Linearly Convergent Frank-Wolfe with Backtracking Line-Search

Fabian Pedregosa **Geoffrey Négiar** **Armin Askari** **Martin Jaggi**
Google Research University of California, Berkeley EPFL, Lausanne

Abstract

Structured constraints in Machine Learning have recently brought the Frank-Wolfe (FW) family of algorithms back into the spotlight. While the classical FW algorithm has poor local convergence properties, Away-steps FW and Pairwise FW have emerged as improved variants with faster convergence. However, these improved variants suffer from two practical limitations: they require at each iteration to solve a 1-dimensional minimization problem to set the step-size and also require the Frank-Wolfe linear subproblems to be solved exactly. In this paper, we propose variants of Away-steps and Pairwise FW that lift both restrictions simultaneously. The proposed methods set the step-size based on a sufficient decrease condition, and do not require prior knowledge of the objective. Furthermore, they inherit all the favorable convergence properties of the exact line-search version, including linear convergence for strongly convex functions over polytopes. Benchmarks on different machine learning problems illustrate large performance gains of the proposed variants.

1 Introduction

The Frank-Wolfe (FW) or conditional gradient algorithm (Frank and Wolfe, 1956; Levitin and Polyak, 1966; Demyanov and Rubinov, 1967) is a method for constrained optimization that solves problems of the form

$$\underset{\boldsymbol{x}\in\operatorname{conv}(\mathcal{A})}{\text{minimize}}\; f(\boldsymbol{x})\,, \tag{OPT}$$

where f is a smooth function for which we have access to its gradient and $\operatorname{conv}(\mathcal{A})$ is the convex hull of $\mathcal{A}$; a bounded but potentially infinite set of elements in $\mathbb{R}^p$ which we will refer to as *atoms*.

The FW algorithm is one of the oldest methods for non-linear constrained optimization and has experienced a renewed interest in recent years due to its applications in machine learning and signal processing (Jaggi, 2013). Despite some favorable properties, the local convergence of the FW algorithm is known to be slow, achieving only a sublinear rate of convergence for strongly convex functions when the solution lies in the boundary (Canon and Cullum, 1968). To overcome these limitations, variants of the FW algorithms with better convergence properties have been proposed. Two of these variants, the Away-steps FW (Guélat and Marcotte, 1986) and Pairwise FW (Lacoste-Julien and Jaggi, 2015) enjoy a linear rate of convergence over polytopes (Lacoste-Julien and Jaggi, 2015).

Despite this theoretical breakthrough, Away-steps and Pairwise FW are not yet practical off-the-shelf solvers due to two main limitations. The first and most important is that both variants rely on an *exact line-search.* That is, they require to solve at each iteration 1-dimensional subproblems of the form

$$\mathop{\mathbf{arg\,min}}_{\gamma\in[0,\gamma_{\max}]} f(\boldsymbol{x}_t+\gamma\boldsymbol{d}_t)\ , \tag{1}$$

where $\boldsymbol{d}_t$ is the update direction and $\gamma_{\max}$ is the maximum admissible step-size. In a few cases like quadratic objectives, the exact line-search subproblem has a closed form solution. In most other cases, it is a costly optimization problem that needs to be solved at each iteration, making these methods impractical. The second limitation is that they require access to an *exact Linear Minimization Oracle* (LMO), which leaves out important cases like minimization over a trace norm ball where the LMO is computed up to some predefined tolerance. In this paper we develop methods that lift both limitations simultaneously.

Our **main contribution** is the design and analysis of variants of Away-steps and Pairwise FW that *i*) don't

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

	Related work	non-convex analysis	approximate subproblems	linear convergence	adaptive step-size	bounded backtracking
Frank-Wolfe	*This work*	✓	✓	✓	✓	✓
	(Lacoste-Julien and Jaggi, 2015)	✗	✗	✓	✗	N/A
	(Beck et al., 2015)	✗	✓†	✗	✓	✗
	(Dunn, 1980)	✓	✗	✗	✓	✗
MP	*This work*	✓	✓	✓	✓	✓
	(Locatello et al., 2017)	✗	✓	✓	✗	N/A

Table 1: **Comparison with related work.** *non-convex analysis*: convergence guarantees for problems with a non-convex objective. *approximate subproblems*: convergence guarantees cover the case in which linear subproblems are solved approximately. *linear convergence*: guaranteed linear rate of convergence (under hypothesis). *adaptive step-size*: step-size is set using local information of the objective. *bounded backtracking*: explicit bound for the total number of inner iterations in adaptive step-size methods. † = assumes cartesian product domain.

require access to an exact line-search or knowledge of properties of the objective like its curvature or Lipschitz constant, and *ii*) admits the FW subproblems to be solved approximately. We describe our approach in §2. Although our main motivation is to develop practical variants of Away-steps and Pairwise FW, we also show that this technique extends to other methods like FW and Matching Pursuit.

We develop in §3 a convergence rate analysis for the proposed methods. The obtained rates match asymptotically the best known bounds on convex, strongly convex and non-convex problems, including linear convergence for strongly convex functions.

Finally, we show in §4 benchmarks between the proposed and related methods, and discuss the importance of large step-sizes in Pairwise FW.

1.1 Related work

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

Away-Steps FW (Guélat and Marcotte, 1986) is a popular variant of FW that adds the option to move away from an atom in the current representation of the iterate. In the case of a polytope domain, it was recently shown to enjoy a linear convergence rate for strongly convex objectives (Garber and Hazan, 2013; Lacoste-Julien and Jaggi, 2015). Pairwise FW (Lacoste-Julien and Jaggi, 2015) simplifies the above-described variant by replacing the two kinds of steps by a single step modifying the weights of only two atoms. It generalizes the algorithm of Mitchell et al. (1974) used in geometry and SMO (Platt, 1998) for training SVMs. These methods all require exact line-search.

Variants of FW that, like the proposed methods, set the step-size based on a local decrease condition have been described by Dunn (1980) and Beck et al. (2015), but none of these methods achieve a linear convergence rate to the best of our knowledge.

Matching Pursuit (MP) (Mallat and Zhang, 1993) is an algorithm for constrained optimization problems of the form (OPT) with $\mathrm{conv}(\mathcal{A})$ replaced by $\mathrm{linspan}(\mathcal{A})$, the linear span of $\mathcal{A}$. Locatello et al. (2018) has recently shown that MP and FW are deeply related. We show that our algorithm and convergence results also extend naturally to MP, and as a byproduct of our analysis we obtain the first convergence rate for MP on non-convex objectives to the best of our knowledge.

Notation. Throughout the paper we denote vectors and vector-valued functions in lowercase boldface (e.g. $\boldsymbol{x}$ or $\mathbf{arg\,min}$), matrices in uppercase boldface letters (e.g. $\boldsymbol{D}$), and sets in caligraphic letters (e.g., $\mathcal{A}$). We say a function f is L-smooth if it is differentiable and its gradient is L-Lipschitz continuous, that is, if it verifies $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y}$ in the domain. A function is μ-strongly convex if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex. $\|\cdot\|$ denotes the euclidean norm.

2 Methods

In this section we describe the core part of our contribution, which is a strategy to select the step-size in FW-type algorithms.

Since this strategy can be applied very broadly to Frank-Wolfe variants including Away-steps, Pairwise, classical FW and Matching Pursuit, we describe it within the context of a generic FW-like algorithm. This generic algorithm is detailed in Algorithm 1 and depends on two key functions: `update_direction` and `step_size`. The first one computes the direction that we will follow to compute the next iterate and its implementation will depend on the FW variant. The second one will choose an appropriate step-size based upon local information of the objective and is the key novelty of this algorithm. We now describe them in more detail.

Algorithm 1: Generic FW with backtracking

```
1 Input: $\boldsymbol{x}_0 \in \operatorname{conv}(\mathcal{A})$, initial Lipschitz estimate
  $L_{-1} > 0$, tolerance $\varepsilon \geq 0$, subproblem quality
  $\delta \in (0, 1]$
2 for $t = 0, 1 \ldots$ do
3     $\boldsymbol{d}_t, \gamma_t^{\max} = \text{update\_direction}(\boldsymbol{x}_t, \nabla f_t)$
4     $g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle$
5     if $g_t \leq \delta\varepsilon$ then return $\boldsymbol{x}_t$;
6     $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \gamma_t^{\max})$
7     $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
```

2.1 Update direction

In this subsection we describe `update_direction` in Algorithm 1, the function that computes the update direction $\boldsymbol{d}_t$ and the maximum allowable step-size $\gamma_t^{\max}$. While its implementation varies according to the FW variant, all of them require to solve one or two linear problems, often referred to as linear minimization oracle (LMO).

The first of these subproblems is the same for all variants and consists in finding $\boldsymbol{s}_t$ in the domain such that:

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle \ . \tag{2}$$

Here, we introduce a *subproblem quality* parameter $\delta \in (0, 1]$ that allows this subproblem to be solved approximately. When $\delta = 1$, the problem is solved exactly and becomes $\mathbf{arg\,min}_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle$, which consists in selecting the atom that correlates the most with the steepest descent direction, $-\nabla f(\boldsymbol{x}_t)$.

Away-steps and Pairwise FW will also require to solve another linear subproblem, this time over the *active set* $\mathcal{S}_t$. This is the set of atoms with non-zero weight in the decomposition of $\boldsymbol{x}_t$. More formally, the active set $\mathcal{S}_t \subseteq \mathcal{A}$ is the set of atoms that have non-zero weight $\boldsymbol{\alpha}_{\boldsymbol{s},t} > 0$ in the expansion $\boldsymbol{x}_t = \sum_{\boldsymbol{s} \in \mathcal{S}_t} \boldsymbol{\alpha}_{\boldsymbol{v},t} \boldsymbol{s}$.

The linear subproblem that needs to be solved consists in finding $\boldsymbol{v}_t$ such that:

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v}_t \rangle \leq \delta \min_{\boldsymbol{v} \in S_t} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v} \rangle \ . \tag{3}$$

Unlike the previous linear subproblem, this time the problem is over the typically much smaller active set $\mathcal{S}_t$. As before, $\delta \in (0, 1]$ allows this subproblem to be solved approximately. When $\delta = 1$, the subproblem becomes $\mathbf{arg\,max}_{\boldsymbol{v} \in \mathcal{S}_t} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle$, which can be interpreted as selecting the atom in the active set that correlates the most with the steepest ascent direction $\nabla f(\boldsymbol{x}_t)$.

Algorithm 2: Backtracking for FW variants

```
1 Procedure step_size($f$, $\boldsymbol{d}_t$, $\boldsymbol{x}_t$, $g_t$, $L_{t-1}$, $\gamma_{\max}$)
2     Choose $\tau > 1$, $\eta \leq 1$
3     Choose $M \in [\eta L_{t-1}, L_{t-1}]$
4     $\gamma = \min\left\{g_t/(M\|\boldsymbol{d}_t\|^2), \gamma_{\max}\right\}$
5     while $f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t) > Q_t(\gamma, M)$ do
6         $M = \tau M$
7         $\gamma = \min\left\{g_t/(M\|\boldsymbol{d}_t\|^2), \gamma_{\max}\right\}$
8     end
9     return $\gamma$, $M$
```

FW, AFW and PFW then combine the solution to these linear subproblems in different ways, and Line 3's `update_direction` is implemented as:

- FW returns $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{x}_t$ and $\gamma_t^{\max} = 1$: the next iterate will be a convex combination of $\boldsymbol{x}_t$ and $\boldsymbol{s}_t$.
- AFW considers directions $\boldsymbol{s}_t - \boldsymbol{x}_t$ and $\boldsymbol{x}_t - \boldsymbol{v}_t$, and chooses the one that correlates the most with $-\nabla f(\boldsymbol{x}_t)$. $\gamma_t^{\max} = 1$ if $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{x}_t$ and $\alpha_{\boldsymbol{v}_t}/(1 - \alpha_{\boldsymbol{v}_t})$ otherwise, where $\alpha_{\boldsymbol{v}_t}$ is the weight associated with $\boldsymbol{v}_t$ in the decomposition of $\boldsymbol{x}_t$ as a convex combination of atoms.
- PFW uses $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{v}_t$, shifting weight from $\boldsymbol{v}_t$ to $\boldsymbol{s}_t$ in our current iterate, and $\gamma_t^{\max} = \alpha_{\boldsymbol{v}_t}$.
- MP uses $\boldsymbol{d}_t = \boldsymbol{s}_t$ and $\gamma^{\max} = +\infty$, since the constraint set is not bounded.

2.2 Backtracking line-search

In this subsection we describe the step-size selection routine `step_size` (Algorithm 2). This is the main novelty in the proposed algorithms, and allows for the step-size to be computed using only *local* properties of the objective, as opposed to other approaches that use global quantities like the gradient's Lipschitz constant. As we will see in §4, this results in step-sizes that are often more than an order of magnitude larger than those estimated using global quantities.

Minimizing the exact line-search objective $\gamma \mapsto f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t)$ yields the highest decrease in objective but can be a costly optimization problem. To overcome this, we will replace the exact line-search objective with the following quadratic approximation:

$$Q_t(\gamma, M) = f(\boldsymbol{x}_t) - \gamma g_t + \frac{\gamma^2 M}{2} \|\boldsymbol{d}_t\|^2 \ . \tag{4}$$

This approximation has the advantage that its minimum over $\gamma \in [0, \gamma^{\max}]$ can be computed in closed form, which gives the step-size used in line 4:

$$\gamma_M^\star = \min\left\{\frac{g_t}{M\|\boldsymbol{d}_t\|^2}, \gamma^{\max}\right\} \ , \tag{5}$$

The quality of this quadratic approximation will depend on the *Lipschitz estimate* parameter M. This parameter needs to be carefully selected to maintain the convergence guaratees of exact line-search, while keeping the number of objective function evaluations to a minimum.

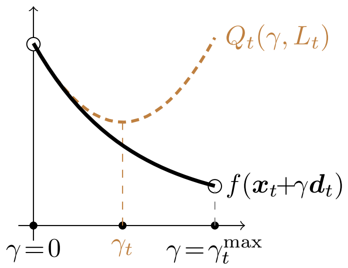
This is achieved through the strategy implemented in Algorithm 2. The algorithm initializes the Lipschitz estimate M to a value between ηL_{t-1} and the previous iterate L_{t-1}, where η is a user-defined parameter (default values discussed later). A value of $\eta = 1$ is admissible but would not allow the Lipschitz estimate to decrease through the optimization, and we have observed empirically a drastic benefit in doing so.

The algorithm then defines a candidate step-size γ (Line 4) and checks whether the following *sufficient decrease* condition is verified for this step-size

$$f(\boldsymbol{x}_t+\gamma\boldsymbol{d}_t) \leq Q_t(\gamma, M)\,, \gamma=\min\left\{g_t/(M\|\boldsymbol{d}_t\|^2), \gamma^{\max}\right\}.$$

If it is not verified, we increase this constant by a power factor of $\tau > 1$ (Line 6). By the properties of L-smooth functions, we know that this condition is verified for all $M \geq L$, and so this loop has a finite termination.

Once this condition is verified, the current step-size is accepted and the value of M is assigned the name L_t. Geometrically, the sufficient decrease condition ensures that the quadratic approximation is an upper bound at its constrained minimum of the line-search objective. We emphasize that this upper bound does not need to be a global one, as it only holds at γ_t. This allows for smaller L_t than the global Lipschitz constant L, and therefore larger step-sizes. As we will see in §3, this translates into faster convergence rates that depend on L_t, as well as faster empirical convergence (§4).



Default and initial parameters. Algorithm 1 requires an (arbitrary) initial value for the Lipschitz estimate L_{-1}. We found the following heuristic using the definition of Lipschitz continuity of the gradient to work well in practice. Select a small constant ε, say 10^{-3}, and compute an initial value as $L_{-1} = \|\nabla f(\boldsymbol{x}_0) - \nabla f(\boldsymbol{x}_0 + \varepsilon\boldsymbol{d}_0)\|/(\varepsilon\|\boldsymbol{d}_t\|)$.

The `step_size` depends on hyperparameters η and τ. Although the algorithm is guaranteed to converge for any $\eta \leq 1, \tau > 1$, we recommend $\eta = 0.9, \tau = 2$, as we found that it performs well in a variety of scenarios. These are the values used throughout benchmarks §4.

This method also requires to choose the initial value of the Lipschitz estimate M to a value between ηL_{t-1} and L_{t-1}. A choice that we found to work remarkably well in practice is to initialize it to

$$M = \text{clip}_{[\eta L_{t-1}, L_{t-1}]}\left(\frac{g_t^2}{2(f_{t-1} - f_t)\|\boldsymbol{d}_t\|^2}\right)\,. \qquad (6)$$

The value inside the clip function corresponds to the optimal value of M for a quadratic interpolation between the previous two iterates and the derivative of the line-search objective $f(\boldsymbol{x}_t + \gamma\boldsymbol{d}_t)$ at $\gamma = 0$. Since this value might be outside of the interval $[\eta L_{t-1}, L_{t-1}]$, we clip the result to this interval.

Pseudocode and implementation details. A practical implementation of these algorithms depends on other details that are not specific to the backtracking variant, such as efficiently maintaining the active-set in the case of Away-steps and Pairwise. For completeness, Appendix A contains a full pseudocode for all these algorithms. A Python implementation of these methods, as well as the benchmarks used in §4 will be made open source upon publication of this manuscript.

3 Analysis

In this section, we provide a convergence rate analysis of the proposed methods, showing that all enjoy a $\mathcal{O}(1/\sqrt{t})$ convergence rate for non-convex objectives (Theorem 2), a stronger $\mathcal{O}(1/t)$ convergence rate for convex objectives (Theorem 3), and for some variants linear convergence for strongly convex objectives over polytopes (Theorem 4).

Notation. In this section we make use of the following extra notation:

- For convenience we will refer to the variants of FW, Away-steps FW, Pairwise FW and MP with backtracking line-search as AdaFW, AdaAFW, AdaPFW and AdaMP respectively.
- We denote the *objective suboptimality* at step t as $h_t = f(\boldsymbol{x}_t) - \min_{\boldsymbol{x}\in\text{conv}(\mathcal{A})} f(\boldsymbol{x})$.
- *Good and bad steps.* Our analysis, as that of Lacoste-Julien and Jaggi (2015), relies on a notion of "good" and "bad" steps. We define bad steps as those that verify $\gamma_t = \gamma_t^{\max}$ and $\gamma_t^{\max} < 1$ and good steps as any step that is not a bad step. The name "bad steps" makes reference to the fact that we won't be able to bound non-trivially the improvement for these steps. For these steps we will only be able to guarantee that the objective is non-increasing. AdaAFW and AdaPFW both may have

Algorithm	Non-convex	Convex	Strongly convex
AdaAFW	$\mathcal{O}(\frac{1}{\delta\sqrt{t}})$	$\mathcal{O}(\frac{1}{\delta^2 t})$	$\mathcal{O}((1-\delta^2\rho)^t)$
AdaPFW	$\mathcal{O}(\frac{1}{\delta\sqrt{t}})$	$\mathcal{O}(\frac{1}{\delta^2 t})$	$\mathcal{O}((1-\delta^2\rho)^t)$
AdaFW	$\mathcal{O}(\frac{1}{\delta\sqrt{t}})$	$\mathcal{O}(\frac{1}{\delta^2 t})$	$\mathcal{O}(\frac{1}{\delta^2 t})$
AdaMP	$\mathcal{O}(\frac{1}{\delta\sqrt{t}})$	$\mathcal{O}(\frac{1}{\delta^2 t})$	$\mathcal{O}((1-\delta^2\rho_{\mathrm{MP}})^t)$

Table 2: **Convergence rate summary** on non-convex, convex and strongly convex objectives. For non-convex objectives, bound is on the minimum FW gap (MP gap in the case of AdaMP), in other cases its on the objective suboptimality.

bad steps. Let us denote by N_t the number of "good steps" up to iteration t. We can lower bound the number of good steps by

$$N_t \geq t/2 \text{ for AdaAFW}\,, \tag{7}$$

$$N_t \geq t/(3|\mathcal{A}|! + 1) \text{ for AdaPFW} \tag{8}$$

where it is worth noting that the last bound for AdaPFW requires the set of atoms $\mathcal{A}$ to be finite. The proof of these bounds can be found in Appendix C.1 and are a direct translation of those in (Lacoste-Julien and Jaggi, 2015). We have found these bounds to be very loose, as in practice the fraction of bad/good steps is negligible, commonly of the order of 10^{-5} (see last column of the table in Figure 1).

- *Average and maximum of Lipschitz estimates.* In order to highlight the better convergence rates that can be obtained by adaptive methods we introduce the average and maximum estimate over good step-sizes. Let $\mathcal{G}_t$ denote the indices of good steps up to iteration t. Then we define the average and maximum Lipschitz estimate as

$$\overline{L}_t \overset{\text{def}}{=} \frac{1}{N_t}\textstyle\sum_{k\in\mathcal{G}_t} L_k \tag{9}$$

$$L_t^{\max} \overset{\text{def}}{=} \max\nolimits_{k\in\mathcal{G}_t} L_k \tag{10}$$

respectively. In the worst case, both quantities can be upper bounded by $\max\{\tau L, L_{-1}\}$ (Proposition 2), which can be used to obtain asymptotic convergence rates. This bound is however very pessimistic. We have found that in practice $\overline{L}_t$ is often more than 100 times smaller than L (see second to last column of the table in Figure 1).

Our new convergence rates are presented in the following theorems, which consider the cases of non-convex, convex and strongly convex objectives. The results are discussed in §3.5 and the proofs can be found in Appendix D, Appendix E and Appendix F respectively.

3.1 Overhead of backtracking

Evaluation of the sufficient decrease condition Algorithm 2 requires two extra evaluations of the objective function. If the condition is verified, then it is only evaluated at the current and next iterate. FW requires anyway to compute the gradient at these iterates, hence in cases in which the objective function is available as a byproduct of the gradient this overhead becomes negligible.

Furthermore, we can provide a bound on the total number of evaluations of the sufficient decrease condition:

Theorem 1. *Let n_t be the total number of evaluations of the sufficient decrease condition up to iteration t. Then we have*

$$n_t \leq \left[1 - \frac{\log\eta}{\log\tau}\right](t+1) + \frac{1}{\log\tau}\max\left\{\log\frac{\tau L}{L_{-1}}, 0\right\}\ ,$$

This result highlights the trade-off faced when choosing η. Minimizing it with respect to η gives $\eta = 1$, in which case $(1 - \log\eta/\log\tau) = 1$ and so there's an asymptotically vanishing number of failures in the sufficient decrease condition. Unfortunately, $\eta = 1$ also forbids the Lipschitz estimate to decrease along the optimization. Ideally, we would like η small enough so that the Lipschitz estimate decreases when it can, but not too small so that we waste too much time in failed sufficient decrease evaluations.

As mentioned before, we recommend parameters $\eta = 0.9$, $\tau = 2$. With these values, we have that $\left[1 - \frac{\log\eta}{\log\tau}\right] \leq 1.16$, and so asymptotically no more than 16% of the iterates will result in more than one evaluations of the sufficient decrease condition.

3.2 Non-convex objectives

Gap function. Convergence rates for convex and strongly convex functions are given in terms of the objective function suboptimality or a primal-dual gap. As the gap upper-bounds (i.e. certifies) the suboptimality, the latter is a stronger result in this scenario. In the case of non-convex objectives, as is common for first order methods, we will only be able to guarantee convergence to a stationary point, defined as any element $\boldsymbol{x}^\star \in \operatorname{conv}(\mathcal{A})$ such that $\langle\nabla f(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star\rangle \geq 0$ for all $\boldsymbol{x} \in \operatorname{conv}(\mathcal{A})$ (Bertsekas, 1999).

Following Lacoste-Julien (2016); Reddi et al. (2016), for FW variants we will use as convergence criterion the FW gap, defined as $g^{\mathrm{FW}}(\boldsymbol{x}) = \max_{\boldsymbol{s}\in\mathrm{conv}(\mathcal{A})}\langle\nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{s}\rangle$. From the definition of stationary point it is clear that the FW gap is zero only at a stationary point. These rates are also valid for AdaMP, albeit for the more appropriate gap function g^{MP} detailed in Appendix D.

Theorem 2. *Let $\boldsymbol{x}_t$ denote the iterate generated by any of the proposed algorithms after t iterations, with $N_{t+1} \geq 1$. Then we have:*

$$\lim_{t\to\infty} g(\boldsymbol{x}_t) = 0 \qquad \textit{and} \tag{11}$$

$$\min_{k=0,\ldots,t} g(\boldsymbol{x}_k) \leq \frac{C_t}{\delta\sqrt{N_{t+1}}} = \mathcal{O}\left(\frac{1}{\delta\sqrt{t}}\right), \tag{12}$$

where $C_t = \max\{2h_0, L_t^{\max}\operatorname{diam}(\mathcal{A})^2\}$ and $g = g^{FW}$ is the FW gap for AdaFW, AdaAFW, AdaPFW and $C_t = \operatorname{radius}(\mathcal{A})\sqrt{2h_0\overline{L}_{t+1}}$ and $g = g^{MP}$ is the MP gap for AdaMP.

3.3 Convex objectives

For convex objectives we will be able to improve the results of Theorem 2. We will first state the convergence results for FW variants and then for MP.

For adaptive FW variants, we will be able to give an $\mathcal{O}(1/\delta^2 t)$ convergence rate on the primal-dual gap, which trivially implies a bound on the objective suboptimality. In order to define the primal-dual gap, we define the following *dual* objective function

$$\psi(\boldsymbol{u}) \overset{\text{def}}{=} -f^*(\boldsymbol{u}) - \sigma_{\mathrm{conv}(\mathcal{A})}(-\boldsymbol{u}) \ , \tag{13}$$

where f^* denotes the convex conjugate of f and $\sigma_{\mathrm{conv}(\mathcal{A})}(\boldsymbol{x}) \overset{\text{def}}{=} \sup\{\langle\boldsymbol{x}, \boldsymbol{a}\rangle : \boldsymbol{a} \in \mathrm{conv}(\mathcal{A})\}$ is the support function over $\mathrm{conv}(\mathcal{A})$, which is the convex conjugate of the indicator function. Note that ψ is concave and that when f convex, we have by duality $\min_{\boldsymbol{x}\in\mathrm{conv}(\mathcal{A})} f(\boldsymbol{x}_t) = \max_{\boldsymbol{u}\in\mathbb{R}^p}\psi(\boldsymbol{u})$.

Theorem 3 (FW variants)**.** *Let f be convex, $\boldsymbol{x}_t$ denote the iterate generated by any of the proposed FW variants (AdaFW, AdaAFW, AdaPFW) after t iterations, with $N_t \geq 1$, and let $\boldsymbol{u}_t$ be defined recursively as $\boldsymbol{u}_0 = \nabla f(\boldsymbol{x}_0)$, $\boldsymbol{u}_{t+1} = (1-\xi_t)\boldsymbol{u}_t + \xi_t\nabla f(\boldsymbol{x}_t)$, where $\xi_t = 2/(\delta N_t + 2)$ if t is a good step and $\xi_t = 0$ otherwise. Then we have:*

$$\begin{aligned} h_t &\leq f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) & (14)\\ &\leq \frac{2\overline{L}_t\operatorname{diam}(\mathcal{A})^2}{\delta^2 N_t + \delta} + \frac{2(1-\delta)}{\delta^2 N_t^2 + \delta N_t}\big(f(\boldsymbol{x}_0) - \psi(\boldsymbol{u}_0)\big) &\\ &= \mathcal{O}\left(\frac{1}{\delta^2 t}\right). & (15)\end{aligned}$$

3.4 Strongly convex objectives

The next result states the linear convergence of some algorithm variants and uses the notions of pyramidal width (PWidth) and minimal directional width (mDW) that have been developed in (Lacoste-Julien, 2016) and (Locatello et al., 2017) respectively, which we state in Appendix B for completeness. We note that the pyramidal width of a set $\mathcal{A}$ is lower bounded by the minimal width over all subsets of atoms, and thus is strictly greater than zero if the number of atoms is finite. The minimal directional width is a much simpler quantity and always strictly greater than zero by the symmetry of our domain.

Theorem 4 (Linear convergence rate for strongly convex objectives)**.** *Let f be μ–strongly convex. Then for AdaAFW, AdaPFW or AdaMP we have the following linear decrease for each good step t:*

$$h_{t+1} \leq (1 - \delta^2\rho_t)h_t, \tag{16}$$

where

$$\rho_t = \frac{\mu}{4L_t}\left(\frac{\mathrm{PWidth}(\mathcal{A})}{\operatorname{diam}(\mathcal{A})}\right)^2 \textit{ for AdaAFW and AdaPFW,}$$

$$\rho_t = \frac{\mu}{L_t}\left(\frac{\mathrm{mDW}(\mathcal{A})}{\operatorname{radius}(\mathcal{A})}\right)^2 \textit{ for AdaMP.}$$

The previous theorem gives a geometric decrease on good steps. Combining this theorem with the bound for the number of bad steps in (7), and noting that the sufficient decrease guarantees that the objective is monotonically decreasing, we obtain a global linear convergence for AdaAFW, AdaPFW and AdaMP.

3.5 Discussion

Non-convex objectives. Lacoste-Julien (2016) studied the convergence of FW assuming the linear subproblems are solved exactly ($\delta = 1$) and obtained a rate of the form (11) with $C_0 = \max\{2h_0, L\operatorname{diam}(\mathrm{conv}(\mathcal{A}))^2\}$ instead. Both rates are similar, although our analysis is more general as it allows to consider the case in which linear subproblems are solved approximately ($\delta < 1$) and also gives rates for the Away-steps and Pairwise variants, for which no rates were previously known.

Theorem 2 also gives the first known convergence rates for a variant of MP on general non-convex functions. Contrary to the case of FW, this bound depends on the mean instead of the maximum of the Lipschitz estimate.

Convex objectives. Compared with (Jaggi, 2013), the primal-dual rates of Theorem 3 are stronger as they hold for the last iterate and not only for the minimum over previous iterates. To the best of our knowledge, primal-dual convergence rates on the last iterate have only been derived in (Nesterov, 2017) and were not extended to approximate linear subproblems nor the Away-steps and Pairwise variants.

Compared to Nesterov (2017) on the special case of exact subproblems ($\delta = 1$), the rates of Theorem 3 are similar but with $\overline{L}_t$ replaced by L. Hence, in the regime $\overline{L}_t \ll L$ (as is often verified in practice), our bounds have a much smaller leading constant.

For MP, Locatello et al. (2018) obtain a similar convergence rate of the form $\mathcal{O}(1/(\delta^2 t))$, but with constants that depend on global properties of ∇f, instead of the adaptive, averaged Lipschitz estimate in our case.

Strongly convex objectives. For the FW variants, the rates are identical to the ones in (Lacoste-Julien and Jaggi, 2015, Theorem 1), with the significant difference of replacing L with the adaptive L_t in the linear rate factor, giving a larger per-iteration decrease whenever $L_t < L$. Our rates are the first also covering approximate subproblems for Away-Steps and Pairwise FW algorithms. It's also worth noticing that both Away-steps FW and Pairwise FW have only been previously analyzed in the presence of exact line-search (Lacoste-Julien and Jaggi, 2015). Additionally, unlike (Lacoste-Julien and Jaggi, 2015), we do not require a smoothness assumption on f outside of the domain. Finally, for the case of MP, we again obtain the same convergence rates as in (Locatello et al., 2017, Theorem 7), but with L replaced by L_t.

4 Benchmarks

We compared the proposed methods across three problems and three datasets. The three datasets are summarized in the table of Figure 1, where density denotes the fraction of nonzero coefficients in data matrix and where the last two columns are quantities that arise during the optimization of AdaPFW and shed light into their empirical value. In both cases t is the number of iterates until 10^{-10} suboptimality is achieved.

4.1 ℓ_1-constrained logistic regression

The first problem that we consider is a logistic regression with an ℓ_1 norm constraint on the coefficients of the form:

$$\operatorname*{arg\,min}_{\|\boldsymbol{x}\|_1 \leq \beta} \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{a}_i^\top \boldsymbol{x}, \boldsymbol{b}_i) + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2 \ , \qquad (17)$$

where φ is the logistic loss. β is chosen to give approximately 1%, 20% of nonzero coefficients respectively. The linear subproblems in this case can be computed exactly ($\delta = 1$) and consist of finding the largest entry of the gradient. The ℓ_2 regularization parameter λ is always set to $\lambda = \frac{1}{n}$.

We applied this problem on two different datasets: Madelon and RCV1. We show the results in Figure 1, subplots A, B, C, D. In this figure we also show the performance of FW, Away-steps FW (AFW) and Pairwise FW (PFW), all of them using the step-size $\gamma_t = \min\left\{g_t L^{-1}\|\boldsymbol{d}_t\|^{-2}, \gamma_t^{\max}\right\}$, as well as the backtracking variants of Dunn (1980) and (Beck et al., 2015), which we denote D-FW and B-FW respectively.

4.2 Nuclear-norm constrained Huber regression

The second problem that we consider is collaborative filtering. We used the MovieLens 1M dataset, which contains 1 million movie ratings, and consider the problem of minimizing a Huber loss, as in (Mehta et al., 2007), between the true known ratings and a matrix $\boldsymbol{X}$. We also constrain the matrix by its nuclear norm $\|\boldsymbol{X}\|_* \leq \beta$, where β is chosen to give approximately 1% and 20% of non-zero singular values respectively. The problem is of the form:

$$\operatorname*{arg\,min}_{\|\boldsymbol{X}\|_* \leq \beta} \frac{1}{n}\sum_{(i,j)\in\mathcal{I}}^{n} L_\xi(\boldsymbol{A}_{i,j} - \boldsymbol{X}_{i,j}) \ , \qquad (18)$$

where H_1 is the Huber loss, defined as

$$L_\xi(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \xi, \\ \xi(|a| - \frac{1}{2}\xi), & \text{otherwise}. \end{cases}$$

The Huber loss is a quadratic for $|a| \leq \xi$ and grows linearly for $|a| > \xi$. The parameter ξ controls this tradeoff and was set to 1 during the experiments.

In this case, the AFW and PFW variants were not considered as they are not directly applicable to this problem as the size of the active set is potentially unbounded. The results of this comparison can be see in subplots E and F of Figure 1. We emphasize that the goal of this experiment is to compare different FW variants and not find the best method for matrix completion. For alternative approaches not based on FW see for instance (Mareček et al., 2017).

We comment on some observed trends from these results:

- **Importance of backtracking.** Across the different datasets, problems and regularization regimes we found that backtracking methods always perform better than their non-backtracking variant.
- **Pairwise FW.** AdaPFW shows a surprisingly good performance when it is applicable, specially in the

Dataset	#samples	#features	density	$\overline{L}_t/L$	$(t - N_t)/t$
Madelon (Guyon et al., 2008)	4400	500	1.	3.3×10^{-3}	5.0×10^{-5}
RCV1 (Lewis et al., 2004)	697641	47236	10^{-3}	1.3×10^{-2}	7.5×10^{-5}
MovieLens 1M (Harper and Konstan, 2015)	6041	3707	0.04	1.1×10^{-2}	–

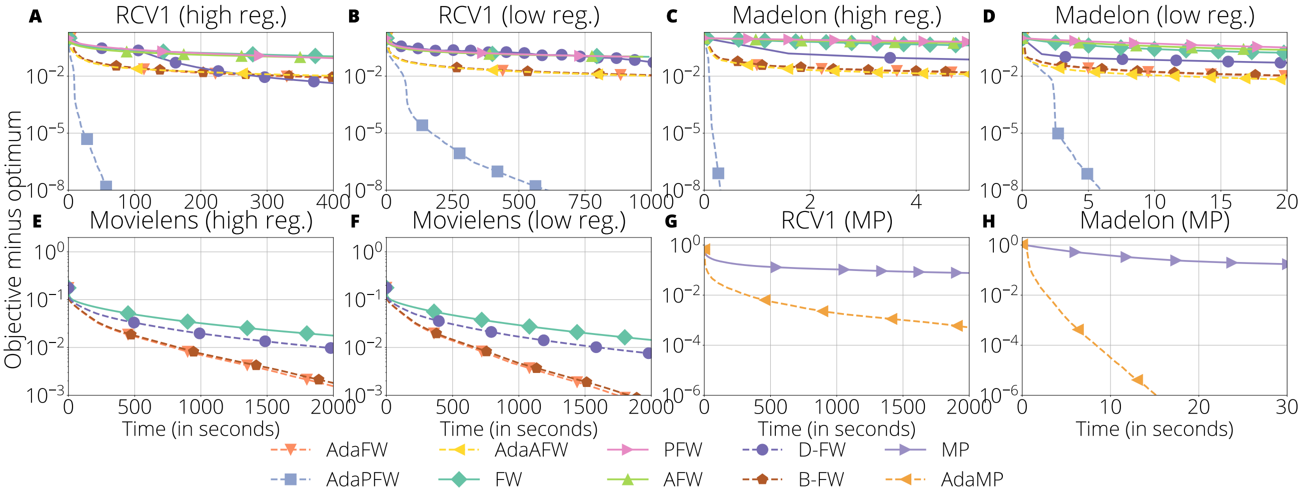


Figure 1: **Top table**: description of the datasets. **Bottom figure**: Benchmark of different FW and MP variants. The variants with backtracking line-search proposed in this paper are in dashed lines. Problem in A, B, C, D = logistic regression with ℓ_1-constrained coefficients, in E, F = Huber regression with on the nuclear norm constrained coefficients and in G, H = unconstrained logistic regression (MP variants). In all the considered datasets and regularization regimes, backtracking variants have a much faster convergence than others.

high regularization regime. A possible interpretation for this is that it is the only variant of FW in which the coefficients associated with previous atoms are not shrunk when adding a new atom, hence large step-sizes are potentially even more beneficial as coefficients that are already close to optimal do not get necessarily modified in subsequent updates.

- **$\overline{L}_t$ vs L.** We compared the average Lipschitz estimate $\overline{L}_t$ and the L, the the gradient's Lipschitz constant. We found that across all datasets the former was more than an order of magnitude smaller, highlighting the need to use a local estimate of the Lipschitz constant to use a large step-size.
- **Bad steps.** Despite the very pessimistic bounds obtained for the number of bad steps in the previous section, we observe that in practice these are extremely rare events, happening less than once every 10,000 iterations.

5 Conclusion and Future Work

In this work we have proposed and analyzed a novel adaptive step-size scheme that can be used in projection-free methods such as FW and MP. The method has minimal computational overhead and does not rely on any step-size hyperparameter (except for an initial estimate). Numerical experiments show large computational gains on a variety of problems.

A possible extension of this work is to develop backtracking step-size strategies for randomized variants of FW such as (Lacoste-Julien et al., 2013; Kerdreux et al., 2018; Mokhtari et al., 2018), in which there is stochasticity in the linear subproblems.

Another area of future research is to improve the convergence rate of the Pairwise FW method. Due to the very pessimistic bound on its number of bad steps, there is still a large gap between its excellent empirical performance and its known convergence rate. Furthermore, convergence of Pairwise and Away-steps for an infinite $\mathcal{A}$, such as the trace norm ball, is still an open problem.

Acknowledgements

The authors would like to thank Vlad Niculae and Courtney Paquette for valuable feedback on the manuscript.

References

Amir Beck, Edouard Pauwels, and Shoham Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM Journal on Optimization*, 2015.

Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

Michael D Canon and Clifton D Cullum. A tight upper bound on the rate of convergence of Frank-Wolfe algorithm. *SIAM Journal on Control*, 1968.

Vladimir Demyanov and Aleksandr Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM Journal on Control*, 1967.

Joseph C Dunn. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM Journal on Control and Optimization*, 1980.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 1956.

Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv preprint arXiv:1301.4666*, 2013.

Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's away step. *Mathematical Programming*, 1986.

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2015.

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

Thomas Kerdreux, Fabian Pedregosa, and Alexandre d'Aspremont. Frank-Wolfe with Subsampling Oracle. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

Simon Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 1966.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 2004.

Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A Unified Optimization View on Generalized Matching Pursuit and Frank-Wolfe. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Francesco Locatello, Anant Raj, Sai Praneeth Karimireddy, Gunnar Raetsch, Bernhard Schölkopf, Sebastian Stich, and Martin Jaggi. On Matching Pursuit and Coordinate Descent. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 1993.

Jakub Mareček, Peter Richtárik, and Martin Takáč. Matrix completion under interval uncertainty. *European Journal of Operational Research*, 2017.

Bhaskar Mehta, Thomas Hofmann, and Wolfgang Nejdl. Robust collaborative filtering. In *Proceedings of the 2007 ACM conference on Recommender systems*, 2007.

BF Mitchell, Vladimir Demyanov, and VN Malozemov. Finding the point of a polyhedron closest to the origin. *SIAM Journal on Control*, 1974.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization. *arXiv*, April 2018.

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 2013.

Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 2017.

John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic Frank-Wolfe methods for nonconvex optimization. In *54th Annual Allerton Conference on Communication, Control, and Computing*, 2016.

Linearly Convergent Frank-Wolfe with Backtracking Line-Search

Supplementary material

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

- Appendix A provides pseudo-code for all FW Variants we consider: AdaFW, AdaAFW, AdaPFW, AdaMP.
- Appendix B contains definitions and properties relative to the objective function and/or the domain, such as the definition of geometric strong convexity and pyramidal width.
- Appendix C we present key inequalities on the abstract algorithm which are used by the different convergence proofs.
- Appendix D provides a proof of convergence for non-convex objectives (Theorem 2).
- Appendix E provides a proof of convergence for convex objectives (Theorem 3).
- Appendix F provides a proof of linear convergence for all variants except FW (Theorem 4).

Appendix A Pseudocode

In this Appendix, we give detailed pseudo-code for the backtracking variants of FW (AdaFW), Away-Steps FW (AdaAFW), Pairwise FW (AdaPFW) and Matching Pursuit (AdaMP).

Appendix A.1 Backtracking FW

Algorithm 3: Backtracking FW (AdaFW)

```
$\boldsymbol{x}_0 \in \mathcal{A}$, initial Lipschitz estimate $L_{-1} > 0$, tolerance $\varepsilon \geq 0$, subproblem quality $\delta \in (0, 1]$, adaptivity params $\tau > 1, \eta \geq 1$
for $t = 0, 1 \ldots$ do
    Choose any $\boldsymbol{s}_t \in \mathcal{A}$ that satisfies $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle$
    Set $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{x}_t$ and $\gamma_{\max} = 1$
    Set $g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle$
    if $g_t \leq \delta\varepsilon$ then return $\boldsymbol{x}_t$;
    $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \gamma_t^{\max})$
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
```

Appendix A.2 Backtracking Away-steps FW

Algorithm 4: Backtracking Away-Steps FW (AdaAFW)

```
$\boldsymbol{x}_0 \in \mathcal{A}$, initial Lipschitz estimate $L_{-1} > 0$, tolerance $\varepsilon \geq 0$, subproblem quality $\delta \in (0, 1]$, adaptivity params
  $\tau > 1, \eta \geq 1$
Let $\mathcal{S}_0 = \{\boldsymbol{x}_0\}$ and $\boldsymbol{\alpha}_{0,\boldsymbol{v}} = 1$ for $\boldsymbol{v} = \boldsymbol{x}_0$ and $\boldsymbol{\alpha}_{0,\boldsymbol{v}} = 0$ otherwise.
for $t = 0, 1 \ldots$ do
    Choose any $\boldsymbol{s}_t \in \mathcal{A}$ that satisfies $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle$
    Choose any $\boldsymbol{v}_t \in \mathcal{S}_t$ that satisfies $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v}_t \rangle \leq \delta \min_{\boldsymbol{v} \in S_t} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v} \rangle$
    if $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v}_t \rangle$ then
        $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{x}_t$ and $\gamma_t^{\max} = 1$
    else
        $\boldsymbol{d}_t = \boldsymbol{x}_t - \boldsymbol{v}_t$, and $\gamma_t^{\max} = \alpha_{\boldsymbol{v}_t,t} / (1 - \alpha_{\boldsymbol{v}_t,t})$
    Set $g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle$
    if $g_t \leq \delta\varepsilon$ then return $\boldsymbol{x}_t$;
    $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \gamma_t^{\max})$
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
    Update active set $\mathcal{S}_{t+1}$ and $\boldsymbol{\alpha}_{t+1}$ (see text)
```

The active set is updated as follows.

- In the case of a FW step, we update the support set $\mathcal{S}_{t+1} = \{\boldsymbol{s}_t\}$ if $\gamma_t = 1$ and otherwise $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{\boldsymbol{s}_t\}$, with coefficients $\alpha_{\boldsymbol{v},t+1} = (1 - \gamma_t)\alpha_{\boldsymbol{v},t}$ for $\boldsymbol{v} \in \mathcal{S}_t \setminus \{\boldsymbol{s}_t\}$ and $\alpha_{\boldsymbol{s}_t,t+1} = (1 - \gamma_t)\alpha_{\boldsymbol{s}_t,t} + \gamma_t$.
- In the case of an Away step: If $\gamma_t = \gamma_{\max}$, then $\mathcal{S}_{t+1} = \mathcal{S}_t \setminus \{\boldsymbol{v}_t\}$, and if $\gamma_t < \gamma_{\max}$, then $\mathcal{S}_{t+1} = \mathcal{S}_t$. Finally, we update the weights as $\alpha_{\boldsymbol{v},t+1} = (1 + \gamma_t)\alpha_{\boldsymbol{v},t}$ for $\boldsymbol{v} \in \mathcal{S}_t \setminus \{\boldsymbol{v}_t\}$ and $\alpha_{\boldsymbol{v}_t,t+1} = (1 + \gamma_t)\alpha_{\boldsymbol{v}_t,t} - \gamma_t$ for the other atoms.

Appendix A.3 Backtracking Pairwise FW

Algorithm 5: Backtracking Pairwise FW (AdaPFW)

```
Input: $\boldsymbol{x}_0 \in \mathcal{A}$, initial Lipschitz estimate $L_{-1} > 0$, tolerance $\varepsilon \geq 0$, subproblem quality $\delta \in (0, 1]$, adaptivity
params $\tau > 1, \eta \geq 1$
Let $\mathcal{S}_0 = \{\boldsymbol{x}_0\}$ and $\boldsymbol{\alpha}_{0,\boldsymbol{v}} = 1$ for $\boldsymbol{v} = \boldsymbol{x}_0$ and $\boldsymbol{\alpha}_{0,\boldsymbol{v}} = 0$ otherwise.
for $t = 0, 1 \ldots$ do
    Choose any $\boldsymbol{s}_t \in \mathcal{A}$ that satisfies $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle$
    Choose any $\boldsymbol{v}_t \in \mathcal{S}_t$ that satisfies $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle$
    $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{v}_t$ and $\gamma_t^{\max} = \alpha_{\boldsymbol{v}_t,t}$
    Set $g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle$
    if $g_t \leq \delta\varepsilon$ then return $\boldsymbol{x}_t$;
    $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \gamma_t^{\max})$
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
    Update active set $\mathcal{S}_{t+1}$ and $\boldsymbol{\alpha}_{t+1}$ (see text)
```

AdaPFW only moves weight from $\boldsymbol{v}_t$ to $\boldsymbol{s}_t$. The active set update becomes $\alpha_{\boldsymbol{s}_t,t+1} = \alpha_{\boldsymbol{s}_t,t} + \gamma_t$, $\alpha_{\boldsymbol{v}_t,t+1} = \alpha_{\boldsymbol{v}_t,t} - \gamma_t$, with $\mathcal{S}_{t+1} = (\mathcal{S}_t \setminus \{\boldsymbol{v}_t\}) \cup \{\boldsymbol{s}_t\}$ if $\alpha_{\boldsymbol{v}_t,t+1} = 0$ and $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{\boldsymbol{s}_t\}$ otherwise.

Appendix A.4 Backtracking Matching Pursuit

Matching Pursuit (Mallat and Zhang, 1993; Locatello et al., 2017) is an algorithm to solve optimization problems of the form

$$\operatorname*{minimize}_{\boldsymbol{x} \in \text{lin}(\mathcal{A})} f(\boldsymbol{x}) \;, \tag{19}$$

where $\mathrm{lin}(\mathcal{A}) \overset{\text{def}}{=} \left\{\sum_{\boldsymbol{v}\in\mathcal{A}} \lambda_{\boldsymbol{v}}\boldsymbol{v} \,\middle|\, \lambda_{\boldsymbol{v}} \in \mathbb{R}\right\}$ is the linear span of the set of *atoms* $\mathcal{A}$. As for the backtracking FW algorithm, we assume that f is L-smooth and $\mathcal{A}$ a potentially infinite but bounded set of elements in $\mathbb{R}^p$.

The MP algorithm relies on solving at each iteration a linear subproblem over the set $\mathcal{B} \overset{\text{def}}{=} \mathcal{A} \cup -\mathcal{A}$, with $-\mathcal{A} = \{-\boldsymbol{a} \,|\, \boldsymbol{a} \in \mathcal{A}\}$. The linear subproblem that needs to be solved at each iteration is the following, where as for previous variants, we allow for an optional quality parameter $\delta \in (0, 1]$:

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t\rangle \leq \delta \min_{\boldsymbol{s}\in\mathcal{B}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}\rangle \ . \tag{20}$$

In Algorithm 6 we detail a novel adaptive variant of the MP algorithm, which we name AdaMP.

Algorithm 6: Backtracking Matching Pursuit (AdaMP)

```
Input: $\boldsymbol{x}_0 \in \mathcal{A}$, initial Lipschitz estimate $L_{-1} > 0$, tolerance $\varepsilon \geq 0$, subproblem quality $\delta \in (0, 1]$
for $t = 0, 1 \ldots$ do
    Choose any $\boldsymbol{s}_t \in \mathcal{A}$ that satisfies (20)
    $\boldsymbol{d}_t = \boldsymbol{s}_t$
    Set $g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t\rangle$
    if $g_t \leq \delta\varepsilon$ then return $\boldsymbol{x}_t$;
    $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \infty)$
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
```

Appendix B Basic definitions and properties

In this section we give basic definitions and properties relative to the objective function and/or the domain, such as the definition of geometric strong convexity and pyramidal width. These definitions are not specific to our algorithms and have appeared in different sources such as Lacoste-Julien and Jaggi (2015); Locatello et al. (2017). We merely gather them here for completeness.

Definition 1 (Geometric strong convexity)**.** *We define the* ***geometric strong convexity constant*** μ_f^A *as*

$$\mu_f^{\mathcal{A}} \overset{def}{=} \inf_{\substack{\boldsymbol{x},\boldsymbol{x}^\star \in \operatorname{conv}(\mathcal{A}) \\ \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle < 0}} \frac{2}{\gamma(\boldsymbol{x},\boldsymbol{x}^\star)^2}\Big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle\Big) \tag{21}$$

$$\textit{where } \gamma(\boldsymbol{x},\boldsymbol{x}^\star) \overset{def}{=} \frac{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle}{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{s}_f(\boldsymbol{x}) - \boldsymbol{v}_f(\boldsymbol{x})\rangle}, \tag{22}$$

where

$$\boldsymbol{s}_f(\boldsymbol{x}) \overset{def}{=} \mathop{\mathbf{arg\,min}}_{\boldsymbol{v}\in\mathcal{A}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle \tag{23}$$

$$\boldsymbol{v}_f(\boldsymbol{x}) \overset{def}{=} \mathop{\mathbf{arg\,min}}_{\substack{\boldsymbol{v}=\boldsymbol{v}_{\mathcal{S}}(\boldsymbol{x}) \\ \mathcal{S}\in\mathcal{S}_{\boldsymbol{x}}}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle \tag{24}$$

$$\boldsymbol{v}_{\mathcal{S}}(\boldsymbol{x}) \overset{def}{=} \mathop{\mathbf{arg\,max}}_{\boldsymbol{v}\in\mathcal{S}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle \tag{25}$$

where $\mathcal{S} \subseteq \mathcal{A}$ *and* $\mathcal{S}_{\boldsymbol{x}} \overset{def}{=} \{\mathcal{S} | \mathcal{S} \subseteq \mathcal{A}$ *such that* $\boldsymbol{x}$ *is a proper convex combination of all the elements in* $\mathcal{S}\}$ *(recall* $\boldsymbol{x}$ *is a proper convex combination of elements in* $\mathcal{S}$ *when* $\boldsymbol{x} = \sum_i \alpha_i \boldsymbol{s}_i$ *where* $\boldsymbol{s}_i \in \mathcal{S}$ *and* $\alpha_i \in (0,1)$*).*

Definition 2 (Pyramidal width)**.** *The* ***pyramidal width*** *of a set* $\mathcal{A}$ *is the smallest pyramidal width of all its faces, i.e.*

$$\operatorname{PWidth}(\mathcal{A}) \overset{def}{=} \min_{\substack{\boldsymbol{x}\in\mathcal{K} \\ \mathcal{K}\in\operatorname{faces}(\operatorname{conv}(\mathcal{A})) \\ \boldsymbol{r}\in\operatorname{cone}(\mathcal{K}-\boldsymbol{x})\setminus\{0\}}} \operatorname{PdirW}(\mathcal{K}\cap\mathcal{A}, \boldsymbol{r}, \boldsymbol{x}) \tag{26}$$

where PdirW *is the pyramidal directional width, defined as*

$$\operatorname{PdirW}(W)(\mathcal{A}, \boldsymbol{r}, \boldsymbol{x}) \overset{def}{=} \min_{\mathcal{S}\in\mathcal{S}_x} \max_{\boldsymbol{s}\in\mathcal{A}, \boldsymbol{v}\in\mathcal{S}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|_2}, \boldsymbol{s}-\boldsymbol{v}\right\rangle \tag{27}$$

We now relate these two geometric quantities together.

Lemma 1 (Lower bounding μ_f^A)**.** *Let* f μ*–strongly convex on* $\operatorname{conv}(\mathcal{A}) = \operatorname{conv}(\mathcal{A})$*. Then*

$$\mu_f^A \geq \mu \cdot (\operatorname{PWidth}(\mathcal{A}))^2 \tag{28}$$

Proof. We refer to (Lacoste-Julien and Jaggi, 2015, Theorem 6). □

Proposition 1. $\operatorname{PWidth}(\mathcal{A}) \leq \operatorname{diam}(\operatorname{conv}(\mathcal{A}))$ *where* $\operatorname{diam}(\mathcal{X}) \overset{def}{=} \sup_{x,y\in\mathcal{X}} \|x-y\|_2$.

Proof. First note that given $\boldsymbol{r}\in\mathcal{R}$, $\boldsymbol{s}\in\mathcal{S}$, $\boldsymbol{v}\in\mathcal{V}$ with $\mathcal{R}$, $\mathcal{S}$, $\mathcal{V} \subseteq \mathbb{R}^n$, we have

$$\langle \boldsymbol{r}/\|\boldsymbol{r}\|_2, \boldsymbol{s}-\boldsymbol{v}\rangle \leq \|\boldsymbol{s}-\boldsymbol{v}\|_2 \qquad \forall \boldsymbol{r}\in\mathcal{R}, \boldsymbol{s}\in\mathcal{S}, \boldsymbol{v}\in\mathcal{V} \tag{29}$$

$$\Rightarrow \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \langle \boldsymbol{r}/\|\boldsymbol{r}\|_2, \boldsymbol{s}-\boldsymbol{v}\rangle \leq \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{s}-\boldsymbol{v}\|_2 \qquad \forall \boldsymbol{r}\in\mathcal{R} \tag{30}$$

$$\Rightarrow \min_{\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \langle \boldsymbol{r}/\|\boldsymbol{r}\|_2, \boldsymbol{s}-\boldsymbol{v}\rangle \leq \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{s}-\boldsymbol{v}\|_2 \tag{31}$$

Applying this result to the definition of pyramidal width we have

$$\begin{aligned}
\mathrm{PWidth}(\mathcal{A}) &= \min_{\substack{\boldsymbol{x}\in\mathcal{K}\\ \mathcal{K}\in\mathrm{faces}(\mathrm{conv}(\mathcal{A}))\\ \boldsymbol{r}\in\mathrm{cone}(\mathcal{K}-\boldsymbol{x})\setminus\{0\}}} \mathrm{PdirW}(\mathcal{K}\cap\mathcal{A},\boldsymbol{r},\boldsymbol{x}) && (32)\\
&= \min_{\substack{\boldsymbol{x}\in\mathcal{K}\\ \mathcal{K}\in\mathrm{faces}(\mathrm{conv}(\mathcal{A}))\\ \boldsymbol{r}\in\mathrm{cone}(\mathcal{K}-\boldsymbol{x})\setminus\{0\}}} \min_{\mathcal{S}\in\mathcal{S}_x} \max_{\boldsymbol{s}\in\mathcal{A},\boldsymbol{v}\in\mathcal{S}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|}, \boldsymbol{s}-\boldsymbol{v}\right\rangle && (33)\\
&= \min_{r\in\mathcal{R}} \max_{s\in\mathcal{A},v\in\mathcal{V}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|}, \boldsymbol{s}-\boldsymbol{v}\right\rangle && (34)\\
& && (35)
\end{aligned}$$

where $\mathcal{R} = \{\mathrm{cone}(\mathcal{K}-\boldsymbol{x})\setminus\{0\} : \text{for some } \boldsymbol{x}\in\mathcal{K},\ \mathcal{K}\in\mathrm{faces}(\mathrm{conv}(\mathcal{A}))\}$ and $\mathcal{V}$ is some subset of $\mathcal{A}$. Applying the derived result we have that

$$\begin{aligned}
\mathrm{PWidth}(\mathcal{A}) &\le \max_{\boldsymbol{s}\in\mathcal{A},\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{s}-\boldsymbol{v}\|_2\\
&\le \max_{\boldsymbol{s},\boldsymbol{v}\in\mathrm{conv}(\mathcal{A})} \|\boldsymbol{s}-\boldsymbol{v}\|_2\\
&= \mathrm{diam}(\mathrm{conv}(\mathcal{A}))
\end{aligned}$$

□

Definition 3. *The* ***minimal directional width*** $\mathrm{mDW}(\mathcal{A})$ *of a set of atoms* $\mathcal{A}$ *is defined as*

$$\mathrm{mDW}(\mathcal{A}) = \min_{\boldsymbol{d}\in\mathrm{lin}(\mathcal{A})} \max_{\boldsymbol{z}\in\mathcal{A}} \frac{\langle \boldsymbol{z},\boldsymbol{d}\rangle}{\|\boldsymbol{d}\|}\ . \tag{36}$$

Note that in contrast to the pyramidal width, the minimal directional width here is a much simpler and robust property of the atom set $\mathcal{A}$, not depending on its combinatorial face structure of the polytope. As can be seen directly from the definition above, the $\mathrm{mDW}(\mathcal{A})$ is robust when adding a duplicate atom or small perturbation of it to $\mathcal{A}$.

Appendix C Preliminaries: Key Inequalities

In this appendix we prove that the sufficient decrease condition verifies a recursive inequality. This key result is used by all convergence proofs.

Lemma 2. *The following inequality is verified for all proposed algorithms (with $\gamma_t^{\max} = +\infty$ for AdaMP):*

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \xi g_t + \frac{\xi^2 L_t}{2}\|\boldsymbol{d}_t\|^2 \quad \textit{for all } \xi \in [0, \gamma_t^{\max}]. \tag{37}$$

Proof. We start the proof by proving an optimality condition of the step-size. Consider the following quadratic optimization problem:

$$\underset{\xi \in [0,\gamma_t^{\max}]}{\text{minimize}} -\xi g_t + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \ . \tag{38}$$

Deriving with respect to ξ and noting that on all the considered algorithms we have $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle \leq 0$, one can easily verify that the global minimizer is achieved at the value

$$\min\left\{\frac{g_t}{L_t\|\boldsymbol{d}_t\|^2}, \gamma_t^{\max}\right\} \ , \tag{39}$$

where $g_t = \langle -\nabla f(\boldsymbol{x}), \boldsymbol{d}_t \rangle$. This coincides with the value of γ_{t+1} computed by the backtracking procedure on the different algorithms and so we have:

$$-\gamma_t g_t + \frac{L_t {\gamma_t}^2}{2}\|\boldsymbol{d}_t\|^2 \leq -\xi g_t + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \quad \text{for all } \xi \in [0, \gamma^{\max}] \ . \tag{40}$$

We can now write the following sequence of inequalities, that combines the sufficient decrease condition with this last inequality:

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \gamma_t g_t + \frac{L_t \gamma_t^2}{2}\|\boldsymbol{d}_t\|^2 \tag{41}$$

$$\overset{(38)}{\leq} f(\boldsymbol{x}_t) - \xi g_t + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \quad \text{for any } \xi \in [0, \gamma^{\max}] \ . \tag{42}$$

□

Proposition 2. *The Lipschitz estimate L_t is bounded as $L_t \leq \max\{\tau L, L_{-1}\}$.*

Proof. If the sufficient decrease condition is verified then we have $L_t = \eta L_{t-1}$ and so $L_t \leq L_{t-1}$. If it's not, we at least have that the Lipschitz estimate cannot larger than τL by definition of Lipschitz constant. Combining both bounds we obtain

$$L_t \leq \max\{\tau L, L_{t-1}\} \ . \tag{43}$$

Applying the same bound recursively on L_{t-1} leads to the claimed bound $L_t \leq \max\{\tau L, L_{-1}\}$. □

Lemma 3. *Let $g(\cdot)$ be as in Theorem 2, i.e., $g(\cdot) = g^{FW}(\cdot)$ for FW variants (AdaFW, AdaAFW, AdaPFW) and $g(\cdot) = g^{MP}(\cdot)$ for MP variants (AdaMP). Then for any of these algorithms we have*

$$g_t \geq \delta g(\boldsymbol{x}_t) \ . \tag{44}$$

Proof. • For AdaFW and AdaMP, Eq. (44) follows immediately from the definition of g_t and $g(\boldsymbol{x}_t)$.

- For AdaAFW, by the way the descent direction is selected in Line 6, we always have

$$g_t \geq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle \geq \delta g(\boldsymbol{x}_t) \ , \tag{45}$$

where the last inequality follows from the definition of $\boldsymbol{s}_t$

- For AdaPFW, we have

$$g_t = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{s}_t \rangle = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{x}_t \rangle \tag{46}$$

$$\geq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle \geq \delta g(\boldsymbol{x}_t) \tag{47}$$

where the term $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{x}_t \rangle$ is positive by definition of $\boldsymbol{v}_t$ since $\boldsymbol{x}_t$ is necessarily in the convex envelope of $\mathcal{S}_t$. The second inequality follows from the definition of $\boldsymbol{s}_t$.

□

Theorem 1. *Let N_t be the total number of evaluations of the sufficient decrease condition up to iteration t. Then we have*

$$n_t \leq \left[1 - \frac{\log \eta}{\log \tau}\right](t+1) + \frac{1}{\log \tau} \max\left\{\log \frac{\tau L}{L_{-1}}, 0\right\} \ . \tag{48}$$

Proof. This proof follows roughly that of (Nesterov, 2013, Lemma 3), albeit with a slightly different bound on L_t due to algorithmic differences.

Denote by $n_i \geq 1$ the number of evaluations of the sufficient decrease condition. Since the algorithm multiplies by τ every time that the sufficient condition is not verified, we have

$$L_i = \eta L_{i-1} \tau^{n_i - 1} \ . \tag{49}$$

Taking logarithms on both sides we obtain

$$n_i \leq 1 - \frac{\log \eta}{\log \tau} + \frac{1}{\tau} \log \frac{L_i}{L_{i-1}} \ . \tag{50}$$

Summing from $i = 0$ to $i = t$ gives

$$n_t \leq \sum_{i=0}^{t} n_i = \left[1 - \frac{\log \eta}{\log \tau}\right](t+1) + \frac{1}{\log \tau} \log\left(\frac{L_t}{L_{-1}}\right) \tag{51}$$

Finally, from Proposition 2 we have the bound $L_t \leq \max\{\tau L, L_{-1}\}$, which we can use to bound the numerator's last term. This gives the claimed bound

$$n_t \leq \sum_{i=0}^{t} n_i = \left[1 - \frac{\log \eta}{\log \tau}\right](t+1) + \frac{1}{\log \tau} \max\left\{\log \frac{\tau L}{L_{-1}}, 0\right\} \ . \tag{52}$$

□

Appendix C.1 A bound on the number of bad steps

To prove the linear rates for the backtracking AFW and backtracking PFW algorithm it is necessary to bound the number of bad steps. There are two different types of bad steps: "drop" steps and "swap" steps. These names come from how the active set $\mathcal{S}_t$ changes. In a drop step, an atom is removed from the active set (i.e. $|\mathcal{S}_{t+1}| < |\mathcal{S}_t|$). In a swap step, the size of the active set remains unchanged (i.e. $|\mathcal{S}_{t+1}| = |\mathcal{S}_t|$) but one atom is swapped with another one not in the active set. Note that drop steps can occur in the (backtracking) Away-steps and Pairwise, but swap steps can only occur in the Pairwise variant.

For the proofs of linear convergence in Appendix F, we show that these two types of bad steps are only problematic when $\gamma_t = \gamma_t^{\max} < 1$. In these scenarios, we cannot provide a meaningful decrease bound. However, we show that the number of bad steps we take is bounded. The following two lemmas adopted from (Lacoste-Julien and Jaggi, 2015, Appendix C) bound the number of drop steps and swap steps the backtracking algorithms can take.

Lemma 4. *After T steps of AdaAFW or AdaPFW, there can only be $T/2$ drop steps. Also, if there is a drop step at step $t+1$, then $f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) < 0$.*

Proof. Let A_t denote the number of steps that added a vertex in the expansion, and let D_t be the number of drop steps. Then $1 \leq |\mathcal{S}_t| = |\mathcal{S}_0| + A_t - D_t$ and we clearly have $A_t - D_t \leq t$. Combining these two inequalities we have that $D_t \leq \frac{1}{2}(|\mathcal{S}_0| - 1 + t) = \frac{t}{2}$.

To show $f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) < 0$, because of Lemma 2, it suffices to show that

$$-\gamma_t g_t + \frac{1}{2}\gamma_t^2 L_t \|\boldsymbol{d}_t\|^2 < 0 \ , \tag{53}$$

with $\gamma_t = \gamma_t^{\max}$ (recall drop steps only occur when $\gamma_t = \gamma_t^{\max}$). Note this is a convex quadratic in γ_t which is precisely less than or equal to 0 when $\gamma_t \in [0, 2g_t/L_t\|\boldsymbol{d}_t\|^2]$. Thus in order to show $f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) < 0$ it suffices to show $\gamma_t^{\max} \in (0, 2g_t/L_t\|\boldsymbol{d}_t\|^2)$. This follows immediately since $0 < \gamma_t^{\max} \leq g_t/L_t\|\boldsymbol{d}_t\|^2$. □

Since in the AdaAFW algorithm all bad steps are drop steps, the previous lemma implies that we can effectively bound the number of bad steps by $t/2$, which is the bound claimed in (7).

Lemma 5. *There are at most $3|\mathcal{A}|!$ bad steps between any two good steps in AdaPFW. Also, if there is a swap step at step $t+1$, then $f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) < 0$.*

Proof. Note that bad steps only occur when $\gamma_t = \gamma_t^{\max} = \alpha_{\boldsymbol{v}_t, t}$. When this happens there are two possibilities; we either move all the mass from $\boldsymbol{v}_t$ to a new atom $\boldsymbol{s}_t \notin \mathcal{S}_t$ (i.e. $\alpha_{\boldsymbol{v}_t, t+1} = 0$ and $\alpha_{\boldsymbol{s}_t, t+1} = \alpha_{\boldsymbol{v}_t, t}$) and preserve the cardinality of our active set ($|\mathcal{S}_{t+1}| = |\mathcal{S}_t|$) or we move all the mass from $\boldsymbol{v}_t$ to an old atom $\boldsymbol{s}_t \in \mathcal{S}_t$ (i.e. $\alpha_{\boldsymbol{s}_t, t+1} = \alpha_{\boldsymbol{s}_t, t} + \alpha_{\boldsymbol{v}_t, t}$) and the cardinality of our active set decreases by 1 ($|\mathcal{S}_{t+1}| < |\mathcal{S}_t|$). In the former case, the possible values of the coordinates $\alpha_{\boldsymbol{v}}$ do not change, but they are simply rearranged in the possible $|\mathcal{A}|$ slots. Note further every time the mass from $\boldsymbol{v}_t$ moves to a new atom $\boldsymbol{s}_t \notin \mathcal{S}_t$ we have strict descent, i.e. $f(\boldsymbol{x}_{t+1}) < f(\boldsymbol{x}_t)$ unless $\boldsymbol{x}_t$ is already optimal (see Lemma 4) and hence we cannot revisit the same point unless we have converged. Thus the maximum number of possible consecutive swap steps is bounded by the number of ways we can assign $|\mathcal{S}_t|$ numbers in $|\mathcal{A}|$ slots, which is $|\mathcal{A}|!/(|\mathcal{A}| - |\mathcal{S}_t|)!$. Furthermore, when the cardinality of our active set drops, in the worst case we will do a maximum number of drop steps before reducing the cardinality of our active set again. Thus starting with $|\mathcal{S}_t| = r$ the maximum number of bad steps B without making any good steps is upper bounded by

$$B \leq \sum_{k=1}^{r} \frac{|\mathcal{A}|!}{(|\mathcal{A}| - k)!} \leq |\mathcal{A}|! \sum_{k=0}^{\infty} \frac{1}{k!} = |\mathcal{A}|! e \leq 3|\mathcal{A}|!$$

□

Appendix D Proofs of convergence for non-convex objectives

In this appendix we provide the convergence proof of Theorem 2. Although this theorem provides a unified convergence proof for both variants of FW and MP, for convenience we split the proof into one for FW variants (Theorem 2.A) and another one for variants of MP (Theorem 2.B)

Theorem 2.A. *Let $\boldsymbol{x}_t$ denote the iterate generated by either AdaFW, AdaAFW or AdaPFW after t iterations. Then for any iteration t with $N_{t+1} \geq 0$, we have the following suboptimality bound in terms of the FW gap:*

$$\lim_{k\to\infty} g^{FW}(\boldsymbol{x}_k) = 0 \quad \text{and} \quad \min_{k=0,\ldots,t} g^{FW}(\boldsymbol{x}_k) \leq \frac{\max\{2h_0, L_t^{\max}\operatorname{diam}(\mathcal{A})^2\}}{\delta\sqrt{N_{t+1}}} = \mathcal{O}\left(\frac{1}{\delta\sqrt{t}}\right) \tag{54}$$

Proof. By Lemma 2 we have the following inequality for any k and any $\xi \in [0, \gamma_k^{\max}]$,

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{\xi^2 C_k}{2}\ , \tag{55}$$

where we define $C_k \overset{\text{def}}{=} L_k\|\boldsymbol{d}_k\|^2$ for convenience. We consider now different cases according to the relative values of γ_k and $\gamma_k^{\max}$, yielding different upper bounds for the right hand side.

Case 1: $\gamma_k < \gamma_k^{\max}$

In this case, γ_k maximizes the right hand side of the (unconstrained) quadratic in inequality (55) which then becomes:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{g_k^2}{2C_k} \leq f(\boldsymbol{x}_k) - \frac{g_k}{2}\min\left\{\frac{g_k}{C_k}, 1\right\} \tag{56}$$

Case 2: $\gamma_k = \gamma_k^{\max} \geq 1$

By the definition of γ_t, this case implies that $C_k \leq g_k$ and so using $\xi = 1$ in (55) gives

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -g_k + \frac{C_k}{2} \leq -\frac{g_k}{2}\ . \tag{57}$$

Case 3: $\gamma_k = \gamma_k^{\max} < 1$

This corresponds to the problematic drop steps for AdaAFW or possibly swap steps for AdaPFW, in which we will only be able to guarantee that the iterates are non-increasing. Choosing $\xi = 0$ in (55) we can at least guarantee that the objective function is non-increasing:

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq 0\ . \tag{58}$$

Combining the previous cases. We can combine the inequalities obtained for the previous cases into the following inequality, valid for all $k \leq t$,

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -\frac{g_k}{2}\min\left\{\frac{g_k}{C_k}, 1\right\}\mathbb{1}\{k \text{ is a good step}\} \tag{59}$$

Adding the previous inequality from $k = 0$ up to t and rearranging we obtain

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}) \geq \sum_{k=0}^{t}\frac{g_k}{2}\min\left\{\frac{g_k}{L_k\|\boldsymbol{d}_k\|^2}, 1\right\}\mathbb{1}\{k \text{ is a good step}\} \tag{60}$$

$$\geq \sum_{k=0}^{t}\frac{g_k}{2}\min\left\{\frac{g_k}{C_k^{\max}}, 1\right\}\mathbb{1}\{k \text{ is a good step}\} \tag{61}$$

with $C_t^{\max} \overset{\text{def}}{=} L_t^{\max} \operatorname{diam}(\operatorname{conv}(\mathcal{A}))^2$. Taking the limit for $t \to +\infty$ we obtain that the left hand side is bounded by the compactness assumption on the domain $\operatorname{conv}(\mathcal{A})$ and L-smoothness on f. The right hand side is an infinite sum, and so a necessary condition for it to be bounded is that $g_k \to 0$, since $g_k \geq 0$ for all k. We have hence proven that $\lim_{k\to\infty} g_k = 0$, which by Lemma 3 implies $\lim_{k\to\infty} g(\boldsymbol{x}_k) = 0$. This proves the first claim of the Theorem.

We will now aim to derive explicit convergence rates for convergence towards a stationary point. Let $\widetilde{g}_t = \min_{0\leq k\leq t} g_k$, then from Eq. (61) we have

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}) \geq \sum_{k=0}^{t} \frac{\widetilde{g}_t}{2} \min\left\{\frac{\widetilde{g}_t}{C_t^{\max}}, 1\right\} \mathbb{1}\{k \text{ is a good step}\} \tag{62}$$

$$= N_{t+1}\frac{\widetilde{g}_t}{2} \min\left\{\frac{\widetilde{g}_t}{C_t^{\max}}, 1\right\} \ . \tag{63}$$

We now make a distinction of cases for the quantities inside the min.

- If $\widetilde{g}_t \leq C_t^{\max}$, then (63) gives $f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}) \geq N_{t+1}\widetilde{g}_t^{\,2}/(2C_t^{\max})$, which reordering gives

$$\widetilde{g}_t \leq \sqrt{\frac{2C_t^{\max}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}))}{N_{t+1}}} \leq \sqrt{\frac{2C_t^{\max} h_0}{N_{t+1}}} \leq \frac{2h_0 + C_t^{\max}}{2\sqrt{N_{t+1}}} \leq \frac{\max\{2h_0, C_t^{\max}\}}{\sqrt{N_{t+1}}} \ . \tag{64}$$

  where in the third inequality we have used the inequality $\sqrt{ab} \leq \frac{a+b}{2}$ with $a = \sqrt{2h_0}$, $b = \sqrt{C_t^{\max}}$.

- If $\widetilde{g}_t > C_t^{\max}$ we can get a better $\frac{1}{N_t}$ rate, trivially bounded by $\frac{1}{\sqrt{N_t}}$.

$$\widetilde{g}_t \leq \frac{2h_0}{N_{t+1}} \leq \frac{2h_0}{\sqrt{N_{t+1}}} \leq \frac{\max\{2h_0, C_t^{\max}\}}{\sqrt{N_{t+1}}} \ . \tag{65}$$

We have obtained the same bound in both cases, hence we always have

$$\widetilde{g}_t \leq \frac{\max\{2h_0, C_t^{\max}\}}{\sqrt{N_{t+1}}} \ . \tag{66}$$

Finally, from Lemma 3 we have $g(\boldsymbol{x}_k) \leq \frac{1}{\delta} g_k$ for all k and so

$$\min_{0\leq k\leq t} g(\boldsymbol{x}_k) \leq \frac{1}{\delta}\min_{0\leq k\leq t} g_k = \frac{1}{\delta}\widetilde{g}_t \leq \frac{\max\{2h_0, C_t^{\max}\}}{\delta\sqrt{N_{t+1}}} \ , \tag{67}$$

and the claimed bound follows by definition of $C_t^{\max}$. The $\mathcal{O}(1/\delta\sqrt{t})$ rate comes from the fact that both $\overline{L}_t$ and h_0 are upper bounded. $\overline{L}_t$ is bounded by Proposition 2 and h_0 is bounded by assumption.

□

Appendix D.1 Matching Pursuit

In the context of Matching Pursuit, we propose the following criterion which we name the MP gap: $g^{\mathrm{MP}}(\boldsymbol{x}) = \max_{\boldsymbol{s}\in\mathcal{B}}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle$, where $\mathcal{B}$ is as defined in Appendix A.4. Note that g^{MP} is always non-negative and $g^{\mathrm{MP}}(\boldsymbol{x}^\star) = 0$ implies $\langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{s}\rangle = 0$ for all $\boldsymbol{s} \in \mathcal{B}$. By linearity of the inner product we then have $\langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star\rangle = 0$ for any $\boldsymbol{x}$ in the domain, since $\boldsymbol{x} - \boldsymbol{x}^\star$ lies in the linear span of $\mathcal{A}$. Hence $\boldsymbol{x}^\star$ is a stationary point and g^{MP} is an appropriate measure of stationarity for this problem.

Theorem 2.B. *Let $\boldsymbol{x}_t$ denote the iterate generated by AdaMP after t iterations. Then for $t \geq 0$ we have the following suboptimality bound in terms of the MP gap:*

$$\lim_{k\to\infty} g^{MP}(\boldsymbol{x}_k) = 0 \qquad \textit{and} \qquad \min_{0\leq k\leq t} g^{MP}(\boldsymbol{x}_k) \leq \frac{\operatorname{radius}(\mathcal{A})}{\delta}\sqrt{\frac{2h_0\overline{L}_t}{t+1}} = \mathcal{O}\left(\frac{1}{\delta\sqrt{t}}\right) . \tag{68}$$

Proof. The proof similar than that of Theorem 2.A, except that in this case the expression of the step-size is simpler and does not depend on the minimum of two quantities. This avoids the case distinction that was necessary in the previous proof, resulting in a much simpler proof.

For all $k = 0, \ldots, t$, using the sufficient decrease condition, and the definitions of γ_k and g_k:

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) &\leq \gamma_k\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \frac{\gamma_k^2 L_k}{2}\|\boldsymbol{d}_k\|^2 && (69)\\ &\leq \min_{\eta\geq 0}\left\{-\eta g_k + \frac{1}{2}\eta^2 L_k\|\boldsymbol{d}_k\|^2\right\} && (70)\\ &\leq -\frac{g_k^2}{2L_k\|\boldsymbol{d}_k\|^2} , && (71)\end{aligned}$$

where the last inequality comes from minimizing with respect to η. Summating over k from 0 to t and negating the previous inequality, we obtain:

$$\sum_{0\leq k\leq t}\frac{g_k^2}{L_k} \leq (f(\boldsymbol{x}_0) - f(\boldsymbol{x}_t))\operatorname{radius}(\mathcal{A})^2 \leq 2h_0\operatorname{radius}(\mathcal{A})^2 . \tag{72}$$

Taking the limit for $t \to \infty$ we obtain that the left hand side has a finite sum since the right hand side is bounded by assumption. Therefore, $g_k \to 0$, which by Lemma 3 implies $\lim_{k\to\infty} g(\boldsymbol{x}_k) = 0$. This proves the first claim of the Theorem.

We now aim to derive explicit convergence rates. Taking the min over the g_ks and taking a square root for the last inequality

$$\min_{0\leq k\leq t} g_k \leq \sqrt{\frac{2h_0\operatorname{radius}(\mathcal{A})^2}{\sum_{0\leq k\leq t} {L_k}^{-1}}} \tag{73}$$

The term $\left(n/\sum_{0\leq k\leq t} {L_k}^{-1}\right)$ is the *harmonic mean* of the L_ks, which is always upper bounded by the average $\overline{L_t}$. Hence we obtain

$$\min_{0\leq k\leq t} g_k \leq \frac{\operatorname{radius}(\mathcal{A})}{\delta}\sqrt{\frac{2h_0\overline{L}_t}{t+1}} . \tag{74}$$

The claimed rate then follows from using the bound $g(\boldsymbol{x}_k) \leq \frac{1}{\delta}g_k$ from Lemma 3, valid for all $k \geq 0$.

The $\mathcal{O}(1/\delta\sqrt{t})$ rate comes from the fact that both $\overline{L}_t$ and h_0 are upper bounded. $\overline{L}_t$ is bounded by Proposition 2 and h_0 is bounded by assumption.

□

Note: Harmonic mean vs arithmetic mean. The convergence rate for MP on non-convex objectives (Theorem 2) also holds by replacing $\overline{L}_t$ by its harmonic mean $H_t \overset{\text{def}}{=} N_t/(\sum_{k=0}^{t-1} L_k^{-1}\mathbb{1}\{k \text{ is a good step}\})$ respectively. The harmonic mean is always less than the arithmetic mean, i.e., $H_t \leq \overline{L}_t$, although for simplicity we only stated both theorems with the arithmetic mean. Note that the Harmonic mean is Schur-concave, implying that $H_t \leq t\min\{L_k : k \leq t\}$, i.e. it is controlled by the smallest Lipschitz estimate encountered so far.

Appendix E Proofs of convergence for convex objectives

In this section we provide a proof the convergence rates stated in the theorem for convex objectives (Theorem 3). The section is structured as follows. We start by proving a technical result which is a slight variation of Lemma 2 and which will be used in the proof of Theorem 3. This is followed by the proof of Theorem 3.

Appendix E.1 Frank-Wolfe variants

Lemma 6. *For any of the proposed FW variants, if t is a good step, then we have*

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \xi g_t + \frac{\xi^2 L_t}{2}\|\boldsymbol{d}_t\|^2 \quad \textit{for all } \xi \in [0,1]. \tag{75}$$

Proof. If $\gamma_t^{\max} \geq 1$, the result is obvious from Lemma 2. If $\gamma_t^{\max} < 1$, then the inequality is only valid in the smaller interval $[0, \gamma_t^{\max}]$. However, since we have assumed that this is a good step, if $\gamma_t^{\max} < 1$ then we must have $\gamma_t < \gamma_t^{\max}$. By Lemma 2, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \min_{\xi \in [0, \gamma_t^{\max}]} \left\{ \xi \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \right\} \tag{76}$$

Because $\gamma_t < \gamma_t^{\max}$ and since the expression inside the minimization term of the previous equation is a quadratic function of ξ, γ_t is the unconstrained minimum and so we have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \min_{\xi \geq 0} \left\{ \xi \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \right\} &(77)\\ &\leq f(\boldsymbol{x}_t) + \min_{\xi \in [0,1]} \left\{ \xi \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \right\} \ . &(78) \end{aligned}$$

The claimed bound then follows from the optimality of the min. □

The following lemma allows to relate the quantity $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle$ with a primal-dual gap and will be essential in the proof of Theorem 3.

Lemma 7. *Let $\boldsymbol{s}_t$ be as defined in any of the FW variants. Then for any iterate $t \geq 0$ we have*

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle \geq \delta(f(\boldsymbol{x}_t) - \psi(\nabla f(\boldsymbol{x}_t))) \ . \tag{79}$$

Proof.

$$\begin{aligned} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle &\overset{(2)}{\geq} \delta \max_{\boldsymbol{s} \in \operatorname{conv}(\mathcal{A})} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s} \rangle &(80)\\ &= \delta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t \rangle + \delta \max_{\boldsymbol{s} \in \operatorname{conv}(\mathcal{A})} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle &(81)\\ &= \delta\big(\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t \rangle + \sigma_{\operatorname{conv}(\mathcal{A})}(-\nabla f(\boldsymbol{x}_t))\big) &(82)\\ &= \delta\big(f(\boldsymbol{x}_t) + \underbrace{f^*(\nabla f(\boldsymbol{x}_t)) + \sigma_{\operatorname{conv}(\mathcal{A})}(-\nabla f(\boldsymbol{x}_t))}_{=-\psi(\nabla f(\boldsymbol{x}_t))}\big) = \delta\big(f(\boldsymbol{x}_t) - \psi(\nabla f(\boldsymbol{x}_t))\big) &(83) \end{aligned}$$

where the first identity uses the definition of $\boldsymbol{s}_t$, the second one the definition of convex conjugate and the last one is a consequence of the Fenchel-Young identity. We recall $\sigma_{\operatorname{conv}(\mathcal{A})}$ is the support function of $\operatorname{conv}(\mathcal{A})$.

□

Theorem 3.A. *Let f be convex, $\boldsymbol{x}_t$ denote the iterate generated by any of the proposed FW variants (AdaFW, AdaAFW, AdaPFW) after t iterations, with $N_t \geq 1$, and let $\boldsymbol{u}_t$ be defined recursively as $\boldsymbol{u}_0 = \nabla f(\boldsymbol{x}_0)$, $\boldsymbol{u}_{t+1} = (1-\xi_t)\boldsymbol{u}_t + \xi_t \nabla f(\boldsymbol{x}_t)$, where $\xi_t = 2/(\delta N_t + 2)$ if t is a good step and $\xi_t = 0$ otherwise. Then we have:*

$$h_t \leq f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2\overline{L}_t \operatorname{diam}(\mathcal{A})^2}{\delta^2 N_t + \delta} + \frac{2(1-\delta)}{\delta^2 N_t^2 + \delta N_t}\big(f(\boldsymbol{x}_0) - \psi(\boldsymbol{u}_0)\big) = \mathcal{O}\left(\frac{1}{\delta^2 t}\right) . \tag{84}$$

Proof. The proof is structured as follows. First, we derive a bound for the case that k is a good step. Second, we derive a bound for the case that k is a bad step. Finally, we add over all iterates to derive the claimed bound.

In both the good and bad step cases, we'll make use of the auxiliary variable σ_k. This is defined recursively as $\sigma_0 = \psi(\nabla f(\boldsymbol{x}_k))$, $\sigma_{k+1} = (1-\delta\xi_k)\sigma_k + \delta\xi_k\psi(\nabla f(\boldsymbol{x}_k))$. Since ψ is concave, Jensen's inequality gives that $\psi(\boldsymbol{u}_k) \geq \sigma_k$ for all k.

Case 1: k is a good step:
By Lemma 6, we have the following sequence of inequalities, valid for all $\xi_t \in [0,1]$:

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \xi_k g_k + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{d}_k\|^2 && (85)\\ &\leq f(\boldsymbol{x}_k) - \xi_k\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{s}_k\rangle + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{d}_k\|^2 && (86)\\ &\leq (1-\delta\xi_k) f(\boldsymbol{x}_k) + \delta\xi_k\psi(\nabla f(\boldsymbol{x}_k)) + \frac{\xi_k^2 L_t}{2}\|\boldsymbol{d}_k\|^2 , && (87) \end{aligned}$$

where the second inequality follows from the definition of g_k (this is an equality for AdaFW but an inequality for the other variants) and the last inequality follows from Lemma 7.

Subtracting $\psi(\boldsymbol{u}_{k+1})$ from both sides of the previous inequality gives

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) - \psi(\boldsymbol{u}_{k+1}) &\leq f(\boldsymbol{x}_{k+1}) - \sigma_{k+1} && (88)\\ &\leq (1-\delta\xi_k)\big[f(\boldsymbol{x}_k) - \sigma_k\big] + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 && (89) \end{aligned}$$

Let $\xi_k = 2/(\delta N_k + 2)$ and $a_k \overset{\text{def}}{=} \frac{1}{2}((N_k - 2)\delta + 2)((N_k - 1)\delta + 2)$. With these definitions, we have the following trivial identities that we will use soon:

$$\begin{aligned} a_{k+1}(1-\delta\xi_k) &= \frac{1}{2}((N_k - 2)\delta + 2)((N_k - 1)\delta + 2) = a_k && (90)\\ a_{k+1}\frac{\xi_k^2}{2} &= \frac{((N_k - 1)\delta + 2)}{(N_k\delta + 2)} \leq 1 && (91) \end{aligned}$$

where in the first inequality we have used that k is a good step and so $N_{k+1} = N_k + 1$.

Multiplying (89) by a_{k+1} we have

$$\begin{aligned} a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - \psi(\boldsymbol{u}_{k+1})\big) &\leq a_{k+1}(1-\delta\xi_k)\big[f(\boldsymbol{x}_k) - \sigma_k\big] + \frac{L_k}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 && (92)\\ &\overset{(90)}{=} a_k\big[f(\boldsymbol{x}_k) - \sigma_k\big] + \frac{L_k}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 && (93)\\ &\leq a_k\big[f(\boldsymbol{x}_k) - \sigma_k\big] + L_k \operatorname{diam}(\mathcal{A})^2 && (94) \end{aligned}$$

Case 2: k is a bad step:
Lemma 2 with $\xi_k = 0$ guarantees that the objective function is non-increasing, i.e., $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$. By construction of σ_k we have $\sigma_{k+1} = \sigma_k$, and so substracting both multiplied by a_{k+1} we obtain

$$\begin{aligned} a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - \psi(\boldsymbol{u}_{k+1})\big) &\leq a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - \sigma_{k+1}\big) && (95)\\ &\leq a_{k+1}\big(f(\boldsymbol{x}_k) - \sigma_k\big) && (96)\\ &= a_k\big(f(\boldsymbol{x}_k) - \sigma_k\big) , && (97) \end{aligned}$$

where in the last identity we have used that its a bad step and so $a_{k+1} = a_k$.

Final: combining cases and adding over iterates:
We can combine (94) and (97) into the following inequality:

$$a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - \psi(\boldsymbol{u}_{k+1})\big) - a_k\big(f(\boldsymbol{x}_k) - \psi(\boldsymbol{u}_k)\big) \leq L_k \operatorname{diam}(\mathcal{A})^2 \mathbb{1}\{k \text{ is a good step}\} \;, \tag{98}$$

where $\mathbb{1}\{\text{condition}\}$ is 1 if condition is verified and 0 otherwise.

Adding this inequality from 0 to $t-1$ gives

$$a_t\big(f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t)\big) \leq \sum_{k=0}^{t-1} L_k Q_A^2 \mathbb{1}\{k \text{ is a good step}\} + a_0(f(\boldsymbol{x}_0) - \sigma_0) \tag{99}$$

$$= N_t \overline{L}_t \operatorname{diam}(\mathcal{A})^2 + (1-\delta)(2-\delta)(f(\boldsymbol{x}_0) - \sigma_0) \tag{100}$$

Finally, dividing both sides by a_t (note that $a_t > 0$ for $N_t \geq 1$) and using $(2-\delta) \leq 2$ we obtain

$$f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2N_t}{((N_t-2)\delta + 2)((N_t-1)\delta+2)} \overline{L}_t Q_{\mathcal{A}}^2 \tag{101}$$

$$+ \frac{4(1-\delta)}{((N_t-2)\delta + 2)((N_t-1)\delta+2)}(f(\boldsymbol{x}_0) - \sigma_0) \tag{102}$$

We will now use the inequalities $(N_t - 2)\delta + 2 \geq N_t\delta$ and $(N_t-1)\delta + 2 \geq N_t\delta + 1$ for the terms in the denominator to obtain

$$f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2\overline{L}_t Q_{\mathcal{A}}^2}{\delta^2 N_t + \delta} + \frac{4(1-\delta)}{\delta_t^2 N_t^2 + \delta N_t}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)) \;. \tag{103}$$

which is the desired bound:

$$f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2\overline{L}_t Q_{\mathcal{A}}^2}{\delta^2 N_t + \delta} + \frac{4(1-\delta)}{\delta_t^2 N_t^2 + \delta N_t}\big[f(\boldsymbol{x}_0) - \psi(\nabla f(\boldsymbol{x}_0))\big]. \tag{104}$$

We will now prove the bound $h_t \leq f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t)$. Let $\boldsymbol{u}^\star$ be an arbitrary maximizer of ψ. Then by duality we have that $f(\boldsymbol{x}^\star) = \psi(\boldsymbol{u}^\star)$ and so

$$f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) = f(\boldsymbol{x}_t) - f*(\boldsymbol{x}^\star) + \psi(\boldsymbol{u}^\star) - \psi(\boldsymbol{u}_t) \geq f(\boldsymbol{x}_t) - f*(\boldsymbol{x}^\star) = h_t \tag{105}$$

Finally, the $\mathcal{O}(\frac{1}{\delta t})$ rate comes from bounding the number of good steps from (7), for which we have $1/N_t \leq \mathcal{O}(1/t)$, and bounding the Lipschitz estimate by a contant (Proposition 2). □

Appendix E.2 Matching Pursuit

Lemma 8. *Let $\boldsymbol{s}_t$ be as defined in AdaMP, $R_{\mathcal{B}}$ be the level set radius defined as*

$$R_{\mathcal{B}} = \max_{\substack{\boldsymbol{x} \in \operatorname{lin}(\mathcal{A}) \\ f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)}} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{B}} \;, \tag{106}$$

and $\boldsymbol{x}^\star$ be any solution to (19). Then we have

$$\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle \geq \frac{\delta}{\max\{\mathcal{R}_{\mathcal{B}}, 1\}}\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big) \tag{107}$$

Proof. By definition of atomic norm we have

$$\frac{\boldsymbol{x}_t - \boldsymbol{x}_t^\star}{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\mathcal{B}}} \in \operatorname{conv}(\mathcal{B}) \tag{108}$$

Since $f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}_0)$, which is a consequence of sufficient decrease condition (Eq. (70)), we have that $R_{\mathcal{B}} \geq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\mathcal{B}}$ and so $\zeta \overset{\text{def}}{=} \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\mathcal{B}}/R_{\mathcal{B}} \leq 1$. By symmetry of $\mathcal{B}$ we have that

$$\frac{\boldsymbol{x}_t - \boldsymbol{x}^\star}{R_{\mathcal{B}}} = \zeta \frac{\boldsymbol{x}_t - \boldsymbol{x}^\star}{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\mathcal{B}}} + (1-\zeta)\mathbf{0} \in \operatorname{conv}(\mathcal{B}) \ . \tag{109}$$

We will now use this fact to bound the original expression. By definition of $\boldsymbol{s}_t$ we have

$$\begin{aligned}
\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t\rangle &\overset{(20)}{\geq} \delta \max_{\boldsymbol{s}\in\mathcal{B}} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}\rangle && (110)\\
&\overset{(109)}{\geq} \frac{\delta}{R_{\mathcal{B}}} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle && (111)\\
&\geq \frac{\delta}{R_{\mathcal{B}}} (f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) && (112)
\end{aligned}$$

where the last inequality follows by convexity.

□

Theorem 3.B. *Let f be convex, $\boldsymbol{x}^\star$ be an arbitrary solution to (19) and let $R_{\mathcal{B}}$ the level set radius:*

$$R_{\mathcal{B}} = \max_{\substack{\boldsymbol{x}\in\operatorname{lin}(\mathcal{A})\\ f(\boldsymbol{x})\leq f(\boldsymbol{x}_0)}} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{B}} \ . \tag{113}$$

If we denote by $\boldsymbol{x}_t$ the iterate generated by AdaMP after $t \geq 1$ iterations and $\beta = \delta/R_{\mathcal{B}}$, then we have:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \frac{2\overline{L}_t \operatorname{radius}(\mathcal{A})^2}{\beta^2 t + \beta} + \frac{2(1-\beta)}{\beta^2 t^2 + \beta t} h_0 = \mathcal{O}\left(\frac{1}{\beta^2 t}\right) \ . \tag{114}$$

Proof. Let $\boldsymbol{x}^\star$ be an arbitrary solution to (19). Then by Lemma 2, we have the following sequence of inequalities, valid for all $\xi_t \geq 0$:

$$\begin{aligned}
f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \xi_k \langle -\nabla f(\boldsymbol{x}_k), \boldsymbol{s}_k\rangle + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{s}_k\|^2 && (115)\\
&\leq f(\boldsymbol{x}_k) - \xi_k \frac{\delta}{R_{\mathcal{B}}}\big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\big] + \frac{\xi_k^2 L_t}{2}\|\boldsymbol{s}_k\|^2 \ , && (116)
\end{aligned}$$

where the second inequality follows from Lemma 8.

Subtracting $f(\boldsymbol{x}^\star)$ from both sides of the previous inequality gives

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq \left(1 - \frac{\delta}{R_{\mathcal{B}}}\xi_k\right)\big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\big] + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{s}_k\|^2 \ . \tag{117}$$

Let $\beta = \delta/R_{\mathcal{B}}$ and $\xi_k = 2/(\beta k + 2)$ and $a_k \overset{\text{def}}{=} \frac{1}{2}((k-2)\beta + 2)((k-1)\beta + 2)$. With these definitions, we have the following trivial results:

$$\begin{aligned}
a_{k+1}(1 - \beta\xi_k) &= \frac{1}{2}((k-2)\beta + 2)((k-1)\beta + 2) = a_k && (118)\\
a_{k+1}\frac{\xi_k^2}{2} &= \frac{((k-1)\beta + 2)}{(k\beta + 2)} \leq 1 \ . && (119)
\end{aligned}$$

Multiplying (117) by a_{k+1} we have

$$\begin{aligned}
a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star)\big) &\leq a_{k+1}(1 - \beta\xi_k)\big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\big] + \frac{L_k}{2}\|\boldsymbol{s}_k\|^2 && (120)\\
&\overset{(90)}{=} a_k\big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\big] + \frac{L_k}{2}\|\boldsymbol{s}_k\|^2 && (121)\\
&\leq \ a_k\big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\big] + L_t \operatorname{radius}(\mathcal{A})^2 && (122)
\end{aligned}$$

Adding this last inequality from 0 to $t-1$ gives

$$a_t\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big) \leq \sum_{k=0}^{t-1} L_k \operatorname{radius}(\mathcal{A})^2 + a_0(f(\boldsymbol{x}_0) - \beta_0) \tag{123}$$

$$= t\overline{L}_t \operatorname{diam}(\mathcal{A})^2 + (1-\delta)(2-\delta)(f(\boldsymbol{x}_0) - \beta_0) \tag{124}$$

Finally, dividing both sides by a_t (note that $a_1 = 2 - \beta \geq 1$ and so a_t is strictly positive for $t \geq 1$), and using $(2-\delta) \leq 2$ we obtain

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \frac{2t}{((t-2)\beta + 2)((t-1)\beta + 2)} \overline{L}_t \operatorname{radius}(\mathcal{A})^2 \tag{125}$$

$$+ \frac{4(1-\beta)}{((t-2)\beta + 2)((t-1)\beta + 2)}(f(\boldsymbol{x}_0) - \beta_0) \tag{126}$$

We will now use the inequalities $(t-2)\beta + 2 \geq t\beta$ and $(t-1)\beta + 2 \geq t\beta + 1$ to simplify the terms in the denominator. With this we obtain to obtain

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \frac{2\overline{L}_t \operatorname{radius}(\mathcal{A})^2}{\beta^2 N_t + \beta} + \frac{4(1-\beta)}{\beta_t^2 N_t^2 + \beta N_t}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)) \;, \tag{127}$$

which is the desired bound. □

Appendix F Proofs of convergence for strongly convex objectives

The following proofs depend on some definitions of geometric constants, which are defined in Appendix B as well as two crucial lemmas from (Lacoste-Julien and Jaggi, 2015, Appendix C).

Appendix F.1 Frank-Wolfe variants

We are now ready to present the convergence rate of the backtracking Frank–Wolfe variants. As we did in Appendix D, although the original proof combines the rates for FW variants and MP, the proof will be split into two, in which we prove separately the linear convergence rates for AdaAFW and AdaPFW (Theorem 4.A) and AdaMP (Theorem 4.B).

Theorem 4.A. *Let f be μ–strongly convex. Then for each good step we have the following geometric decrease:*

$$h_{t+1} \leq (1-\rho_t)h_t, \tag{128}$$

with

$$\rho_t = \frac{\mu\delta^2}{4L_t}\left(\frac{\text{PWidth}(\mathcal{A})}{\text{diam}(\text{conv}(\mathcal{A}))}\right)^2 \qquad \textit{for AdaAFW} \tag{129}$$

$$\rho_t = \min\left\{\frac{\delta}{2}, \delta^2\frac{\mu}{L_t}\left(\frac{\text{PWidth}(\mathcal{A})}{\text{diam}(\text{conv}(\mathcal{A}))}\right)^2\right\} \qquad \textit{for AdaPFW} \tag{130}$$

Note. In the main paper we provided the simplified bound $\rho_t = \dfrac{\mu}{4L_t}\left(\dfrac{\text{PWidth}(\mathcal{A})}{\text{diam}(\mathcal{A})}\right)^2$ for both algorithms AdaAFW and AdaPFW for simplicity. It is easy to see that the bound for AdaPFW above can be trivially bounded by this quantity by noting that $\delta^2 \leq \delta$ and that μ/L_t and $\text{PWidth}(\mathcal{A})/\text{diam}(\text{conv}(\mathcal{A}))$ are necessarily smaller than 1.

Proof. The structure of this proof is similar to that of (Lacoste-Julien and Jaggi, 2015, Theorem 8). We begin by upper bounding the suboptimality h_t. Then we derive a lower bound on $h_{t+1} - h_t$. Combining both we arrive at the desired geometric decrease.

Upper bounding h_t

Assume $\boldsymbol{x}_t$ is not optimal, ie $h_t > 0$. Then we have $\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{x}^\star - \boldsymbol{x}_t\rangle > 0$. Using the definition of the geometric strong convexity bound and letting $\overline{\gamma} \overset{\text{def}}{=} \gamma(\boldsymbol{x}_t, \boldsymbol{x}^\star)$ we have

$$\begin{aligned}
\frac{\overline{\gamma}^2}{2}\mu_f^A &\leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t) + \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{x}^\star - \boldsymbol{x}_t\rangle & (131)\\
&= -h_t + \overline{\gamma}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_f(\boldsymbol{x}_t) - \boldsymbol{v}_f(\boldsymbol{x}_t)\rangle & (132)\\
&\leq -h_t + \overline{\gamma}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t\rangle & (133)\\
&= -h_t + \overline{\gamma} q_t\ , & (134)
\end{aligned}$$

where $q_t \overset{\text{def}}{=} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t\rangle$. For the last inequality we have used the definition of $\boldsymbol{v}_f(\boldsymbol{x})$ which implies $\langle f(\boldsymbol{x}_t), \boldsymbol{v}_f(\boldsymbol{x}_t)\rangle \leq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t\rangle$ and the fact that $\boldsymbol{s}_t = \boldsymbol{s}_f(\boldsymbol{x}_t)$. Therefore

$$h_t \leq -\frac{\overline{\gamma}^2}{2}\mu_f^A + \overline{\gamma} q_t\ , \tag{135}$$

which can always be upper bounded by taking $\overline{\gamma} = \mu^{-1} q_t$ (since this value of $\overline{\gamma}$ maximizes the expression on the right hand side of the previous inequality) to arrive at

$$\begin{aligned}
h_t &\leq \frac{q_t^2}{2\mu_f^A} & (136)\\
&\leq \frac{q_t^2}{2\mu\Delta^2}\ , & (137)
\end{aligned}$$

with $\Delta \overset{\text{def}}{=} \text{PWidth}(\mathcal{A})$ and where the last inequality follows from Lemma 1.

Lower bounding progress $h_t - h_{t+1}$.
Let G be defined as $G = 1/2$ for AdaAFW and $G = 1$ for AdaPFW. We will now prove that for both algorithms we have

$$\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle \geq \delta G q_t \ . \tag{138}$$

For AdaAFW, by the way the direction $\boldsymbol{d}_t$ is chosen on Line 6, we have the following sequence of inequalities:

$$\begin{aligned} 2\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle &\geq \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t^{FW} \rangle + \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t^{A} \rangle \\ &\geq \delta \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle + \delta \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v}_t \rangle \\ &= \delta \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t \rangle \\ &= \delta q_t \ , \end{aligned}$$

For AdaPFW, since $\boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{v}_t$, it follows from the definition of q_t that $\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle \geq \delta q_t$.

We split the rest of the analysis into three cases: $\gamma_t < \gamma_t^{\max}, \gamma_t = \gamma_t^{\max} \geq 1$ and $\gamma_t = \gamma_t^{\max} < 1$. We prove a geometric descent in the first two cases. In the case where $\gamma_t = \gamma_t^{\max} < 1$ (a bad step) we show that the number of bad steps is bounded.

Case 1: $\gamma_t < \gamma_t^{\max}$:
By Lemma 2, we have

$$f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t) \leq f(\boldsymbol{x}_t) + \min_{\eta \in [0, \gamma_t^{\max}]} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} \tag{139}$$

Because $\gamma_t < \gamma_t^{\max}$ and since the expression inside the minimization term (139) is a convex function of η, the minimizer is unique and it coincides with the minimum of the unconstrained problem. Hence we have

$$\min_{\eta \in [0, \gamma_t^{\max}]} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} = \min_{\eta \geq 0} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} \tag{140}$$

Replacing in (2), our bound becomes

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t) &\leq f(\boldsymbol{x}_t) + \min_{\eta \geq 0} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} && (141) \\ &\leq f(\boldsymbol{x}_t) + \min_{\eta \geq 0} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} M^2 \right\} && (142) \\ &\leq f(\boldsymbol{x}_t) + \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} M^2, \ \ \forall \eta \geq 0 && (143) \end{aligned}$$

where the second inequality comes from bounding $\|\boldsymbol{d}_t\|$ by $M \overset{\text{def}}{=} \text{diam}(\text{conv}(\mathcal{A}))$. Subtracting $f(\boldsymbol{x}^\star)$ from both sides and rearranging we have

$$h_t - h_{t+1} \geq \eta \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle - \frac{1}{2} \eta^2 L_t M^2, \ \ \forall \eta \geq 0 \ . \tag{144}$$

Using the gap inequality (138) our lower bound becomes

$$h_t - h_{t+1} \geq \eta \delta G q_t - \frac{1}{2} \eta^2 L_t M^2, \ \ \forall \eta \geq 0 \ . \tag{145}$$

Noting that the lower bound in (145) is a concave function of η, we maximize the bound by selecting $\eta^\star = (L_t M^2)^{-1} \delta G q_t$. Plugging $\eta^\star$ into the bound in (145) and then using the strong convexity bound (137) we have

$$h_t - h_{t+1} \geq \frac{\mu G^2 \Delta^2 \delta^2}{L_t M^2} h_t \Longrightarrow h_{t+1} \leq \Big(1 - \frac{\mu G^2 \Delta^2 \delta^2}{L_t M^2}\Big) h_t \ . \tag{146}$$

Then we have geometric convergence with rate $1-\rho$ where $\rho = (4L_tM^2)^{-1}\mu\Delta^2\delta^2$ for AdaAFW and $\rho = (L_tM^2)^{-1}\mu\Delta^2\delta^2$ for AdaPFW.

Case 2: $\gamma_t = \gamma_t^{\max} \geq 1$

By Lemma 2 and the gap inequality (138), we have

$$h_t - h_{t+1} = f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \geq \eta\delta G q_t - \frac{1}{2}\eta^2 L_t M^2, \;\; \forall \eta \leq \gamma_t^{\max} \; . \tag{147}$$

Since the lower bound (147) is true for all $\eta \leq \gamma_t^{\max}$, we can maximize the bound with $\eta^\star = \min\{(L_tM^2)^{-1}\delta G q_t, \gamma_t^{\max}\}$. In the case when $\eta^\star = (L_tM^2)^{-1}\delta G q_t$ we get the same bound as we do in (146) and hence have linear convergence with rate $1-\rho$ where $\rho = (4L_tM^2)^{-1}\mu\Delta^2\delta^2$ for AdaAFW and $\rho = (L_tM^2)^{-1}\mu\Delta^2\delta^2$ for AdaPFW. If $\eta^\star = \gamma_t^{\max}$ then this implies $L_tM^2 \leq \delta G q_t$. Since $\gamma_t^{\max}$ is assumed to be greater than 1 and the bound holds for all $\eta \leq \gamma_t^{\max}$ we have in particular that it holds for $\eta = 1$ and hence

$$h_t - h_{t+1} \geq \delta G q_t - \frac{1}{2} L_t M^2 \tag{148}$$

$$\geq \delta G q_t - \frac{\delta G q_t}{2} \tag{149}$$

$$\geq \frac{\delta G h_t}{2} \; , \tag{150}$$

where in the second line we use the inequality $L_tM^2 \leq \delta G q_t$ and in the third we use the inequality $h_t \leq q_t$ which is an immediate consequence of convexity of f. Then we have

$$h_{t+1} \leq (1-\rho) h_t \; , \tag{151}$$

where $\rho = \delta/4$ for AdaAFW and $\rho = \delta/2$ for AdaPFW. Note by Proposition 1 and the fact $\mu \leq L_t$ we have $\delta/4 \geq (4L_tM^2)^{-1}\mu\Delta^2\delta^2$.

Case 3: $\gamma_t = \gamma_t^{\max} < 1$ (bad step)

In this case, we have either a drop or swap step and can make no guarantee on the progress of the algorithm (drop and swap are defined in Appendix C). For AdaAFW, $\gamma_t = \gamma_t^{\max} < 1$ is a drop step. From lines 6–9 of AdaAFW we can make the following distinction of cases. In case of a FW step, then $\mathcal{S}_{t+1} = \{\boldsymbol{s}_t\}$ and $\gamma_t = \gamma_t^{\max} = 1$, otherwise $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{\boldsymbol{s}_t\}$. In case of an Away step, $\mathcal{S}_{t+1} = \mathcal{S}_t \backslash \{\boldsymbol{v}_t\}$ if $\gamma_t = \gamma_t^{\max} < 1$, otherwise $\mathcal{S}_{t+1} = \mathcal{S}_t$. Note a drop step can only occur at an Away step. For AdaPFW, $\gamma_t = \gamma_t^{\max} < 1$ will be a drop step when $\boldsymbol{s}_t \in \mathcal{S}_t$ and will be a swap step when $\boldsymbol{s}_t \notin \mathcal{S}_t$.

Even though at these bad steps we do not have the same geometric decrease, Lemma 4 yields that the sequence $\{h_t\}$ is a non-increasing sequence, i.e., $h_{t+1} \leq h_t$. Since we are guaranteed a geometric decrease on steps that are not bad steps, the bounds on the number of bad steps of Eq. (7) is sufficient to conclude that AdaAFW and AdaPFW exhibit a global linear convergence.

□

Appendix F.2 Matching Pursuit

We start by proving the following lemma, which will be crucial in the proof of the backtracking MP's linear convergence rate.

Lemma 9. *Suppose that $\mathcal{A}$ is a non-empty compact set and that f is μ–strongly convex. Let $\nabla_{\mathcal{B}} f(\boldsymbol{x})$ denote the orthogonal projection of $\nabla f(\boldsymbol{x})$ onto $\operatorname{lin}(\mathcal{B})$. Then for all $\boldsymbol{x}^\star - \boldsymbol{x} \in \operatorname{lin}(\mathcal{A})$, we have*

$$f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x}) - \frac{1}{2\mu \operatorname{mDW}(\mathcal{B})^2} \|\nabla_{\mathcal{B}} f(\boldsymbol{x})\|^2_{\mathcal{B}^\star} \; . \tag{152}$$

Proof. From Locatello et al. (2018, Theorem 6), we have that if f is μ-strongly convex, then

$$\mu_{\mathcal{B}} \overset{\text{def}}{=} \inf_{\boldsymbol{x},\boldsymbol{y} \in \operatorname{lin}(\mathcal{B}), \boldsymbol{x} \neq \boldsymbol{y}} \frac{2}{\|\boldsymbol{y} - \boldsymbol{x}\|^2_{\mathcal{B}}} [f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle] \tag{153}$$

is positive and verifies $\mu_{\mathcal{B}} \geq \mathrm{mDW}(\mathcal{B})^2\mu$. Replacing $\boldsymbol{y} = \boldsymbol{x} + \gamma(\boldsymbol{x}^\star - \boldsymbol{x})$ in the definition above we have

$$f(\boldsymbol{x} + \gamma(\boldsymbol{x}^\star - \boldsymbol{x})) \geq f(\boldsymbol{x}) + \gamma\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle + \gamma^2 \frac{\mu_{\mathcal{B}}}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\mathcal{B}} \ . \tag{154}$$

We can fix $\gamma = 1$ on the left hand side and since the expression on the right hand side is true for all γ, we minimize over γ to find $\gamma^* = -\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle / \mu_{\mathcal{B}}\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\mathcal{B}}$. Thus the lower bound becomes

$$\begin{aligned} f(\boldsymbol{x}^\star) &\geq f(\boldsymbol{x}) - \frac{1}{2\mu_{\mathcal{B}}}\frac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\mathcal{B}}} && (155)\\ &\geq f(\boldsymbol{x}) - \frac{1}{2\mu\,\mathrm{mDW}(\mathcal{B})^2}\frac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\mathcal{B}}} && (156)\\ &= f(\boldsymbol{x}) - \frac{1}{2\mu\,\mathrm{mDW}(\mathcal{B})^2}\frac{\langle \nabla_{\mathcal{B}} f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\mathcal{B}}} && (157)\\ &\geq f(\boldsymbol{x}) - \frac{1}{2\mu\,\mathrm{mDW}(\mathcal{B})^2}\|\nabla_{\mathcal{B}} f(\boldsymbol{x})\|^2_{\mathcal{B}^*} \ , && (158) \end{aligned}$$

where the last inequality follows by $|\langle \boldsymbol{y}, \boldsymbol{z}\rangle| \leq \|\boldsymbol{y}\|_{\mathcal{B}^*}\|\boldsymbol{z}\|_{\mathcal{B}}$ □

Theorem 4.B. *(Convergence rate backtracking MP) Let f be μ–strongly convex and suppose $\mathcal{B}$ is a non-empty compact set. Then AdaMP verifies the following geometric decrease for each $t \geq 0$:*

$$h_{t+1} \leq \Big(1 - \delta^2 \rho_t\Big) h_t, \quad \textit{with } \rho_t = \frac{\mu}{L_t}\left(\frac{\mathrm{mDW}(\mathcal{B})}{\mathrm{radius}(\mathcal{B})}\right)^2 \ , \tag{159}$$

where $\mathrm{mDW}(\mathcal{B})$ *the minimal directional width of* $\mathcal{B}$.

Proof. By Lemma 2 and bounding $\|\boldsymbol{d}_t\|$ by $R = \mathrm{radius}(\mathcal{B})$ we have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \min_{\eta \in \mathbb{R}}\left\{\eta\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t\rangle + \frac{\eta^2 L_t R^2}{2}\right\} && (160)\\ &= f(\boldsymbol{x}_t) - \frac{\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t\rangle^2}{2L_t R^2} && (161)\\ &\leq f(\boldsymbol{x}_t) - \delta^2\frac{\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t^\star\rangle^2}{2L_t R^2} && (162) \end{aligned}$$

where $\boldsymbol{s}_t^\star$ is any element such that $\boldsymbol{s}_t^\star \in \mathbf{arg\,min}_{\boldsymbol{s} \in \mathcal{B}}\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}\rangle$ and the inequality follows from the optimality of min and the fact that $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t^\star\rangle \leq 0$. Let $\nabla_{\mathcal{B}} f(\boldsymbol{x}_t)$ denote as in Lemma 9 the orthogonal projection of $\nabla f(\boldsymbol{x}_t)$ onto $\mathrm{lin}(\mathcal{B})$. Then the previous inequality simplifies to

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \delta^2\frac{\langle \nabla_{\mathcal{B}} f(\boldsymbol{x}_t), \boldsymbol{s}_t^\star\rangle^2}{2L_t R^2} \ . \tag{163}$$

By definition of dual norm, we also have $\langle -\nabla_{\mathcal{B}} f(\boldsymbol{x}_t), \boldsymbol{s}_t^\star\rangle = \|\nabla_{\mathcal{B}} f(\boldsymbol{x}_t)\|^2_{\mathcal{B}*}$. Subtracting $f(\boldsymbol{x}^\star)$ from both sides we obtain the upper-bound:

$$h_{t+1} \leq h_t - \delta^2\frac{\|\nabla_{\mathcal{B}} f(\boldsymbol{x}_t)\|^2_{\mathcal{B}*}}{2L_t R^2} \tag{164}$$

To derive the lower-bound, we use Lemma 9 with $\boldsymbol{x} = \boldsymbol{x}_t$ and see that

$$\|\nabla_{\mathcal{B}} f(\boldsymbol{x}_t)\|_{\mathcal{B}^*} \geq 2\mu\,\mathrm{mDW}(\mathcal{B})^2 h_t \tag{165}$$

Combining the upper and lower bound together we have

$$h_{t+1} \leq \Big(1 - \delta^2\frac{\mu\,\mathrm{mDW}(\mathcal{B})^2}{L_t R^2}\Big) h_t \ , \tag{166}$$

which is the claimed bound. □

Appendix G Experiments

In this appendix we give give some details on the experiments which were omitted from the main text, as well as an extended set of results.

Appendix G.1 ℓ_1-regularized logistic regression, Madelon dataset

For the first experiment, we consider an ℓ_1-regularized logistic regression of the form

$$\underset{\|\boldsymbol{x}\|_1 \leq \beta}{\textbf{arg min}} \; \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{a}_i^\top \boldsymbol{x}, \boldsymbol{b}_i) + \frac{\lambda}{2} \|\boldsymbol{x}\|_2^2 \;, \tag{167}$$

where φ is the logistic loss. The linear subproblems in this case can be computed exactly ($\delta = 1$) and consists of finding the largest entry of the gradient. The regularization parameter λ is always set to $\lambda = \frac{1}{n}$.

We first consider the case in which the data $\boldsymbol{a}_i, \boldsymbol{b}_i$ is the Madelon datset. Below are the curves objective suboptimality vs time for the different methods considered. The regularization parmeter, denoted ℓ_1 ball radius in the figure, is chosen as to give 1%, 5% and 20% of non-zero coefficients (the middle figure is absent from the main text).

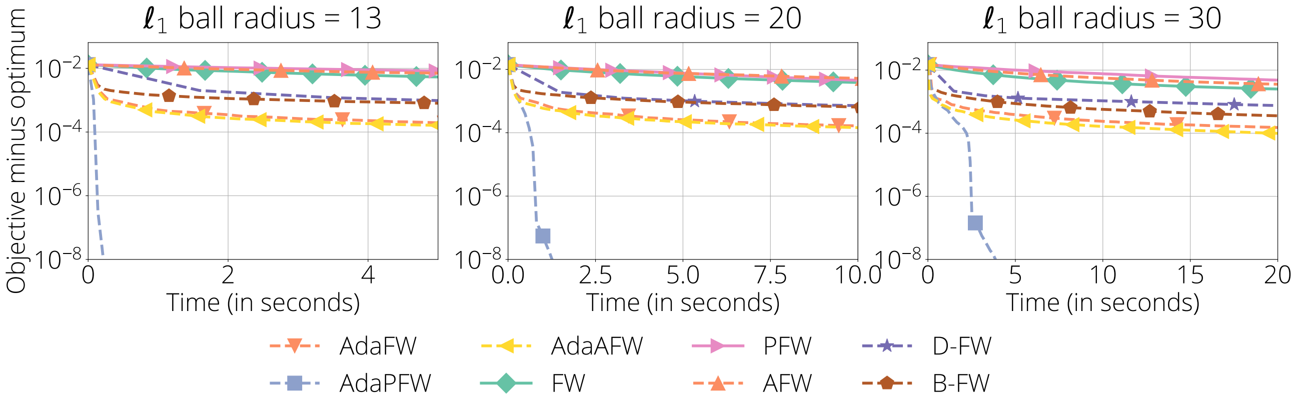


Figure 2: **Comparison of different FW variants**. Problem is ℓ_1-regularized logistic regression and dataset is Madelon in the first, RCV1 in the second figure.

Appendix G.2 ℓ_1-regularized logistic regression, RCV1 dataset

The second experiment is identical to the first one, except the madelon datset is replaced by the larger RCV1 datset. Below we display the results of the comparison in this dataset:

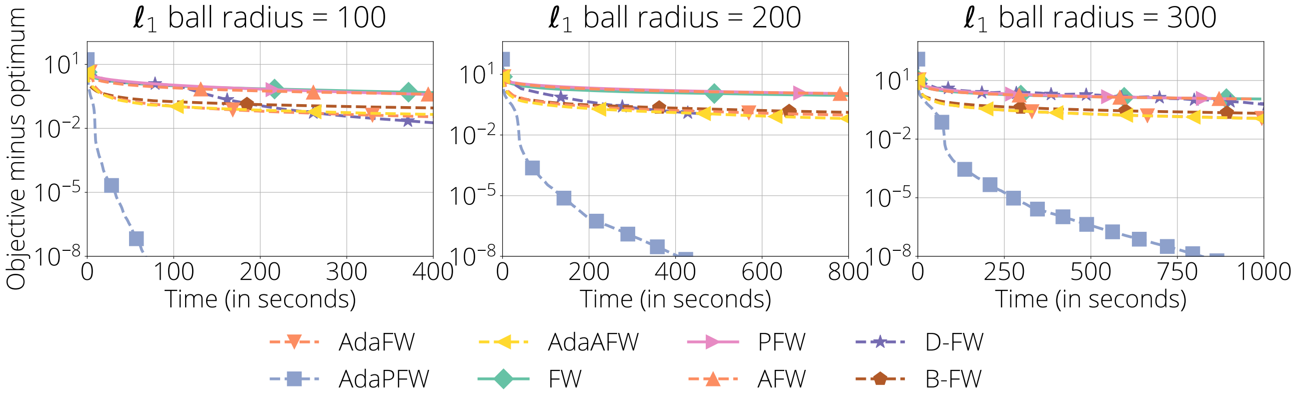


Figure 3: **Comparison of different FW variants**. Problem is ℓ_1-regularized logistic regression and dataset is RCV1.

Appendix G.3 Nuclear norm-regularized Huber regression, MovieLens dataset

For the third experiment, we consider a collaborative filtering problem with the Movielens 1M dataset (Harper and Konstan, 2015) as provided by the spotlight[1] Python package.

In this case the dataset consists of a sparse matrix $\boldsymbol{A}$ representing the ratings for the different movies and users. We denote by $\mathcal{I}$ the non-zero indices of this matrix. Then the optimization probllem that we consider is the following

$$\operatorname*{\mathbf{arg\,min}}_{\|\boldsymbol{X}\|_* \leq \beta} \frac{1}{n} \sum_{(i,j)\in\mathcal{I}}^{n} L_\xi(\boldsymbol{A}_{i,j} - \boldsymbol{X}_{i,j}) \ , \tag{168}$$

where L_ξ is the Huber loss, defined as

$$L_\xi(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \xi, \\ \xi(|a| - \frac{1}{2}\xi), & \text{otherwise}. \end{cases} \tag{169}$$

The Huber loss is a quadratic for $|a| \leq \xi$ and grows linearly for $|a| > \xi$. The parameter ξ controls this tradeoff and was set to 1 during the experiments.

We compared the variant of FW that do not require to store the active set on this problem (as these are the only competitive variants for this problem).

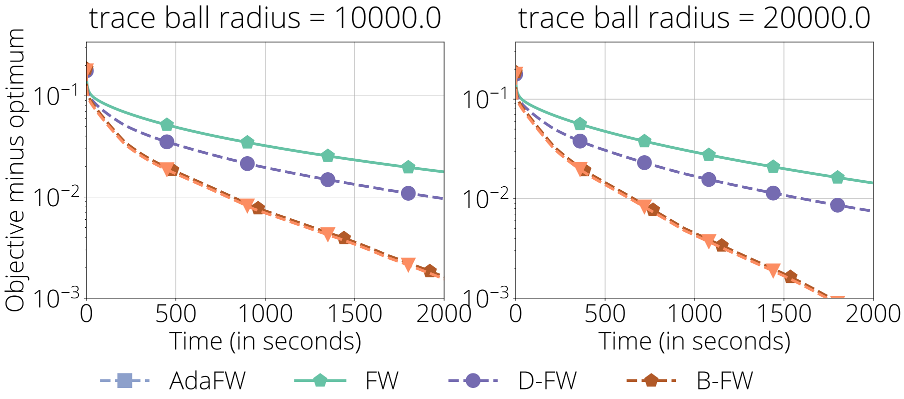


Figure 4: **Comparison of different FW variants**. Comparison of FW variants on the Movielens 1M dataset.

[1] https://github.com/maciejkula/spotlight