a consequence of the definition of $\widetilde{\alpha}_0$, using that by definition $\delta_{-1} = 0$. $\qquad\square$

Appendix C.5   Convergence rates comparison

We compare the obtained convergence rate against those in Davis & Yin (2015) for the fixed step-size strategy.

Sublinear convergence.   Our own Corollary 1 and (Davis & Yin, 2015, Corollary D.5.2) provide a convergence rate of the form

$$P(\overline{\boldsymbol{x}}_{t+1}) - P(\boldsymbol{x}^\star) \leq \frac{L_f}{2(t+1)} Q_0 , \tag{152}$$

with a different definition of Q_0 in both cases:

$$\text{This work: } Q_0 \overset{\text{def}}{=} \|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{2}{L_f^2}\left(\|\boldsymbol{u}_0\|^2 + \beta_h^2\right) \tag{153}$$

$$\text{(Davis \& Yin, 2015): } Q_0 \overset{\text{def}}{=} \|\boldsymbol{y}_0 - \boldsymbol{x}^\star\|^2 + \frac{1}{2}\|\boldsymbol{y}_0 - \boldsymbol{y}^*\|^2 + \frac{4}{L_f}\|\boldsymbol{y}_0 - \boldsymbol{y}^*\|\|\nabla f(\boldsymbol{x}^\star)\| + \frac{4\beta_h}{L_f}\|\boldsymbol{y}_0 - \boldsymbol{y}^*\| , \tag{154}$$

where $\boldsymbol{y}_0 \overset{\text{def}}{=} \boldsymbol{z}_0 + \frac{1}{L_f}\boldsymbol{u}_0$, $\boldsymbol{y}^* \overset{\text{def}}{=} \boldsymbol{x}^\star + \frac{1}{L}\boldsymbol{u}^\star$ the value of Q_0 for (Davis & Yin, 2015) was obtained by plugging the step-size $\gamma = 1/L_f$ in their Corollary D.5.2, and optimizing with respect to their ε parameter.

Both quantities are difficult to compare, but it is instructive to compare them on specific cases. If $h = 0$ and $\boldsymbol{u}_0 = \boldsymbol{u}^\star = 0$ we have

$$\text{This work: } Q_0 = \|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 \tag{155}$$

$$\text{(Davis \& Yin, 2015): } Q_0 = \frac{3}{2}\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|^2 + \frac{4}{L_f}\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\|\|\nabla f(\boldsymbol{x}^\star)\| + \frac{4\beta_h}{L_f}\|\boldsymbol{z}_0 - \boldsymbol{x}^\star\| , \tag{156}$$

And in this case it is clear that our bound is better. Consider now the case $f = g = 0$ and $\boldsymbol{z}_0 = \boldsymbol{x}^\star = 0$. Then we have the following value of Q_0:

$$\text{This work: } Q_0 \stackrel{\text{def}}{=} \frac{2}{L_f^2} \left(\|\boldsymbol{u}_0\|^2 + \beta_h^2 \right) \tag{157}$$

$$\text{(Davis \& Yin, 2015): } Q_0 \stackrel{\text{def}}{=} \frac{1}{L_f^2} \|\boldsymbol{u}_0\|^2 + \frac{1}{2L_f^2} \|\boldsymbol{u}_0 - \boldsymbol{u}^\star\|^2 + \frac{4\beta_h}{L_f^{3/2}} \|\boldsymbol{u}_0 - \boldsymbol{u}^\star\| , \tag{158}$$

in this case however the bound of (Davis & Yin, 2015) can be better as it avoids the quadratic depency on β_h, although the quantity $\|\boldsymbol{u}_0 - \boldsymbol{u}^\star\|$ is also in the worse case in the order of $2\beta_h$.

Linear convergence. For a step-size $\gamma = 1/L_f$, the convergence rate of (Davis & Yin, 2017, Theorem D.6, point 6) gives a convergence of the form (using our notation and the step-size $\gamma = 1/L_f$):

$$\frac{2\mu_f(1 - \eta)}{L_f(1 + \gamma L_h)^2} \tag{159}$$

where η is a quantity that verifies the inequalities

$$\eta > \frac{1}{\varepsilon}, \text{ with } \varepsilon > \frac{1}{2} \tag{160}$$

by the choice of step-size (see their Theorem D.6) for a definition of these quantities. Optimizing with respect to η gives $\eta < 1/2$ and so the rate can be arbitrarily close to

$$\frac{\mu_f}{L_f(1 + \gamma L_h)^2} = \rho\sigma^2 . \tag{161}$$

This bound is clearly *strictly worse* than our $\min\{\rho, \sigma\}$ of Theorem 3 for Variant 1 with $\tau = 1$. The constant factor D_0 is worse than the one obtained by Davis & Yin (2015), but only by a factor $6(1 - \sigma)^{-1}$, and note also that the rate of Davis & Yin (2015) is not given in terms of the primal variables but in terms of the less interpretable quantity $\|\boldsymbol{x}_t + \gamma\boldsymbol{u}_{t+1} - \boldsymbol{x}^\star - \gamma\boldsymbol{u}^\star\|^2$.

The difference between our convergence rate and that of Davis & Yin (2015) can be quite large, as illustrated in the plot below for values in the interval $[0, \frac{1}{2}]^2$:
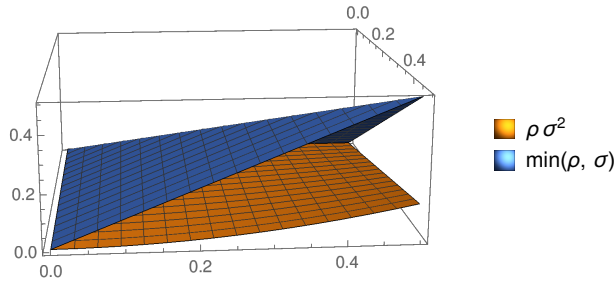
Appendix D Proximal operators

In this section we provide a more detailed description of the different splitting methods. We start with a reminder of some penalties that have a known closed form proximal operator.

- ℓ_1 **norm.** The proximal operator of the ℓ_1 norm $\|x\|_1 \overset{\text{def}}{=} \sum_{j=1}^p |x_j|$ norm is given by the soft-thresholding operator. More precisely, for $x \in \mathbb{R}^p$ it is given componentwise as

$$[\mathbf{prox}_{\gamma\|\cdot\|}(x)]_j = \begin{cases} \left(1 - \frac{\gamma}{|x_j|}\right) x_j, & \text{if } |x_j| \geq \gamma \\ 0 & \text{otherwise} \end{cases} \tag{162}$$

  for $j \in \{1, \ldots, p\}$.

- **Group lasso norm.** Let $\mathcal{G}$ be a partition of $\{1, \ldots, p\}$. Then the group lasso or ℓ_1/ℓ_2 norm for the $\mathcal{G}$ partition is defined as $\|x\|_{\mathcal{G}} \overset{\text{def}}{=} \sum_{g \in \mathcal{G}} \|[x]_g\|_2$. Its proximal operator is given as

$$[\mathbf{prox}_{\gamma\|\cdot\|_{\mathcal{G}}}(x)]_g = \begin{cases} \left(1 - \frac{\gamma}{\|[x]_g\|}\right)_+ [x]_g, & \text{if } \|[x]_g\| \geq \gamma \\ 0 & \text{otherwise} \end{cases} \tag{163}$$

  A proof of this can be found in (Yuan & Lin, 2006, Proposition 1).

Appendix D.1 ℓ_1 trend filtering.

This penalty can be split as the sum of three proximal terms h_1, h_2 and h_3 as follows:

$$\|x\|_{\text{TF}} \overset{\text{def}}{=} \sum_{i=1}^{p-2} |x_i - 2x_{i+1} + x_{i+2}| \tag{164}$$

$$= \underbrace{\sum_{i=1}^{(p-2)/3} |x_i - 2x_{i+1} + x_{i+2}|}_{=h_1(x)} + \underbrace{\sum_{i=1}^{(p-2)/3} |x_{i+1} - 2x_{i+2} + x_{i+3}|}_{=h_2(x)} + \underbrace{\sum_{i=1}^{(p-2)/3} |x_{i+2} - 2x_{i+3} + x_{i+4}|}_{=h_3(x)}$$
$$\tag{165}$$

The term h_1 can be written as $\|Lx\|_1$ for an L matrix of the form

$$L = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 1 & -2 & 1 & \ddots \\ \vdots & \ddots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} \tag{166}$$

Usually, penalties of the form $\|Lx\|_1$ do not have a closed form proximal operator. However, this matrix L has a very special property that will allow to compute its proximal operator in closed form. In particular, this matrix is a *semi-orthogonal* matrix, i.e., it verifies $LL^T = \nu I$ for some $\nu > 0$ ($\nu = 6$ in our case). For these matrices, the proximal operator of $\varphi(Lx)$ can be computed as (see e.g. (Combettes & Pesquet, 2011, Table 1)):

$$\mathbf{prox}_{\varphi(L\cdot)} = x + \nu^{-1}L^T(\mathbf{prox}_{\nu\varphi}(Lx) - Lx), \tag{167}$$

where in this case φ is the ℓ_1 norm and so its proximal operator is the soft thresholding operator of Eq. (162).

Appendix D.2   Isotonic and nearly-isotonic penalties

The isotonic constraint $\{x_1 \leq x_2 \leq \ldots \leq x_p\}$ can be enforced through the use of the indicator function $\imath\{x_1 \leq x_2 \leq x_3 \leq x_4 \leq \cdots \leq x_p\}$. Suppose that $p = 2$, i.e., we only have the constraint $\{x_1 \leq x_2\}$. In this case, the projection of (z_1, z_2) can be computed as follows: if $z_1 \leq z_2$ the projection does obviously nothing, otherwise it can be computed as the projection onto the line generated by the vector $(1, 1)$. This gives:

$$\mathbf{prox}_{\imath\{x_1 \leq x_2\}}(x_1, x_2) = \begin{cases} (x_1, x_2) & \text{if } x_1 \leq x_2 \\ ((x_1 + x_2)/2, (x_1 + x_2)/2) & \text{otherwise.)} \end{cases} \tag{168}$$

Now lets consider the general case. The full indicator function can be decomposed into a sum of the two following terms

$$\imath\{x_1 \leq x_2 \leq x_3 \leq x_4 \leq \cdots\} = \underbrace{\imath\{x_1 \leq x_2; x_3 \leq x_4; \cdots\}}_{=g(x)} + \underbrace{\imath\{x_2 \leq x_3; x_4 \leq x_5; \cdots\}}_{=h(x)} . \tag{169}$$

where the both terms are block separable with blocks of size 2. We can hence use the proximal operator of Eq. (168) to derive the proximal operator of g and h, which by block-separability is merely a concatenation of the previous one. These proximal operators are given block-wise as

$$[\mathbf{prox}_{\gamma g}(z)]_{\{2i,2i+1\}} = \begin{cases} (x_{2i}, x_{2i+1}) & \text{if } x_{2i} \leq x_{2i+1} \\ ((x_{2i} + x_{2i+1})/2, (x_{2i} + x_{2i+1})/2) & \text{otherwise.)} \end{cases} \tag{170}$$

$$[\mathbf{prox}_{\gamma h}(z)]_{\{2i+1,2i+2\}} = \begin{cases} (x_{2i+1}, x_{2i+2}) & \text{if } x_{2i+1} \leq x_{2i+2} \\ ((x_{2i+1} + x_{2i+2})/2, (x_{2i+1} + x_{2i+2})/2) & \text{otherwise.)} \end{cases} \tag{171}$$

The nearly isotonic penalty.   The nearly isotonic penalty was proposed by Tibshirani et al. (2011) as a relaxation of the aforementioned isotonic constraints and is defined as

$$\|x\|_{\text{iso}} \stackrel{\text{def}}{=} \sum_{i=1}^{p-1} \max\{x_i - x_{i+1}, 0\} . \tag{172}$$

The penalty is zero if $x_i \leq x_{i+1}$ and so it encourages the coefficients to be non-decreasing. It can be split and similarly as we did for the isotonic constraints:

$$\sum_{i=1}^{p-1} \max\{x_i - x_{i+1}, 0\} = \underbrace{\sum_{i=1}^{\lfloor p/2 \rfloor} \max\{x_{2i} - x_{2i+1}, 0\}}_{=g(x)} + \underbrace{\sum_{i=1}^{\lfloor (p-1)/2 \rfloor} \max\{x_{2i+1} - x_{2i+2}, 0\}}_{=h(x)} . \tag{173}$$

Let $\psi(x_1, x_2) \stackrel{\text{def}}{=} \max\{x_1 - x_2, 0\}$. If we denote by ∂_i the subgradient with respect to the i-th coordinate we have

$$\partial_1 \psi(x_1, x_2) = \begin{cases} 0 & \text{if } x_1 < x_2 \\ [0, 1] & \text{if } x_1 = x_2 \\ 1 & \text{if } x_1 > x_2 \end{cases} \quad \text{and} \quad \partial_2 \psi(x_1, x_2) = \begin{cases} 0 & \text{if } x_1 < x_2 \\ [-1, 0] & \text{if } x_1 = x_2 \\ -1 & \text{if } x_1 > x_2 \end{cases} \tag{174}$$

then it is easy to verify using the subdifferential inclusion of Lemma 1 that its proximal operator is given by

$$\mathbf{prox}_{\gamma \psi}(x_1, x_2) = \begin{cases} (x_1, x_2) & \text{if } x_1 \leq x_2 \\ (x_1 - \gamma, x_2 + \gamma) & \text{if } x_1 - \gamma \geq x_2 + \gamma \\ ((x_1 + x_2)/2, (x_1 + x_2)/2) & \text{otherwise .} \end{cases} \tag{175}$$

The proximal operator of the nearly isotonic penalty can then be computed by applying block-wise the above proximal operator.

Appendix D.3   Doubly stochastic constraints

The set of double stochastic matrices is composed of square matrices with nonnegative entries, each of whose rows and columns sum to 1, i.e., $\{X^T \mathbf{1} = \mathbf{1}, \mathbf{1}^T X = \mathbf{1}, X \geq 0\}$. The indicator function over this set can be split as

$$\underbrace{\imath\{X^T \mathbf{1} = \mathbf{1}, \mathbf{1}^T X = \mathbf{1}\}}_{=g(X)} + \underbrace{\imath\{X \geq 0\}}_{=h(X)} \quad . \tag{176}$$

For any $X \in \mathbb{R}^{p \times p}$, the proximal operator of g and h is then given in closed form as

$$\mathbf{prox}_{\gamma g}(X) = X + \left(\frac{1}{n}I + \frac{\mathbf{1}^T X \mathbf{1}}{n^2}I - \frac{1}{n}X\right)\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T X \tag{177}$$

$$\mathbf{prox}_{\gamma h}(X) = (X + |X|)/2 , \tag{178}$$

where $|\cdot|$ takes the absolute value componentwise and $\mathbf{1}$ denotes the p-dimensional vector of ones. The proof of this result can be found in (Lu et al., 2016).

Appendix E   Benchmarks

In this appendix we provide a more detailed description of the experiments, as well as present an extended set of benchmarks.

- Appendix E.1 discusses the implementation of the different algorithms.

- Appendix E.2 details the synthetic data generation process that we used in the synthetic benchmarks.

- Each of the experiments we run has a subsection in this appendix with an extended discussion and set of results. These are organized in the order they appear in Figure 1: Overlapping group lasso (Appendix E.3), total variation (Appendix E.4), sparse and low rank matrix recovery with trace norm ℓ_1 penalty (Appendix E.5), and nearly isotonic penalty (Appendix E.6).

Appendix E.1   Implementation details

Particular care has been taken to ensure a fair and extensive comparison. We have implemented all algorithms in Python, and used the just-in-time compiler Numba for some non-vectorizable computationally demanding parts such as computing the fused lasso proximal operator in the total variation-regularized problems. We profiled and optimized all algorithms equally, and in some cases made modifications that depart from the canonical description when this improved significantly performance (see TOS-AOLS below).

All methods have been implemented in Python and except TOS-AOLS are part of the C-OPT optimization package (Pedregosa, 2018).

PDHG.   The PDHG or Condat-Vũ algorithm (Condat, 2013b; Vũ, 2013) depends on two step-size parameters τ and σ. We parametrize the second step-size as $\sigma = \beta/\tau$ (this gave a better results than the parametrization $\sigma = \beta\tau$ used by Malitsky & Pock (2018) for the range β considered) then by (Condat, 2013b, Theorem 3.1) τ needs to be of the form

$$\tau \leq \frac{2(1-\beta)}{L_f} \ , \ \ \text{with } \beta < 1 \ . \tag{179}$$

With the above value of τ, we tested three different values for β, $\beta = 0.1, 0.5, 0.9$, and selected the one that had overall better performance (we found this to be $\beta = 0.5$) hence giving it a slight advantage with respect to other methods that do not require step-size tuning.

Adaptive PDHG.   Although one of the step-sizes is computed by a line search method, due to the two step-sizes it still requires to select the parametrization constant β. This parameter was computed the same way than for PDHG, except in this case we found the best performing value was $\beta = 0.1$. In this method the step-size is allowed to grow, and the initial value for the step-size in next iterate needs to be selected from an interval of admissible values. Following the experimental section of (Malitsky & Pock, 2018), we always take the maximum admissible step-size.

As it is visible from some of the benchmarks (see e.g., Figures 3, 6, 7), the method seems to suffer more than other methods from numerical instabilities in the high precision regime (typically when the suboptimality has reached 10^{-10}). We investigated this issue and developed an implementation that uses Moreau's decomposition to replaces $\mathbf{prox}_{\tau h^*}$ by $\mathbf{prox}_{h/\tau}$, thinking that perhaps the different scales between τ and β/τ would be the cause of the instability, but this did not solve the issue. Interestingly, these instabilities did not appear on the non adaptive variant, despite both algorithms share much of the same code.

TOS-AOLS.   We implemented the averaged operator line search method of Giselsson et al. (2016), using TOS as the averaged operator and name it TOS-AOLS. Our implementation has a crucial improvement with respect to the version detailed in the original reference that we found was crucial to obtain competitive results. In our implementation, the line search multiplicative factor α_t (in their notation) is not bound to start from an upper bound $\alpha_{\max}$ as is described in the reference but is instead incremented as $\alpha_{t+1} = 1.05\alpha_t$ in case of success of the line search condition. This gave much better empirical results.

Adaptive TOS. As described in §2.

Appendix E.2   Synthetic data generation

For the experiments with synthetic data (except for the image deblurring task), we followed the same data generation process as in (Agarwal et al., 2010), which generates a design matrix $\{a_i\}_{i=1}^n$ according to the recursion

$$a_1 = z_1 \tag{180}$$
$$a_i = z_i + p a_{j-1} \text{ for } i > 1 \tag{181}$$

with z_i sampled from a standardized Gaussian distribution. $p \in [0,1)$ is a correlation parameter that makes the problem more ill-conditioned as $p \to 1$. This is necessary as otherwise the resulting dataset has an unrealistic perfect spectrum. In the experiments we chose $p = 0.95$, which for a matrix of size 65×65 gives a condition number of around 60. Although this value of p might seem high, the resulting condition number is in fact smaller than the one found in the real datasets. For example, the `real-sim` dataset has a condition over 200.

Appendix E.3   Overlapping group lasso benchmarks

We consider an overlapping group lasso penalty with the following groups of size 10: $\{\{1,\ldots,10\}, \{8,\ldots,18\}, \{16,\ldots,26\},\ldots\}$, where each group has an overlap of 2 coefficient with the previous groups. This is modeled after the synthetic experiments of in (Jacob et al., 2009, §9.1). This set of groups can be split into two set $\mathcal{G}$ and $\mathcal{H}$ of disjoint groups by including in $\mathcal{G}$ the odd groups and in $\mathcal{H}$ the even ones, i.e., $\mathcal{G} = \{\{\{1,\ldots,10\}, \{16,\ldots,26\},\ldots\}, \mathcal{H} = \{\{8,\ldots,18\}, \{24,\ldots,34\},\ldots\}$.

With this notation, we can write the overlapping group lasso-penalized logistic regression problem as follows,

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \underbrace{\frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-b_i a_i^T x))}_{=f(x)} + \lambda \underbrace{\sum_{G \in \mathcal{G}} \|[x]_G\|}_{=g(x)} + \lambda \underbrace{\sum_{H \in \mathcal{H}} \|[x]_H\|}_{=h(x)} , \tag{182}$$

where λ is a regularization parameter and $\{b_i, a_i\}_{i=1}^n$ is the dataset that we will discuss later.

Lipschitz constant of the proximal term. For any set of disjoint groups $\mathcal{G}$ and any vector x we have $\sum_{G \in \mathcal{G}} \|x\| \leq \sqrt{|\mathcal{G}|}\|x\|$, where $|\mathcal{G}|$ denotes the cardinality of $\mathcal{G}$. Hence, the Lipschitz constant of h is upper bounded by $\lambda\sqrt{|\mathcal{G}|}$. This is the value we used in the experiments.

Dasets. We consider the following three datasets:

- The `real-sim` dataset, retrieved from the libsvm dataset collection.[2] The size of this dataset is $n = 72,309, p = 20,958$ and a density of 2% (i.e., 98% of zero coefficients in the data matrix).

- The `RCV1` dataset (Lewis et al., 2004), also retrieved from the `libsvm` dataset collection. The dimensions in this case are $n = 20,242$, $p = 677,399$, with a density of 0.1%.

- A synthetic dataset, in which a_i is generated according to Appendix E.2 and a vector of "ground truth" is generated by randomly selecting 10 groups and setting them to a random value generated from a standard Gaussian distribution (in this case the data matrix is fully dense). In Python, for a problem of size $n = 100$, $p = 1002$:

```python
import numpy as np
np.random.seed(0)

n_samples, n_features = 100, 1002
```

```
groups = [np.arange(8 * i, 8 * i + 10) for i in range(125)]

ground_truth = np.zeros(n_features)
g = np.random.randint(0, len(groups), 10)
for i in g:
    ground_truth[groups[i]] = np.random.randn()
```

Benchmarks. The results for this model and the above datasets can be seen in Figure 2 for the `real-sim` and `RCV1` datasets and in Figure 3 for the synthetic dataset. On 11 out of 12 cases the Adaptive TOS (variant 2) algorithm is the best performing method, and only marginally slower in the other case.
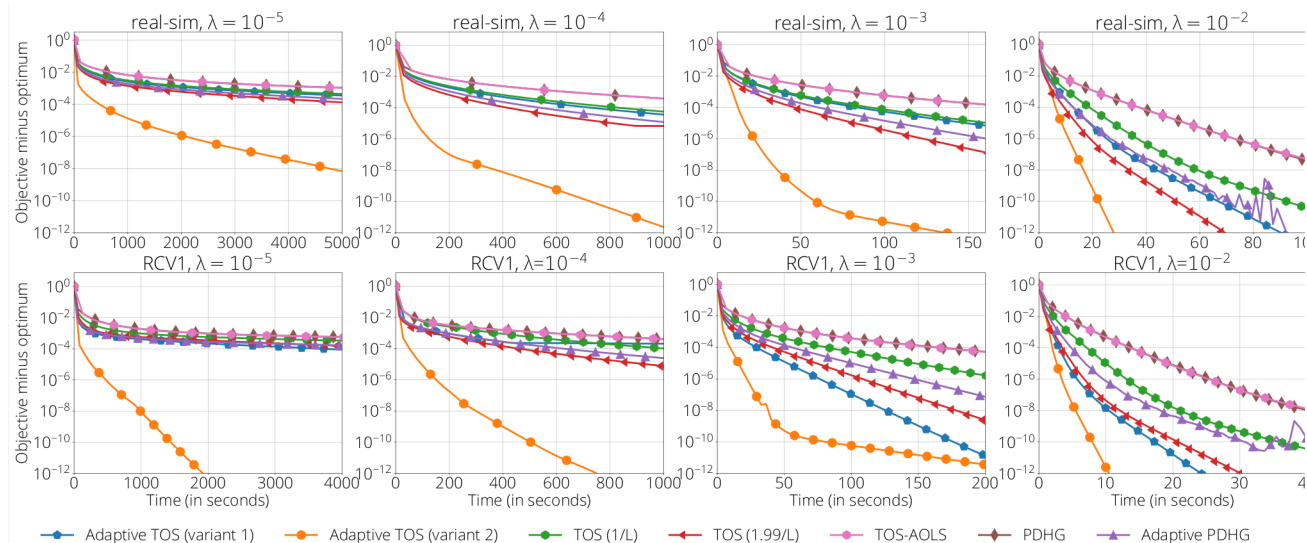
Figure 2: **Logistic regression with overlapping group lasso penalty, text datasets**. Top row: `real-sim` dataset, bottom row: `RCV1` dataset. The columns denote the amount of regularization, from a parameter λ giving $\approx 50\%$ of zero coefficients (left) to a parameter giving $\approx 5\%$ of zero coefficients
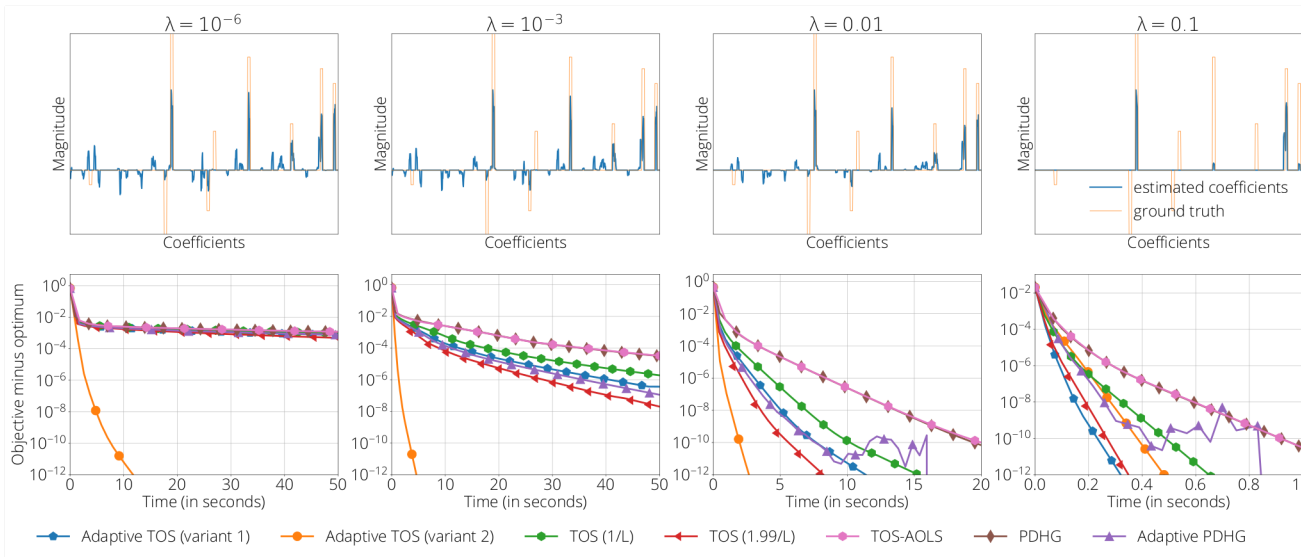
Figure 3: **Logistic regression with overlapping group lasso penalty, simulation dataset**. Top row: ground truth and estimated coefficients. Bottom row: time vs suboptimality comparison. Columns represent different values of the λ regularization parameter

Appendix E.4   Total variation benchmarks (image deblurring with known blur operator)

In this benchmark we consider a classical image deblurring task with known blur operator. For this we choose a natural image shown below as "Original image" and generated another image by convolving the original image with a blur kernel and adding standardized Gaussian noise, shown below as "Observed image". The blur kernel is displayed in bottom left corner of Observed image
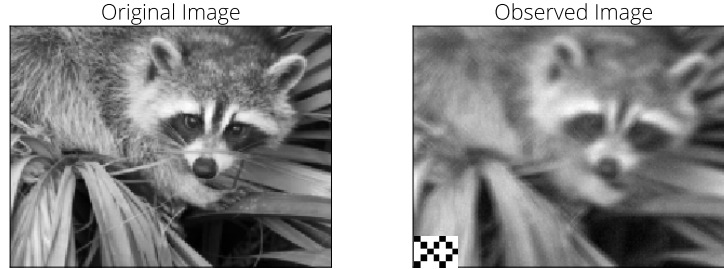

Figure 4: **Left**: Original image. **Right**: Observed image, obtained by convolving with the kernel shown in lower left corner and with added Gaussian noise. Columns represent different values of the λ regularization parameter

The image recovery can then be posed as least squares problem with a total variation penalty (Rudin et al., 1992; Chambolle & Pock, 2016) of the form:

$$\underset{\boldsymbol{X}\in\mathbb{R}^{p\times q}}{\text{minimize}} \|\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{X}\|^2 + \lambda\|\boldsymbol{X}\|_{\text{TV}} \,, \tag{183}$$

$$\equiv \underset{\boldsymbol{X}\in\mathbb{R}^{p\times q}}{\text{minimize}} \underbrace{\|\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{X}\|^2}_{=f(\boldsymbol{X})} + \lambda\underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q-1}|\boldsymbol{X}_{i,j+1}-\boldsymbol{X}_{i,j}|}_{=g(\boldsymbol{X})} + \underbrace{\sum_{j=1}^{q}\sum_{j=1}^{p-1}|\boldsymbol{X}_{i+1,j}-\boldsymbol{X}_{i,j}|}_{=h(\boldsymbol{X})}, \tag{184}$$

where $\boldsymbol{Y}$ is the observed image, $\boldsymbol{B}$ is the blurring (linear) operator and g and h are fused lasso (also known as 1-dimensional total variation) penalties acting on the columns and rows of $\boldsymbol{X}$ respectively. Their proximal operator was implemented using the algorithm of Condat (2013a).

Alternatively, PDHG can solve this problem using a different splitting in which $g(\boldsymbol{x}) = 0$ and $\lambda\|\boldsymbol{X}\|_{\text{TV}} = h(\boldsymbol{K}\boldsymbol{x})$, where h is the ℓ_1 penalty and $\boldsymbol{K}$ is the matrix of first order finite differences. We compared both splittings and concluded that the splitting of Eq. (184) gave better results.

Lipschitz constant of the proximal term.   For a matrix $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ we have

$$\|\boldsymbol{X}\|_{\text{FL}} = \|\boldsymbol{K}\text{vec}(\boldsymbol{x})\|_1 \leq \|\boldsymbol{K}\|_1\|\text{vec}(\boldsymbol{x})\|_1 \leq 2\|\text{vec}(\boldsymbol{x})\|_1 \leq 2\sqrt{pq}\|\text{vec}(\boldsymbol{x})\| \,, \tag{185}$$

and so the Lipschitz constant of h can be bounded by $2\sqrt{pq}$. This is the quantity we used in the experiments.

Benchmarks.   We show recovered image as well as benchmarks in Figure 5. In this case, we found that the fixed step-size strategy with step-size $1.99/L$ marginally outperformed the Adaptive TOS on 3 out of 4 cases (but never by more than a factor of 1.5). Interestingly, this same method performed the worse on the high regularization setting (right column).
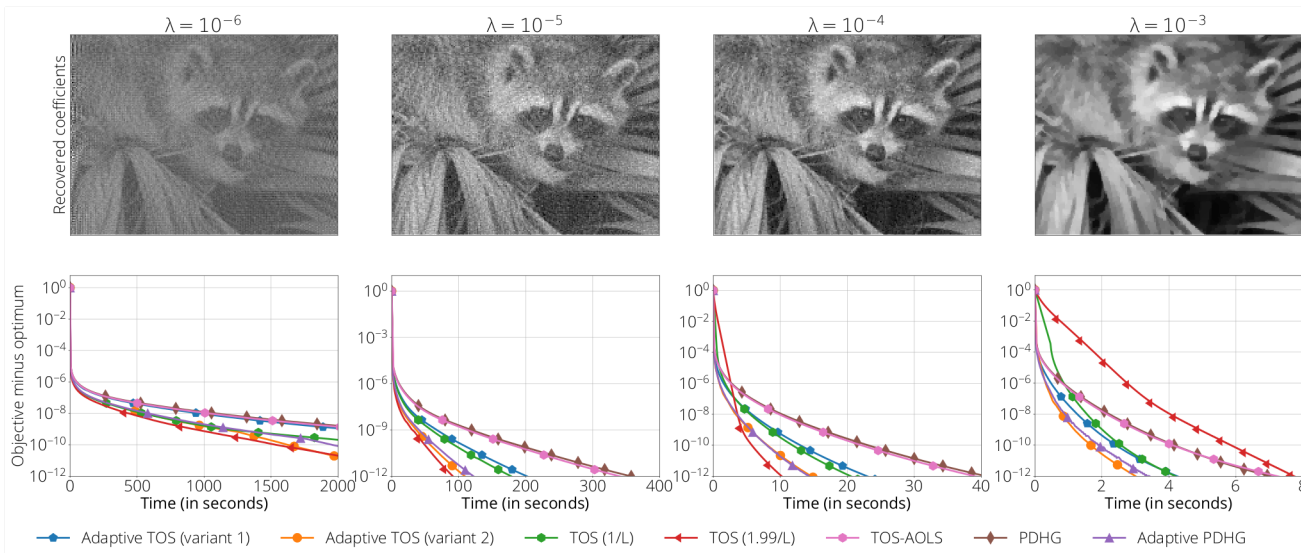
Figure 5: **Least Squares regression with total variation penalty**. Top row: recovered image. Bottom row: time vs suboptimality comparison. Columns represent different values of the λ regularization parameter

Appendix E.5 Sparse and low rank matrix recovery benchmarks

We generated a sparse and low rank 20×20 symmetric matrix as the one in the right using the procedure detailed in (Richard et al., 2012). This matrix can be seen in the right. We will denote this matrix by $\boldsymbol{X}_{\text{truth}}$. We then generate the target values as b_i as $b_i = \boldsymbol{a}_i^T \text{vec}(\boldsymbol{X}_{\text{truth}}) + \varepsilon_i$, where ε is a random noise, generated from a zero-mean, unit variance Gaussian distribution. We generated 200 of these samples, so that the problem has twice as many samples as features.
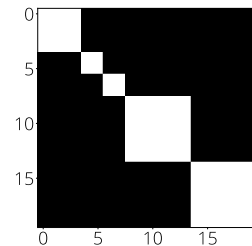
It has been shown that it is possible to promote sparse and low rank solutions by using a penalty composed of a trace (or nuclear) norm and a (vector) ℓ_1 norm (see e.g., (Richard et al., 2012)), where the trace norm is the sum of the absolute values and we will denote it $\| \cdot \|_*$.

In all, we consider the following problem:

$$\underset{\boldsymbol{X} \in \mathbb{R}^{20 \times 20}}{\text{minimize}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (b_i - \boldsymbol{a}_i^T \text{vec}(\boldsymbol{X}))^2}_{=f(\boldsymbol{X})} + \underbrace{\lambda \|\boldsymbol{X}\|_*}_{=g(\boldsymbol{X})} + \underbrace{\mu \|\text{vec}(\boldsymbol{X})\|_1}_{=h(\boldsymbol{X})} , \tag{186}$$

where the proximal operator of g is given by soft thresholding the singular values (Cai et al., 2010).

Lipschitz constant of proximal term. By the properties of vector norms, for a matrix $\boldsymbol{X}$ of size $p \times q$ we have that $\|\text{vec}(\boldsymbol{X})\|_1 \leq \sqrt{pq}\|\text{vec}(\boldsymbol{X})\|$. Hence, we can bound the Lipschitz constant of h by $\lambda \sqrt{pq}$.

Benchmarks. The results of the time comparison and recovered coefficients can be see in the Figure 6 below. For simplicity we consider $\lambda = \mu$, but of course more accurate coefficients could be recovered by tuning both regularization parameters.
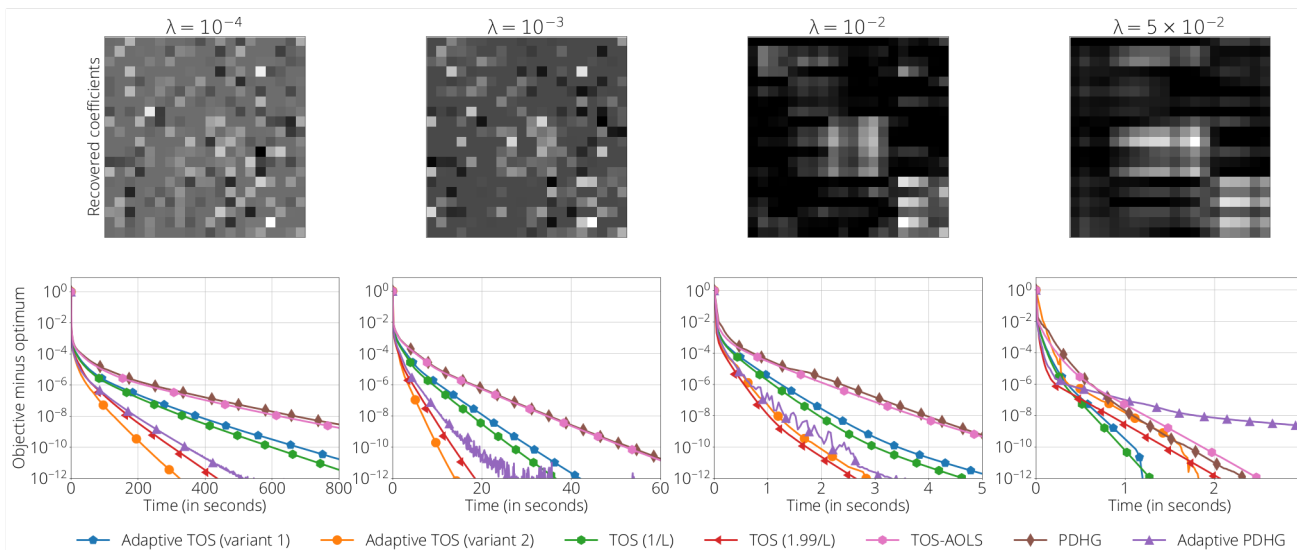

Figure 6: **Least Squares with trace norm + ℓ_1 penalty.** Top row: recovered coefficients from problem (186). Bottom row: time vs suboptimality comparison. Columns represent different values of the λ regularization parameter.

Appendix E.6 Nearly isotonic penalty benchmarks

Given an input dataset $\{(b_i, \boldsymbol{a}_i)\}_{i=1}^n$, which we generated according to Appendix E.2, we consider a logistic regression problem with nearly isotonic penalty on the weights:

$$\underset{\boldsymbol{x} \in \mathbb{R}^p}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-b_i \boldsymbol{a}_i^T \boldsymbol{x})) + \lambda \sum_{i=1}^{p-1} \max\{\boldsymbol{x}_i - \boldsymbol{x}_{i+1}\} \tag{187}$$

$$\equiv \underset{\boldsymbol{x} \in \mathbb{R}^p}{\text{minimize}} \; \underbrace{\frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-b_i \boldsymbol{a}_i^T \boldsymbol{x}))}_{=f(\boldsymbol{x})} + \underbrace{\lambda \sum_{i=1}^{\lfloor p/2 \rfloor} \max\{\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i+1}\}}_{=g(\boldsymbol{x})} + \underbrace{\lambda \sum_{i=1}^{\lfloor (p-1)/2 \rfloor} \max\{\boldsymbol{x}_{2i+1} - \boldsymbol{x}_{2i+2}\}}_{=h(\boldsymbol{x})} \;,$$

$$\tag{188}$$

where λ is a regularization parameter. The problem dimension in this case is $n = 100$, $p = 50$, and an noise of variance of 5 (instead of 1 as described in Appendix E.2). We increased the noise variance to make the estimation of the ordering of coefficients more challenging (see figure below).

Benchmarks. We show recovered coefficients as well as the time benchmarks in Figure 5. The Adaptive TOS method is the best performing method in all four cases, and the difference is larger in the low regularization setting.

Lipschitz constant of proximal term. Let $\|\cdot\|_{\text{ISO}}$ denote the nearly isotonic pseudo-norm. Then we can bound it in terms of the euclidean norm as follows, where as in Appendix E.4, $\boldsymbol{K}$ is the matrix of first order differences:

$$\|\boldsymbol{x}\|_{\text{ISO}} \le \|\boldsymbol{K}\boldsymbol{x}\|_1 \le \|\boldsymbol{K}\|_1 \|\boldsymbol{x}\|_1 = 2\|\boldsymbol{x}\|_1 \le 2\sqrt{p}\|\boldsymbol{x}\| \tag{189}$$
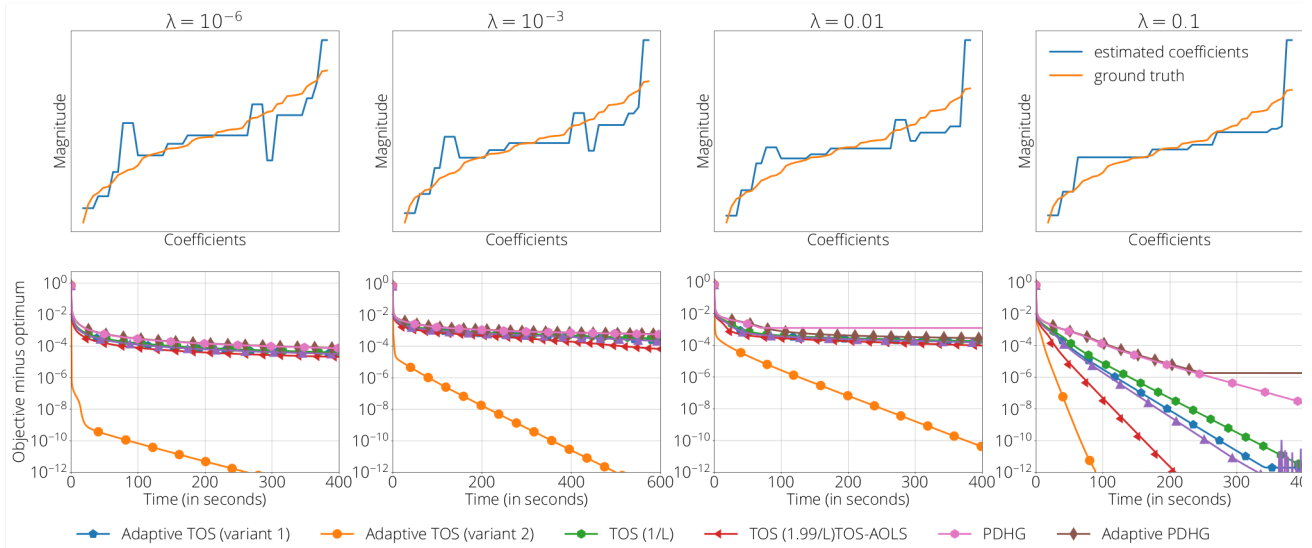

Figure 7: **Logistic regression with nearly isotonic penalty**. Top row: ground truth and recovered coefficients. Bottom row: time vs suboptimality comparison. Columns represent different values of the λ regularization parameter