OPT2019: 11th Annual Workshop on Optimization for Machine Learning

Linearly Convergent Frank-Wolfe Made Practical

Geoffrey Négiar GEOFFREY_NEGIAR@BERKELEY.EDU
Armin Askari AASKARI@BERKELEY.EDU
Martin Jaggi MARTIN.JAGGI@EPFL.CH
Fabian Pedregosa PEDREGOSA@GOOGLE.COM

Abstract

Structured constraints in Machine Learning have recently brought the Frank-Wolfe (FW) family of algorithms back in the spotlight. Recently, the Away-steps (A) and Pairwise (P) FW variants have been shown to converge linearly for polytopic constraints. However, these improved variants suffer from two practical limitations: each iteration requires solving a 1-dimensional minimization problem to determine the step-size along with an exact solution to the Frank-Wolfe linear subproblems. In this paper, we propose simple modifications of AFW and PFW that lift both restrictions simultaneously. Our method relies on a sufficient decrease condition to determine the step-size. It only requires evaluation and gradient oracles on the objective, along with an approximate solution to the Frank-Wolfe linear subproblems. Furthermore, the theoretical convergence rates of our methods match ones for the exact line-search versions. Benchmarks on different machine learning problems illustrate large practical performance gains of the proposed variants.

1. Introduction

The Frank-Wolfe (FW) or conditional gradient [3, 5] is a method to solve problems of the form

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

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

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 [10]. Although the original FW algorithm only achieves a sublinear convergence rate, other variants like the Away-steps (AFW) and Pairwise (PFW) achieve linear convergence for strongly convex functions over a polytope domain [12]. Unfortunately, both variants rely on an exact line-search, that is, at each iteration, they require the solution of 1-dimensional subproblems of the form $\mathbf{arg\,min}_{\gamma\in[0,\gamma_{\max}]} f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t)$, where $\boldsymbol{d}_t$ is the update direction and $\gamma_{\max}$ is the maximum admissible step-size. This can be a costly optimization problem if the objective is not quadratic, making these methods impractical for more general objectives. It is therefore of great practical interest to have linearly-convergent practical variants of AFW and PFW.

Contributions. Our main contribution is a variant of AFW and PFW for which only we only require access to an oracle for evaluating the objective and its gradient, along with an approximate oracle for linear minimization over the constraint set. In particular we do not know the Lipschitz constant of its gradient. We provide an open source implementation of our adaptive variants, AdaAFW and AdaPFW in the copt library.

Related work	non-convex analysis	approximate subproblems	linear convergence	adaptive step size	bounded backtracking
This work	✓	✓	✓	✓	✓
Lacoste-Julien and Jaggi [12]	✗	✗	✓	✗	N/A
Beck et al. [1]	✗	✓†	✗	✓	✗
Dunn [4]	✓	✗	✗	✓	✗

Table 1: **Comparison with related work.** *non-convex analysis*: convergence guarantees for non-convex objectives. *approximate subproblems*: convergence guarantees when solving linear subproblems approximately. *linear convergence*: guaranteed linear rate of convergence (under hypothesis). *bounded backtracking*: explicit bound for the number of inner loops in adaptive step size methods. †: assumes domain with cartesian product structure.

Related Work. We improve on the Away-Steps [7] and Pairwise [12] variants of FW. In the case of polytope constraints, they were recently shown to converge linearly for strongly convex objectives [6, 12]. These methods require solving the exact line-search at every iteration. In practice, this limits these methods to a small class of objective functions: quadratic objectives. Adaptive step size variants for the classical FW have been described in [4] and [1], but no method to the best of our knowledge derived them for the linearly-convergent Frank-Wolfe variants.

2. Methods

Notation. 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 for $\mu > 0$. $\|\cdot\|$ denotes the euclidean norm.

A general Frank-Wolfe-like algorithm. Our contributions can be applied broadly to variants of FW algorithms, such as AFW, PFW, vanilla FW and Matching Pursuit (MP), therefore we describe it in the context of a general FW-like algorithm, detailed in Alg 1. It relies on two key subroutines: `update_direction` and `step_size`. The first solves a linear subproblem to decide on the direction $\boldsymbol{d}_t$ we then follow to compute the next iterate. The second yields how far along this line we move our iterate: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$ and is detailed in Alg 2.

The `update_direction` routine. This routine varies according to the FW variant. All of them require to solve one or two linear problems, often referred to as linear minimization oracles (LMOs).

These subproblems consist in finding atoms $\boldsymbol{s}_t$ and $\boldsymbol{v}_t$ in the domain such that:

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle \ , \qquad (1)$$

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v}_t \rangle \leq \delta \min_{\boldsymbol{v} \in S_t} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{v} \rangle \ . \qquad (2)$$

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

Algorithm 1: FW variants with adaptive step-size

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

Algorithm 2: Adaptive step-size for FW variants

Note that the second LMO is only useful for the Away-steps and Pairwise variants of FW. The atom $\boldsymbol{v}_t$ belongs to the typically small subset of the atoms called the active set: the set of atoms with non-zero weight $\boldsymbol{\alpha}_{\boldsymbol{s},t} > 0$ in the expansion $\boldsymbol{x}_t = \sum_{\boldsymbol{s}\in\mathcal{S}_t} \boldsymbol{\alpha}_{\boldsymbol{v},t}\boldsymbol{s}$.

FW, AFW, PFW and MP then combine the solution to these linear subproblems in different ways, which we describe in the pseudo code of Appendix A.

The `step_size` routine. To set the step size, we find a local quadratic approximation of our function. We then perform exact line search on this quadratic, which amounts to computing the minimum of a second degree polynomial. The form of our local approximation is

$$Q_t(\gamma, M) = f(\boldsymbol{x}_t) + \gamma\langle\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t\rangle + \frac{\gamma^2 M}{2}\|\boldsymbol{d}_t\|^2 \,,$$

where γ belongs to the interval $[0, \gamma^{\max}]$ to stay in our constraint set. This gives the step size $\gamma_t = \min\left\{g_t/(M\|\boldsymbol{d}_t\|^2), \gamma^{\max}\right\}$. We consider that our approximation is satisfactory when the following condition is verified:

$$f(\boldsymbol{x}_t + \gamma\boldsymbol{d}_t) \leq Q_t(\gamma, M)\,, \gamma = \min\left\{g_t/(M\|\boldsymbol{d}_t\|^2), \gamma^{\max}\right\}. \tag{3}$$

We call this condition the *sufficient decrease condition*. 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 surrogate $Q_t(\cdot, M)$ at its constrained minimum γ_t is an upper bound of $\gamma \mapsto f(\boldsymbol{x}_t + \gamma\boldsymbol{d}_t)$. We emphasize that unlike the "exact line search on quadratic upper bound" approach [3], in this case the surrogate Q_t need not be a global upper bound on the objective. This allows for smaller L_t, and therefore larger step sizes, which empirically induce faster convergence.

3. Theoretical Results

In this section, we provide convergence rates for the proposed methods. We show that they enjoy a $\mathcal{O}(1/\sqrt{t})$ convergence rate for non-convex objectives (Theorem 1), a stronger $\mathcal{O}(1/t)$ convergence rate for convex objectives (Theorem 2), and linear convergence for strongly convex objectives for polytope domains (Theorem 3).

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

- We denote the *objective suboptimality* at step t as $h_t = f(\boldsymbol{x}_t) - \min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$.
- *Good and bad steps*. Following Lacoste-Julien and Jaggi [12], our analysis relies on a notion of "good" and "bad" steps: bad steps verify $\gamma_t = \gamma_t^{\max}$ and $\gamma_t^{\max} < 1$ and good steps don't. For bad steps, we can only guarantee that the objective won't increase, but not lower bound the improvement. However, as in [12], we can lower bound the number of good steps N_t by

$$N_t \geq t/2 \text{ for AdaAFW}\,, \quad \text{and} \quad N_t \geq t/(3|\mathcal{A}|! + 1) \text{ for AdaPFW}. \tag{4}$$

  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).

As a byproduct of our adaptive scheme, our convergence rates make use of the average of the previous Lipschitz estimates over good steps. Let $\mathcal{G}_t$ denote the indices of good steps up to iteration t.

We define the average Lipschitz estimate as $\overline{L}_t \stackrel{\text{def}}{=} \frac{1}{N_t}\sum_{k\in\mathcal{G}_t} L_k$. 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), *which greatly improves convergence in practice*.

Our new convergence rates are presented in the following theorems, which consider the cases of non-convex, convex and strongly convex objectives. Proofs can be found in our full length paper.

Gap function and non-convex objectives. 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 $\boldsymbol{x}^\star \in \mathcal{D}$ such that $\langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star\rangle \geq 0$ for all $\boldsymbol{x} \in \mathcal{D}$ [2]. Following Lacoste-Julien [11], Reddi et al. [18], we express our convergence rate in terms of the FW gap, $g^{\text{FW}}(\boldsymbol{x}) = \max_{\boldsymbol{s}\in\mathcal{D}}\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{s}\rangle$. It is clear that the FW gap is nonnegative and zero only at a stationary point.

Overhead of the adaptive step-size strategy. Evaluation of the sufficient decrease condition requires two extra evaluations of the objective function. If the condition is verified, then it is only evaluated at the current and next iterate. This makes for negligible overhead in this case. On the other hand, we can bound the total number of evaluations of the sufficient decrease condition by $\left[1 + \frac{\log\eta}{\log\tau}\right](t+1) + \frac{1}{\log\tau}\max\left\{\log\frac{\tau L}{L_{-1}}, 0\right\}$. Using this bound, we recommend $\eta = 1.001$, $\tau = 2$. Using these values and for $L_{-1} \geq L/10$ the above bounds imply that for $t \geq 1000$, 99% of the iterations will only perform one evaluation of the sufficient decrease condition.

Theorem 1 (General objectives) *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:*

$$\lim_{t\to\infty} g(\boldsymbol{x}_t) = 0 \qquad \textit{and} \qquad \min_{k=0,\dots,t} g(\boldsymbol{x}_k) \leq \frac{C_t}{\delta\sqrt{N_{t+1}}} = \mathcal{O}\left(\frac{1}{\delta\sqrt{t}}\right), \tag{5}$$

where $C_t = \max\{2h_0, L_t^{\max}\operatorname{diam}(\mathcal{A})^2\}$ and $g = g^{FW}$ is the FW gap for AdaAFW, AdaPFW

Convex Objectives. In the convex case, we need to define the primal-dual gap. We define the *dual* objective function $\psi(\boldsymbol{u}) \stackrel{\text{def}}{=} -f^*(\boldsymbol{u}) - \sigma_{\mathcal{D}}(-\boldsymbol{u})$. f^* denotes the convex conjugate of f and $\sigma_{\mathcal{D}}(\boldsymbol{x}) \stackrel{\text{def}}{=} \sup\{\boldsymbol{x}\cdot\boldsymbol{a} : \boldsymbol{a} \in \mathcal{D}\}$ is the support function over $\mathcal{D}$, which is the convex conjugate of the indicator function. Note that ψ is concave and that when f convex, we have by duality $\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}_t) = \max_{\boldsymbol{u}\in\mathbb{R}^p}\psi(\boldsymbol{u})$.

Theorem 2 (Convex objectives) *Let f be convex, $\boldsymbol{x}_t$ denote the iterate generated by any of the proposed FW variants (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:*

$$h_t \leq f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2\overline{L}_t\operatorname{diam}(\mathcal{A})^2}{\delta^2 N_t + \delta} + \frac{2(1-\delta)}{\delta^2 N_t^2 + \delta N_t}\big(f(\boldsymbol{x}_0) - \psi(\boldsymbol{u}_0)\big) = \mathcal{O}\left(\frac{1}{\delta^2 t}\right). \tag{6}$$

Strongly convex objectives and polytope constraints. For strongly convex objectives and polytope constraints, we use the notions of pyramidal width (PWidth) [11]. We note that the pyramidal width of a set $\mathcal{A}$ is strictly greater than zero if the number of atoms is finite.

Theorem 3 (Linear convergence rate for strongly convex objectives) *Let f be μ–strongly convex. Then for AdaAFW and AdaPFW, we have the following linear decrease for each good step t:*

$$h_{t+1} \leq (1 - \delta^2 \rho_t) h_t, \quad \text{where} \quad \rho_t = \frac{\mu}{4L_t} \left(\frac{\text{PWidth}(\mathcal{A})}{\text{diam}(\mathcal{A})} \right)^2 \text{for AdaAFW and AdaPFW}, \quad (7)$$

The previous theorem gives a geometric decrease on good steps. Combining this theorem with the bound for the number of bad steps in (4), and noting that the sufficient decrease guarantees that the objective is monotonically decreasing, we obtain a global linear convergence for AdaAFW, AdaPFW.

4. Empirical results

We also apply our method to vanilla FW and to Matching Pursuit (MP). Cf Appendix and [15] for details.

Dataset	#samples	#features	density	$\overline{L}_t/L$	$(t - N_t)/t$
Madelon [8]	4400	500	1.	3.3×10^{-3}	5.0×10^{-5}
RCV1 [13]	697641	47236	10^{-3}	1.3×10^{-2}	7.5×10^{-5}
MovieLens 1M [9]	6041	3707	0.04	1.1×10^{-2}	–

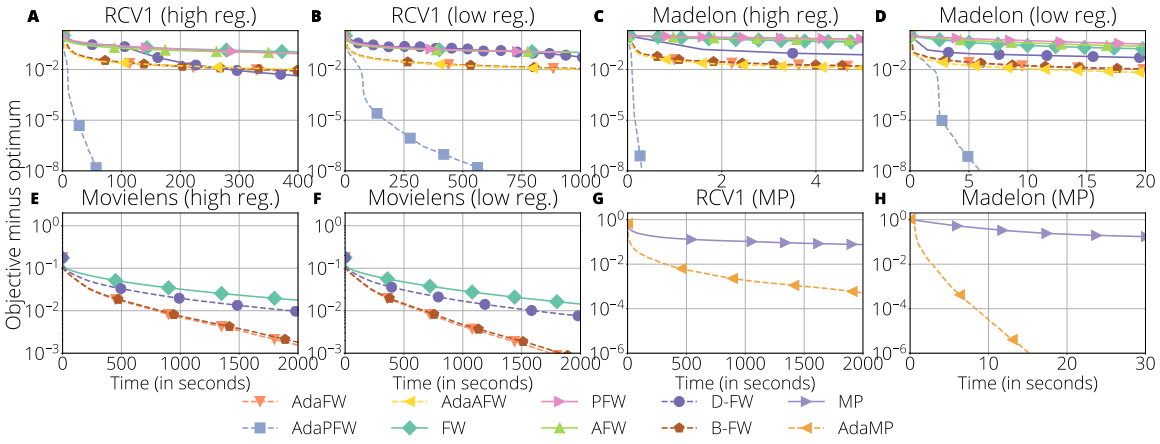

Figure 1: **Top table**: description of the datasets. **Bottom figure**: Benchmark of different FW and MP variants. Adaptive variants proposed in this paper are in dashed lines. Problem in A, B, C, D = logistic regression with ℓ_1-constrained coefficients, in E, F = Huber regression with on the nuclear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

REFERENCES

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

[2] Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

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

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

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

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

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

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

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

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

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

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

[13] 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*, 5(Apr):361–397, 2004.

[14] 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.

[15] 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.

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

[17] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 2013.

[18] 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 without Prior Knowledge

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 1).
- Appendix E provides a proof of convergence for convex objectives (Theorem 2).
- Appendix F provides a proof of linear convergence for all variants except FW (Theorem 3).

Appendix A Pseudocode

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

Appendix A.1 Adaptive FW

$\boldsymbol{x}_0 \in \mathcal{A}$, 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** $\boldsymbol{x}_t$;
  $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \gamma_t^{\max})$
  $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
end

Algorithm 3: Adaptive FW (AdaFW)

Appendix A.2 Adaptive Away-steps FW

$\boldsymbol{x}_0 \in \mathcal{A}$, 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 $\gamma_t^{\max} = \alpha_{\boldsymbol{v}_t,t} / (1 - \alpha_{\boldsymbol{v}_t,t})$
  end
  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)
end

Algorithm 4: Adaptive Away-Steps FW (AdaAFW)

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 Adaptive Pairwise FW

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)
end

Algorithm 5: Adaptive Pairwise FW (AdaPFW)

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 Adaptive Matching Pursuit

Matching Pursuit [14, 16] is an algorithm to solve optimization problems of the form

$$\underset{\boldsymbol{x} \in \text{lin}(\mathcal{A})}{\text{minimize}} \; f(\boldsymbol{x}) \,, \tag{8}$$

where $\text{lin}(\mathcal{A}) \stackrel{\text{def}}{=} \left\{ \sum_{\boldsymbol{v} \in \mathcal{A}} \lambda_{\boldsymbol{v}} \boldsymbol{v} \,\middle|\, \lambda_{\boldsymbol{v}} \in \mathbb{R} \right\}$ is the linear span of the set of *atoms* $\mathcal{A}$. As for the Adaptive 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 $\mathcal{B} \stackrel{\text{def}}{=} \mathcal{A} \cup -\mathcal{A}$, 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 $\delta \in (0, 1]$:

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{B}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle \,. \tag{9}$$

In Algorithm Appendix A.4 we detail a novel adaptive variant of the MP algorithm, which we name 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 (9)
  $\boldsymbol{d}_t = \boldsymbol{s}_t$
  Set $g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle$
  if $g_t \leq \delta\varepsilon$ **then return** $\boldsymbol{x}_t$;
  $\gamma_t, L_t = \text{step\_size}(f, \boldsymbol{d}_t, \boldsymbol{x}_t, g_t, L_{t-1}, \infty)$
  $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
end

Algorithm 6: Adaptive Matching Pursuit (AdaMP)

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 [12], Locatello et al. [14]. We merely gather them here for completeness.

Definition 4 (Geometric strong convexity) *We define the **geometric strong convexity constant** μ_f^A as*

$$\mu_f^{\mathcal{A}} \overset{def}{=} \inf_{\substack{\boldsymbol{x},\boldsymbol{x}^\star \in \operatorname{conv}(\mathcal{A}) \\ \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle < 0}} \frac{2}{\gamma(\boldsymbol{x}, \boldsymbol{x}^\star)^2} \Big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle \Big) \tag{10}$$

$$\text{where } \gamma(\boldsymbol{x}, \boldsymbol{x}^\star) \overset{def}{=} \frac{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle}{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{s}_f(\boldsymbol{x}) - \boldsymbol{v}_f(\boldsymbol{x}) \rangle}, \tag{11}$$

where

$$\boldsymbol{s}_f(\boldsymbol{x}) \overset{def}{=} \operatorname*{\mathbf{arg\,min}}_{\boldsymbol{v} \in \mathcal{A}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle \tag{12}$$

$$\boldsymbol{v}_f(\boldsymbol{x}) \overset{def}{=} \operatorname*{\mathbf{arg\,min}}_{\substack{\boldsymbol{v} = \boldsymbol{v}_{\mathcal{S}}(\boldsymbol{x}) \\ \mathcal{S} \in \mathcal{S}_{\boldsymbol{x}}}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle \tag{13}$$

$$\boldsymbol{v}_{\mathcal{S}}(\boldsymbol{x}) \overset{def}{=} \operatorname*{\mathbf{arg\,max}}_{\boldsymbol{v} \in \mathcal{S}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle \tag{14}$$

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

$$\operatorname{PWidth}(\mathcal{A}) \overset{def}{=} \min_{\substack{\mathcal{K} \in \operatorname{faces}(\operatorname{conv}(\mathcal{A})) \\ \boldsymbol{r} \in \operatorname{cone}(\mathcal{K} - \boldsymbol{x}) \setminus \{0\}}} \operatorname{PdirW}(\mathcal{K} \cap \mathcal{A}, \boldsymbol{r}, \boldsymbol{x}) \tag{15}$$

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

$$\operatorname{PdirW}(W)(\mathcal{A}, \boldsymbol{r}, \boldsymbol{x}) \overset{def}{=} \min_{\mathcal{S} \in \mathcal{S}_x} \max_{\boldsymbol{s} \in \mathcal{A}, \boldsymbol{v} \in \mathcal{S}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|_2}, \boldsymbol{s} - \boldsymbol{v} \right\rangle \tag{16}$$

We now relate these two geometric quantities together.

Lemma 6 (Lower bounding μ_f^A) *Let f μ–strongly convex on $\operatorname{conv}(\mathcal{A}) = \operatorname{conv}(\mathcal{A})$. Then*

$$\mu_f^A \geq \mu \cdot (\operatorname{PWidth}(\mathcal{A}))^2 \tag{17}$$

Proof We refer to [12, Theorem 6]. ■

Proposition 7 $\mathrm{PWidth}(\mathcal{A}) \leq \mathrm{diam}(\mathrm{conv}(\mathcal{A}))$ *where* $\mathrm{diam}(\mathcal{X}) \stackrel{def}{=} \sup_{x,y\in\mathcal{X}} \|x-y\|_2$.

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

$$\langle \boldsymbol{r}/\|\boldsymbol{r}\|_2, \boldsymbol{s}-\boldsymbol{v}\rangle \leq \|\boldsymbol{s}-\boldsymbol{v}\|_2 \qquad \forall \boldsymbol{r}\in\mathcal{R}, \boldsymbol{s}\in\mathcal{S}, \boldsymbol{v}\in\mathcal{V} \tag{18}$$

$$\Rightarrow \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \langle \boldsymbol{r}/\|\boldsymbol{r}\|_2, \boldsymbol{s}-\boldsymbol{v}\rangle \leq \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{s}-\boldsymbol{v}\|_2 \qquad \forall \boldsymbol{r}\in\mathcal{R} \tag{19}$$

$$\Rightarrow \min_{\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \langle \boldsymbol{r}/\|\boldsymbol{r}\|_2, \boldsymbol{s}-\boldsymbol{v}\rangle \leq \max_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{s}-\boldsymbol{v}\|_2 \tag{20}$$

Applying this result to the definition of pyramidal width we have

$$\mathrm{PWidth}(\mathcal{A}) = \min_{\substack{\boldsymbol{x}\in\mathcal{K} \\ \mathcal{K}\in\mathrm{faces}(\mathrm{conv}(\mathcal{A})) \\ \boldsymbol{r}\in\mathrm{cone}(\mathcal{K}-\boldsymbol{x})\backslash\{0\}}} \mathrm{PdirW}(\mathcal{K}\cap\mathcal{A}, \boldsymbol{r}, \boldsymbol{x}) \tag{21}$$

$$= \min_{\substack{\boldsymbol{x}\in\mathcal{K} \\ \mathcal{K}\in\mathrm{faces}(\mathrm{conv}(\mathcal{A})) \\ \boldsymbol{r}\in\mathrm{cone}(\mathcal{K}-\boldsymbol{x})\backslash\{0\}}} \min_{\mathcal{S}\in\mathcal{S}_x} \max_{\boldsymbol{s}\in\mathcal{A},\boldsymbol{v}\in\mathcal{S}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|}, \boldsymbol{s}-\boldsymbol{v} \right\rangle \tag{22}$$

$$= \min_{r\in\mathcal{R}} \max_{s\in\mathcal{A},v\in\mathcal{V}} \left\langle \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|}, \boldsymbol{s}-\boldsymbol{v} \right\rangle \tag{23}$$

$$\tag{24}$$

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

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

■

Definition 8 *The* ***minimal directional width*** $\mathrm{mDW}(\mathcal{A})$ *of a set of atoms* $\mathcal{A}$ *is defined as*

$$\mathrm{mDW}(\mathcal{A}) = \min_{\boldsymbol{d}\in\mathrm{lin}(\mathcal{A})} \max_{\boldsymbol{z}\in\mathcal{A}} \frac{\langle \boldsymbol{z}, \boldsymbol{d}\rangle}{\|\boldsymbol{d}\|}\ . \tag{25}$$

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

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \xi g_t + \frac{\xi^2 L_t}{2}\|\boldsymbol{d}_t\|^2 \text{ for all } \xi \in [0, \gamma_t^{\max}]. \tag{26}$$

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

$$\underset{\xi \in [0,\gamma_t^{\max}]}{\text{minimize}} -\xi g_t + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \,. \tag{27}$$

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

$$\min\left\{ \frac{g_t}{L_t \|\boldsymbol{d}_t\|^2}, \gamma_t^{\max} \right\} \,, \tag{28}$$

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:

$$-\gamma_t g_t + \frac{L_t {\gamma_t}^2}{2}\|\boldsymbol{d}_t\|^2 \leq -\xi g_t + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \text{ for all } \xi \in [0, \gamma^{\max}] \,. \tag{29}$$

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

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) - \gamma_t g_t + \frac{L_t \gamma_t^2}{2}\|\boldsymbol{d}_t\|^2 && (30)\\ &\overset{(27)}{\leq} f(\boldsymbol{x}_t) - \xi g_t + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \text{ for any } \xi \in [0, \gamma^{\max}] \,. && (31) \end{aligned}$$

■

Proposition 10 *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 its not, we at least have that the Lipschitz estimate cannot larger than τL by definition of Lipschitz constant. Combining both bounds we obtain

$$L_t \leq \max\{\tau L, L_{t-1}\} \,. \tag{32}$$

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

Lemma 11 *Let $g(\cdot)$ be as in Theorem 1, 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*

$$g_t \geq \delta g(\boldsymbol{x}_t) \ . \tag{33}$$

Proof

- For AdaFW and AdaMP, Eq. (33) 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 Appendix A.2, we always have

$$g_t \geq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle \geq \delta g(\boldsymbol{x}_t) \ , \tag{34}$$

  where the last inequality follows from the definition of $\boldsymbol{s}_t$
- For AdaPFW, we have

$$\begin{aligned} g_t &= \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{s}_t \rangle = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{x}_t \rangle & (35) \\ &\geq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle \geq \delta g(\boldsymbol{x}_t) & (36) \end{aligned}$$

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

$$n_t \leq \left[1 - \frac{\log \eta}{\log \tau}\right](t+1) + \frac{1}{\log \tau} \max\left\{\log \frac{\tau L}{L_{-1}}, 0\right\} \ . \tag{37}$$

Proof This proof follows roughly that of [17, 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

$$L_i = \eta L_{i-1} \tau^{n_i - 1} \ . \tag{38}$$

Taking logarithms on both sides we obtain

$$n_i \leq 1 - \frac{\log \eta}{\log \tau} + \frac{1}{\tau} \log \frac{L_i}{L_{i-1}} \ . \tag{39}$$

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

$$n_t \leq \sum_{i=0}^{t} n_i = \left[1 - \frac{\log \eta}{\log \tau}\right](t+1) + \frac{1}{\log \tau} \log \left(\frac{L_t}{L_{-1}}\right) \tag{40}$$

Finally, from Proposition 10 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

$$n_t \leq \sum_{i=0}^{t} n_i = \left[1 - \frac{\log \eta}{\log \tau}\right](t+1) + \frac{1}{\log \tau} \max\left\{\log \frac{\tau L}{L_{-1}}, 0\right\} \ . \tag{41}$$

■

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

To prove the linear rates for the adaptive AFW and adaptive 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 (adaptive) 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 [12, Appendix C] bound the number of drop steps and swap steps the adaptive algorithms can take.

Lemma 12 *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 9, it suffices to show that

$$-\gamma_t g_t + \frac{1}{2}\gamma_t^2 L_t \|\boldsymbol{d}_t\|^2 < 0\,, \tag{42}$$

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 (4).

Lemma 13 *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 12) 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 1. 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 1.A) and another one for variants of MP (Theorem 1.B)

Theorem 1.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:*

$$\lim_{k\to\infty} g^{FW}(\boldsymbol{x}_k) = 0 \quad \textit{and} \quad \min_{k=0,\dots,t} g^{FW}(\boldsymbol{x}_k) \leq \frac{\max\{2h_0, L_t^{\max}\operatorname{diam}(\mathcal{A})^2\}}{\delta\sqrt{N_{t+1}}} = \mathcal{O}\left(\frac{1}{\delta\sqrt{t}}\right) \tag{43}$$

Proof By Lemma 9 we have the following inequality for any k and any $\xi \in [0, \gamma_k^{\max}]$,

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{\xi^2 C_k}{2}\,, \tag{44}$$

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 (44) which then becomes:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{g_k^2}{2C_k} \leq f(\boldsymbol{x}_k) - \frac{g_k}{2}\min\left\{\frac{g_k}{C_k}, 1\right\} \tag{45}$$

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 (44) gives

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -g_k + \frac{C_k}{2} \leq -\frac{g_k}{2}\,. \tag{46}$$

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 (44) we can at least guarantee that the objective function is non-increasing:

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) < 0\,. \tag{47}$$

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

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -\frac{g_k}{2}\min\left\{\frac{g_k}{C_k}, 1\right\}\mathbb{1}\{k \text{ is a good step}\} \tag{48}$$

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

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}) \geq \sum_{k=0}^{t} \frac{g_k}{2} \min\left\{\frac{g_k}{L_k\|\boldsymbol{d}_k\|^2}, 1\right\} \mathbb{1}\{k \text{ is a good step}\} \tag{49}$$

$$\geq \sum_{k=0}^{t} \frac{g_k}{2} \min\left\{\frac{g_k}{C_k^{\max}}, 1\right\} \mathbb{1}\{k \text{ is a good step}\} \tag{50}$$

with $C_t^{\max} \stackrel{\text{def}}{=} L_t^{\max} \operatorname{diam}(\operatorname{conv}(\mathcal{A}))^2$. Taking the limit for $t \to +\infty$ we obtain that the right hand side is bounded by the compactness assumption on the domain $\operatorname{conv}(\mathcal{A})$ and L-smoothness on f. The left 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 11 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. (50) we have

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}) \geq \sum_{k=0}^{t} \frac{\widetilde{g}_t}{2} \min\left\{\frac{\widetilde{g}_t}{C_t^{\max}}, 1\right\} \mathbb{1}\{k \text{ is a good step}\} \tag{51}$$

$$= N_{t+1} \frac{\widetilde{g}_t}{2} \min\left\{\frac{\widetilde{g}_t}{C_t^{\max}}, 1\right\} . \tag{52}$$

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

- If $\widetilde{g}_t \leq C_t^{\max}$, then (52) gives $f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}) \geq N_{t+1}\widetilde{g}_t^2/(2C_t^{\max})$, which reordering gives

$$\widetilde{g}_t \leq \sqrt{\frac{2C_t^{\max}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{t+1}))}{N_{t+1}}} \leq \sqrt{\frac{2C_t^{\max}h_0}{N_{t+1}}} \leq \frac{2h_0 + C_t^{\max}}{2\sqrt{N_{t+1}}} \leq \frac{\max\{2h_0, C_t^{\max}\}}{\sqrt{N_{t+1}}} . \tag{53}$$

  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 $\frac{1}{\sqrt{N_t}}$.

$$\widetilde{g}_t \leq \frac{2h_0}{N_{t+1}} \leq \frac{2h_0}{\sqrt{N_{t+1}}} \leq \frac{\max\{2h_0, C_t^{\max}\}}{\sqrt{N_{t+1}}} . \tag{54}$$

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

$$\widetilde{g}_t \leq \frac{\max\{2h_0, C_t^{\max}\}}{\sqrt{N_{t+1}}} . \tag{55}$$

Finally, from Lemma 11 we have $g(\boldsymbol{x}_k) \leq \frac{1}{\delta} g_k$ for all k and so

$$\min_{0\leq k\leq t} g(\boldsymbol{x}_k) \leq \frac{1}{\delta} \min_{0\leq k\leq t} g_k = \frac{1}{\delta}\widetilde{g}_t \leq \frac{\max\{2h_0, C_t^{\max}\}}{\delta\sqrt{N_{t+1}}} , \tag{56}$$

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 10 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: $g^{\text{MP}}(\boldsymbol{x}) = \max_{\boldsymbol{s}\in\mathcal{B}}\langle\nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle$, where $\mathcal{B}$ is as defined in Appendix A.4. Note that g^{MP} is always non-negative and $g^{\text{MP}}(\boldsymbol{x}^\star) = 0$ 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 1.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:*

$$\lim_{k\to\infty} g^{MP}(\boldsymbol{x}_k) = 0 \qquad \textit{and} \qquad \min_{0\leq k\leq t} g^{MP}(\boldsymbol{x}_k) \leq \frac{\operatorname{radius}(\mathcal{A})}{\delta}\sqrt{\frac{2h_0\overline{L}_t}{t+1}} = \mathcal{O}\left(\frac{1}{\delta\sqrt{t}}\right). \tag{57}$$

Proof The proof similar than that of Theorem 1.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:

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq \gamma_k\langle\nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \frac{\gamma_k^2 L_k}{2}\|\boldsymbol{d}_k\|^2 \tag{58}$$

$$\leq \min_{\eta\geq 0}\left\{-\eta g_k + \frac{1}{2}\eta^2 L_k\|\boldsymbol{d}_k\|^2\right\} \tag{59}$$

$$\leq -\frac{g_k^2}{2L_k\|\boldsymbol{d}_k\|^2}\,, \tag{60}$$

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

$$\sum_{0\leq k\leq t}\frac{g_k^2}{L_k} \leq (f(\boldsymbol{x}_0) - f(\boldsymbol{x}_t))\operatorname{radius}(\mathcal{A})^2 \leq 2h_0\operatorname{radius}(\mathcal{A})^2\,. \tag{61}$$

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 11 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

$$\min_{0\leq k\leq t} g_k \leq \sqrt{\frac{2h_0\operatorname{radius}(\mathcal{A})^2}{\sum_{0\leq k\leq t} {L_k}^{-1}}} \tag{62}$$

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

$$\min_{0\leq k\leq t} g_k \leq \frac{\operatorname{radius}(\mathcal{A})}{\delta}\sqrt{\frac{2h_0\overline{L}_t}{t+1}}\,. \tag{63}$$

The claimed rate then follows from using the bound $g(\boldsymbol{x}_k) \leq \frac{1}{\delta} g_k$ from Lemma 11, 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 10 and h_0 is bounded by assumption.

■

Note: Harmonic mean vs arithmetic mean. The convergence rate for MP on non-convex objectives (Theorem 1) also holds by replacing $\overline{L}_t$ by its harmonic mean $H_t \stackrel{\text{def}}{=} N_t/(\sum_{k=0}^{t-1} \overline{L_k}^{-1}\mathbb{1}\{k \text{ is a good step}\})$ 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 2). The section is structured as follows. We start by proving a technical result which is a slight variation of Lemma 9 and which will be used in the proof of Theorem 2. This is followed by the proof of Theorem 2.

Appendix E.1 Frank-Wolfe variants

Lemma 14 *For any of the proposed FW variants, if t is a good step, then we have*

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \xi g_t + \frac{\xi^2 L_t}{2}\|\boldsymbol{d}_t\|^2 \;\; \textit{for all } \xi \in [0,1]. \tag{64}$$

Proof If $\gamma_t^{\max} \geq 1$, the result is obvious from Lemma 9. 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 9, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \min_{\xi \in [0,\gamma_t^{\max}]} \left\{ \xi \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \right\} \tag{65}$$

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

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \min_{\xi \geq 0} \left\{ \xi \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \right\} && (66) \\ &\leq f(\boldsymbol{x}_t) + \min_{\xi \in [0,1]} \left\{ \xi \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \xi^2}{2}\|\boldsymbol{d}_t\|^2 \right\} . && (67) \end{aligned}$$

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 2.

Lemma 15 *Let $\boldsymbol{s}_t$ be as defined in any of the FW variants. Then for any iterate $t \geq 0$ we have*

$$\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle \geq \delta(f(\boldsymbol{x}_t) - \psi(\nabla f(\boldsymbol{x}_t))) . \tag{68}$$

Proof

$$\begin{aligned} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle &\overset{(1)}{\geq} \delta \max_{\boldsymbol{s} \in \operatorname{conv}(\mathcal{A})} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s} \rangle && (69) \\ &= \delta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t \rangle + \delta \max_{\boldsymbol{s} \in \operatorname{conv}(\mathcal{A})} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle && (70) \\ &= \delta\big(\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t \rangle + \sigma_{\operatorname{conv}(\mathcal{A})}(-\nabla f(\boldsymbol{x}_t))\big) && (71) \\ &= \delta\big(f(\boldsymbol{x}_t) + \underbrace{f^*(\nabla f(\boldsymbol{x}_t)) + \sigma_{\operatorname{conv}(\mathcal{A})}(-\nabla f(\boldsymbol{x}_t))}_{=-\psi(\nabla f(\boldsymbol{x}_t))}\big) = \delta\big(f(\boldsymbol{x}_t) - \psi(\nabla f(\boldsymbol{x}_t))\big) && (72) \end{aligned}$$

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 $\sigma_{\mathrm{conv}(\mathcal{A})}$ is the support function of $\mathrm{conv}(\mathcal{A})$.

■

Theorem 2.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:*

$$h_t \leq f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2\overline{L}_t \operatorname{diam}(\mathcal{A})^2}{\delta^2 N_t + \delta} + \frac{2(1-\delta)}{\delta^2 N_t^2 + \delta N_t}\big(f(\boldsymbol{x}_0) - \psi(\boldsymbol{u}_0)\big) = \mathcal{O}\left(\frac{1}{\delta^2 t}\right) . \quad (73)$$

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.

Case 1: k is a good step:

By Lemma 14, we have the following sequence of inequalities, valid for all $\xi_t \in [0,1]$:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi_k g_k + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{d}_k\|^2 \quad (74)$$

$$\leq f(\boldsymbol{x}_k) - \xi_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{s}_k\rangle + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{d}_k\|^2 \quad (75)$$

$$\leq (1-\delta\xi_k) f(\boldsymbol{x}_k) + \delta\xi_k \psi(\nabla f(\boldsymbol{x}_k)) + \frac{\xi_k^2 L_t}{2}\|\boldsymbol{d}_k\|^2 , \quad (76)$$

where the second inequality follows from the definition of g_k (this is an equality for AdaFP but an inequality for the other variants) and the last inequality follows from Lemma 15.

We now introduce the auxiliary variable σ_k. This is defined recursively as $\sigma_0 = \psi(\nabla f(\boldsymbol{x}_k))$, $\sigma_{k+1} = (1-\delta\xi_k)\sigma_k + \delta\xi_k \psi(\nabla f(\boldsymbol{x}_k))$. Subtracting σ_{k+1} from both sides of the previous inequality gives

$$f(\boldsymbol{x}_{k+1}) - \sigma_{k+1} \leq (1-\delta\xi_k)\big[f(\boldsymbol{x}_k) - \sigma_k\big] + \frac{\xi_k^2 L_k}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 \quad (77)$$

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:

$$a_{k+1}(1-\delta\xi_k) = \frac{1}{2}((N_k-2)\delta+2)((N_k-1)\delta+2) = a_k \quad (78)$$

$$a_{k+1}\frac{\xi_k^2}{2} = \frac{((N_k-1)\delta+2)}{(N_k\delta+2)} \leq 1 \quad (79)$$

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

Multiplying (77) by a_{k+1} we have

$$a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - \sigma_{k+1}\big) \leq a_{k+1}(1-\delta\xi_k)\big[f(\boldsymbol{x}_k) - \sigma_k\big] + \frac{L_k}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 \quad (80)$$

$$\overset{(78)}{=} a_k\big[f(\boldsymbol{x}_k) - \sigma_k\big] + \frac{L_k}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 \quad (81)$$

$$\leq a_k\big[f(\boldsymbol{x}_k) - \sigma_k\big] + L_k \operatorname{diam}(\mathcal{A})^2 \quad (82)$$

Case 2: k is a bad step:
Lemma 9 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 adding both multiplied by a_{k+1} we obtain

$$a_{k+1}\big(f(\boldsymbol{x}_{k+1}) - \sigma_{k+1}\big) \leq a_{k+1}\big(f(\boldsymbol{x}_k) - \sigma_k\big) \tag{83}$$
$$= a_k\big(f(\boldsymbol{x}_k) - \sigma_k\big)\,, \tag{84}$$

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 (82) and (84) into the following inequality:

$$a_{k+1}\big(f(\boldsymbol{x}_k) - \sigma_k\big) - a_k\big(f(\boldsymbol{x}_k) - \sigma_k\big) \leq L_k \operatorname{diam}(\mathcal{A})^2 \mathbb{1}\{k \text{ is a good step}\}\,, \tag{85}$$

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

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

$$a_t\big(f(\boldsymbol{x}_t) - \sigma_t\big) \leq \sum_{k=0}^{t-1} L_k Q_A^2 \mathbb{1}\{k \text{ is a good step}\} + a_0(f(\boldsymbol{x}_0) - \sigma_0) \tag{86}$$
$$= N_t \overline{L}_t \operatorname{diam}(\mathcal{A})^2 + (1-\delta)(2-\delta)(f(\boldsymbol{x}_0) - \sigma_0) \tag{87}$$

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

$$f(\boldsymbol{x}_t) - \sigma_t \leq \frac{2N_t}{((N_t-2)\delta+2)((N_t-1)\delta+2)} \overline{L}_t Q_{\mathcal{A}}^2 \tag{88}$$
$$+ \frac{4(1-\delta)}{((N_t-2)\delta+2)((N_t-1)\delta+2)} (f(\boldsymbol{x}_0) - \sigma_0) \tag{89}$$

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

$$f(\boldsymbol{x}_t) - \sigma_t \leq \frac{2\overline{L}_t Q_{\mathcal{A}}^2}{\delta^2 N_t + \delta} + \frac{4(1-\delta)}{\delta_t^2 N_t^2 + \delta N_t} (f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star))\,. \tag{90}$$

In order to prove the claimed bound we just need to prove the bound $-\psi(\boldsymbol{u}_t) \leq -\sigma_t$. We will prove this by induction. For $t = 0$ we have $\psi(\boldsymbol{u}_t) = \sigma_t$ by definition and so the bound is trivially verified. Suppose its true for t, then for $t+1$ we have

$$-\psi(\boldsymbol{u}_{t+1}) = -\psi((1-\xi_t)\boldsymbol{u}_t + \xi_t \nabla f(\boldsymbol{x}_t)) \tag{91}$$
$$\leq -(1-\xi_t)\psi(\boldsymbol{u}_t) - \xi_t \psi(\nabla f(\boldsymbol{x}_t)) \tag{92}$$
$$\leq -(1-\xi_t)\sigma_t - \xi_t \psi(\nabla f(\boldsymbol{x}_t)) \tag{93}$$
$$= -\sigma_{t+1} \tag{94}$$

where the first inequality is true by convexity of $-\psi$ and the second one by the induction hypothesis. Using this bound in (90) yields the desired bound

$$f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) \leq \frac{2\overline{L}_t Q_{\mathcal{A}}^2}{\delta^2 N_t + \delta} + \frac{4(1-\delta)}{\delta_t^2 N_t^2 + \delta N_t} \big[f(\boldsymbol{x}_0) - \psi(\nabla f(\boldsymbol{x}_0))\big] \tag{95}$$

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

$$f(\boldsymbol{x}_t) - \psi(\boldsymbol{u}_t) = f(\boldsymbol{x}_t) - f*(\boldsymbol{x}^\star) + \psi(\boldsymbol{u}^\star) - \psi(\boldsymbol{u}_t) \geq f(\boldsymbol{x}_t) - f*(\boldsymbol{x}^\star) = h_t \tag{96}$$

Finally, the $\mathcal{O}(\frac{1}{\delta t})$ rate comes from bounding the number of good steps from (4), for which we have $1/N_t \leq \mathcal{O}(1/t)$, and bounding the Lipschitz estimate by a contant (Proposition 10). ■

Appendix E.2 Matching Pursuit

Lemma 16 *Let $\boldsymbol{s}_t$ be as defined in AdaMP, $R_{\mathcal{B}}$ be the level set radius defined as*

$$R_{\mathcal{B}} = \max_{\substack{\boldsymbol{x} \in \operatorname{lin}(\mathcal{A}) \\ f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)}} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{B}} \,, \tag{97}$$

and $\boldsymbol{x}^\star$ be any solution to (8). Then we have

$$\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle \geq \frac{\delta}{\max\{\mathcal{R}_{\mathcal{B}}, 1\}} \big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big) \tag{98}$$

Proof By definition of atomic norm we have

$$\frac{\boldsymbol{x}_t - \boldsymbol{x}_t^\star}{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\mathcal{B}}} \in \operatorname{conv}(\mathcal{B}) \tag{99}$$

Since $f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}_0)$, which is a consequence of sufficient decrease condition (Eq. (59)), 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

$$\frac{\boldsymbol{x}_t - \boldsymbol{x}^\star}{R_{\mathcal{B}}} = \zeta \frac{\boldsymbol{x}_t - \boldsymbol{x}^\star}{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\mathcal{B}}} + (1 - \zeta)\mathbf{0} \in \operatorname{conv}(\mathcal{B}) \,. \tag{100}$$

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

$$\begin{aligned}
\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle &\overset{(9)}{\geq} \delta \max_{\boldsymbol{s} \in \mathcal{B}} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle && (101) \\
&\overset{(100)}{\geq} \frac{\delta}{R_{\mathcal{B}}} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle && (102) \\
&\geq \frac{\delta}{R_{\mathcal{B}}} (f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) && (103)
\end{aligned}$$

where the last inequality follows by convexity.

■

Theorem 2.B *Let f be convex, $\boldsymbol{x}^\star$ be an arbitrary solution to (8) and let $R_{\mathcal{B}}$ the level set radius:*

$$R_{\mathcal{B}} = \max_{\substack{\boldsymbol{x} \in \operatorname{lin}(\mathcal{A}) \\ f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)}} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{B}} \,. \tag{104}$$

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:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \frac{2\overline{L}_t \operatorname{radius}(\mathcal{A})^2}{\beta^2 t + \beta} + \frac{2(1-\beta)}{\beta^2 t^2 + \beta t} h_0 = \mathcal{O}\left(\frac{1}{\beta^2 t}\right) . \tag{105}$$

Proof Let $\boldsymbol{x}^\star$ be an arbitrary solution to (8). Then by Lemma 9, we have the following sequence of inequalities, valid for all $\xi_t \geq 0$:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi_k \langle -\nabla f(\boldsymbol{x}_k), \boldsymbol{s}_k \rangle + \frac{\xi_k^2 L_k}{2} \|\boldsymbol{s}_k\|^2 \tag{106}$$

$$\leq f(\boldsymbol{x}_k) - \xi_k \frac{\delta}{R_{\mathcal{B}}} \big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \big] + \frac{\xi_k^2 L_t}{2} \|\boldsymbol{s}_k\|^2 , \tag{107}$$

where the second inequality follows from Lemma 16.

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

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq \left(1 - \frac{\delta}{R_{\mathcal{B}}} \xi_k\right) \big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \big] + \frac{\xi_k^2 L_k}{2} \|\boldsymbol{s}_k\|^2 . \tag{108}$$

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:

$$a_{k+1}(1 - \beta \xi_k) = \frac{1}{2}((k-2)\beta + 2)((k-1)\beta + 2) = a_k \tag{109}$$

$$a_{k+1} \frac{\xi_k^2}{2} = \frac{((k-1)\beta + 2)}{(k\beta + 2)} \leq 1 . \tag{110}$$

Multiplying (108) by a_{k+1} we have

$$a_{k+1} \big(f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \big) \leq a_{k+1}(1 - \beta \xi_k) \big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \big] + \frac{L_k}{2} \|\boldsymbol{s}_k\|^2 \tag{111}$$

$$\overset{(78)}{=} a_k \big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \big] + \frac{L_k}{2} \|\boldsymbol{s}_k\|^2 \tag{112}$$

$$\leq a_k \big[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \big] + L_t \operatorname{radius}(\mathcal{A})^2 \tag{113}$$

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

$$a_t \big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \big) \leq \sum_{k=0}^{t-1} L_k \operatorname{radius}(\mathcal{A})^2 + a_0 (f(\boldsymbol{x}_0) - \beta_0) \tag{114}$$

$$= t \overline{L}_t \operatorname{diam}(\mathcal{A})^2 + (1-\delta)(2-\delta)(f(\boldsymbol{x}_0) - \beta_0) \tag{115}$$

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

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \frac{2t}{((t-2)\beta + 2)((t-1)\beta + 2)} \overline{L}_t \operatorname{radius}(\mathcal{A})^2 \tag{116}$$

$$+ \frac{4(1-\beta)}{((t-2)\beta + 2)((t-1)\beta + 2)} (f(\boldsymbol{x}_0) - \beta_0) \tag{117}$$

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

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq \frac{2\overline{L}_t \operatorname{radius}(\mathcal{A})^2}{\beta^2 N_t + \beta} + \frac{4(1-\beta)}{\beta_t^2 N_t^2 + \beta N_t}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)) \,, \tag{118}$$

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 [12, Appendix C].

Appendix F.1 Frank-Wolfe variants

We are now ready to present the convergence rate of the adaptive 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 3.A) and AdaMP (Theorem 3.B).

Theorem 3.A *Let f be μ–strongly convex. Then for each good step we have the following geometric decrease:*

$$h_{t+1} \leq (1-\rho_t)h_t, \tag{119}$$

with

$$\rho_t = \frac{\mu\delta^2}{4L_t}\left(\frac{\mathrm{PWidth}(\mathcal{A})}{\mathrm{diam}(\mathrm{conv}(\mathcal{A}))}\right)^2 \qquad \textit{for AdaAFW} \tag{120}$$

$$\rho_t = \min\left\{\frac{\delta}{2}, \delta^2\frac{\mu}{L_t}\left(\frac{\mathrm{PWidth}(\mathcal{A})}{\mathrm{diam}(\mathrm{conv}(\mathcal{A}))}\right)^2\right\} \qquad \textit{for AdaPFW} \tag{121}$$

Note. In the main paper we provided the simplified bound $\rho_t = \frac{\mu}{4L_t}\left(\frac{\mathrm{PWidth}(\mathcal{A})}{\mathrm{diam}(\mathcal{A})}\right)^2$ 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 $\mathrm{PWidth}(\mathcal{A})/\mathrm{diam}(\mathrm{conv}(\mathcal{A}))$ are necessarily smaller than 1.

Proof The structure of this proof is similar to that of [12, 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 $\overline{\gamma} \stackrel{\text{def}}{=} \gamma(\boldsymbol{x}_t, \boldsymbol{x}^\star)$ we have

$$\begin{aligned}
\frac{\overline{\gamma}^2}{2}\mu_f^A &\leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t) + \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{x}^\star - \boldsymbol{x}_t\rangle && (122)\\
&= -h_t + \overline{\gamma}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_f(\boldsymbol{x}_t) - \boldsymbol{v}_f(\boldsymbol{x}_t)\rangle && (123)\\
&\leq -h_t + \overline{\gamma}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t\rangle && (124)\\
&= -h_t + \overline{\gamma} q_t\,, && (125)
\end{aligned}$$

where $q_t \stackrel{\text{def}}{=} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t\rangle$. 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

$$h_t \leq -\frac{\overline{\gamma}^2}{2}\mu_f^A + \overline{\gamma} q_t\,, \tag{126}$$

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

$$h_t \leq \frac{q_t^2}{2\mu_f^A} \tag{127}$$

$$\leq \frac{q_t^2}{2\mu\Delta^2} \,, \tag{128}$$

with $\Delta \stackrel{\text{def}}{=} \text{PWidth}(\mathcal{A})$ and where the last inequality follows from Lemma 6.

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

$$\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle \geq \delta G q_t \,. \tag{129}$$

For AdaAFW, by the way the direction $\boldsymbol{d}_t$ is chosen on Line Appendix A.2, 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 9, we have

$$f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t) \leq f(\boldsymbol{x}_t) + \min_{\eta \in [0, \gamma_t^{\max}]} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} \tag{130}$$

Because $\gamma_t < \gamma_t^{\max}$ and since the expression inside the minimization term (130) is a convex function of η, the minimizer is unique and it coincides with the minimum of the unconstrained problem. Hence we have

$$\min_{\eta \in [0, \gamma_t^{\max}]} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} = \min_{\eta \geq 0} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} \tag{131}$$

Replacing in (9), our bound becomes

$$f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t) \leq f(\boldsymbol{x}_t) + \min_{\eta \geq 0} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} \|\boldsymbol{d}_t\|^2 \right\} \tag{132}$$

$$\leq f(\boldsymbol{x}_t) + \min_{\eta \geq 0} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} M^2 \right\} \tag{133}$$

$$\leq f(\boldsymbol{x}_t) + \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L_t \eta^2}{2} M^2, \quad \forall \eta \geq 0 \tag{134}$$

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

$$h_t - h_{t+1} \geq \eta\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t\rangle - \frac{1}{2}\eta^2 L_t M^2, \quad \forall \eta \geq 0\,. \tag{135}$$

Using the gap inequality (129) our lower bound becomes

$$h_t - h_{t+1} \geq \eta\delta G q_t - \frac{1}{2}\eta^2 L_t M^2, \quad \forall \eta \geq 0\,. \tag{136}$$

Noting that the lower bound in (136) 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 (136) and then using the strong convexity bound (128) we have

$$h_t - h_{t+1} \geq \frac{\mu G^2 \Delta^2 \delta^2}{L_t M^2} h_t \implies h_{t+1} \leq \Big(1 - \frac{\mu G^2 \Delta^2 \delta^2}{L_t M^2}\Big) h_t\,. \tag{137}$$

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

Case 2: $\gamma_t = \gamma_t^{\max} \geq 1$
By Lemma 9 and the gap inequality (129), we have

$$h_t - h_{t+1} = f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \geq \eta\delta G q_t - \frac{1}{2}\eta^2 L_t M^2, \quad \forall \eta \leq \gamma_t^{\max}\,. \tag{138}$$

Since the lower bound (138) is true for all $\eta \leq \gamma_t^{\max}$, we can maximize the bound with $\eta^\star = \min\{(L_t M^2)^{-1}\delta G q_t, \gamma_t^{\max}\}$. In the case when $\eta^\star = (L_t M^2)^{-1}\delta G q_t$ we get the same bound as we do in (137) and hence have linear convergence with rate $1-\rho$ where $\rho = (4L_t M^2)^{-1}\mu\Delta^2\delta^2$ for AdaAFW and $\rho = (L_t M^2)^{-1}\mu\Delta^2\delta^2$ for AdaPFW. If $\eta^\star = \gamma_t^{\max}$ then this implies $L_t M^2 \leq \delta G q_t$. 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

$$\begin{aligned} h_t - h_{t+1} &\geq \delta G q_t - \frac{1}{2} L_t M^2 && (139)\\ &\geq \delta G q_t - \frac{\delta G q_t}{2} && (140)\\ &\geq \frac{\delta G h_t}{2}\,, && (141)\end{aligned}$$

where in the second line we use the inequality $L_t M^2 \leq \delta G q_t$ 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

$$h_{t+1} \leq (1-\rho) h_t\,, \tag{142}$$

where $\rho = \delta/4$ for AdaAFW and $\rho = \delta/2$ for AdaPFW. Note by Proposition 7 and the fact $\mu \leq L_t$ we have $\delta/4 \geq (4L_t M^2)^{-1}\mu\Delta^2\delta^2$.

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 Appendix A.2–Appendix A.2 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 12 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. (4) 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 Adaptive MP's linear convergence rate.

Lemma 17 *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*

$$f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x}) - \frac{1}{2\mu \operatorname{mDW}(\mathcal{B})^2} \|\nabla_{\mathcal{B}} f(\boldsymbol{x})\|_{\mathcal{B}^*}^2 . \tag{143}$$

Proof

From Locatello et al. [15, Theorem 6], we have that if f is μ-strongly convex, then

$$\mu_{\mathcal{B}} \stackrel{\text{def}}{=} \inf_{\boldsymbol{x},\boldsymbol{y} \in \operatorname{lin}(\mathcal{B}), \boldsymbol{x} \neq \boldsymbol{y}} \frac{2}{\|\boldsymbol{y} - \boldsymbol{x}\|_{\mathcal{B}}^2} [f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle] \tag{144}$$

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

$$f(\boldsymbol{x} + \gamma(\boldsymbol{x}^\star - \boldsymbol{x})) \geq f(\boldsymbol{x}) + \gamma \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle + \gamma^2 \frac{\mu_{\mathcal{B}}}{2} \|\boldsymbol{x}^\star - \boldsymbol{x}\|_{\mathcal{B}}^2 . \tag{145}$$

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 $\gamma^* = -\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle / \mu_{\mathcal{B}} \|\boldsymbol{x}^\star - \boldsymbol{x}\|_{\mathcal{B}}^2$. Thus the lower bound becomes

$$\begin{aligned}
f(\boldsymbol{x}^\star) &\geq f(\boldsymbol{x}) - \frac{1}{2\mu_{\mathcal{B}}} \frac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|_{\mathcal{B}}^2} && (146)\\
&\geq f(\boldsymbol{x}) - \frac{1}{2\mu \operatorname{mDW}(\mathcal{B})^2} \frac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|_{\mathcal{B}}^2} && (147)\\
&= f(\boldsymbol{x}) - \frac{1}{2\mu \operatorname{mDW}(\mathcal{B})^2} \frac{\langle \nabla_{\mathcal{B}} f(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x} \rangle}{\|\boldsymbol{x}^\star - \boldsymbol{x}\|_{\mathcal{B}}^2} && (148)\\
&\geq f(\boldsymbol{x}) - \frac{1}{2\mu \operatorname{mDW}(\mathcal{B})^2} \|\nabla_{\mathcal{B}} f(\boldsymbol{x})\|_{\mathcal{B}^*}^2 , && (149)
\end{aligned}$$

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

Theorem 3.B *(Convergence rate Adaptive 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 $t \geq 0$:*

$$h_{t+1} \leq \Big(1 - \delta^2 \rho_t\Big) h_t, \quad \textit{with } \rho_t = \frac{\mu}{L_t} \left(\frac{\mathrm{mDW}(\mathcal{B})}{\mathrm{radius}(\mathcal{B})} \right)^2 , \tag{150}$$

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

Proof By Lemma 9 and bounding $\|\boldsymbol{d}_t\|$ by $R = \mathrm{radius}(\mathcal{B})$ we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \min_{\eta \in \mathbb{R}} \left\{ \eta \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle + \frac{\eta^2 L_t R^2}{2} \right\} \tag{151}$$

$$= f(\boldsymbol{x}_t) - \frac{\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle^2}{2 L_t R^2} \tag{152}$$

$$\leq f(\boldsymbol{x}_t) - \delta^2 \frac{\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t^\star \rangle^2}{2 L_t R^2} \tag{153}$$

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 17 the orthogonal projection of $\nabla f(\boldsymbol{x}_t)$ onto $\mathrm{lin}(\mathcal{B})$. Then the previous inequality simplifies to

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \delta^2 \frac{\langle \nabla_{\mathcal{B}} f(\boldsymbol{x}_t), \boldsymbol{s}_t^\star \rangle^2}{2 L_t R^2} \, . \tag{154}$$

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

$$h_{t+1} \leq h_t - \delta^2 \frac{\|\nabla_{\mathcal{B}} f(\boldsymbol{x}_t)\|_{\mathcal{B}*}^2}{2 L_t R^2} \tag{155}$$

To derive the lower-bound, we use Lemma 17 with $\boldsymbol{x} = \boldsymbol{x}_t$ and see that

$$\|\nabla_{\mathcal{B}} f(\boldsymbol{x}_t)\|_{\mathcal{B}^*} \geq 2\mu \, \mathrm{mDW}(\mathcal{B})^2 h_t \tag{156}$$

Combining the upper and lower bound together we have

$$h_{t+1} \leq \Big(1 - \delta^2 \frac{\mu \, \mathrm{mDW}(\mathcal{B})^2}{L_t R^2}\Big) h_t \, , \tag{157}$$

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

$$\underset{\|\boldsymbol{x}\|_1 \leq \beta}{\arg\min} \ \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{a}_i^\top \boldsymbol{x}, \boldsymbol{b}_i) + \frac{\lambda}{2} \|\boldsymbol{x}\|_2^2 \,, \tag{158}$$

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).

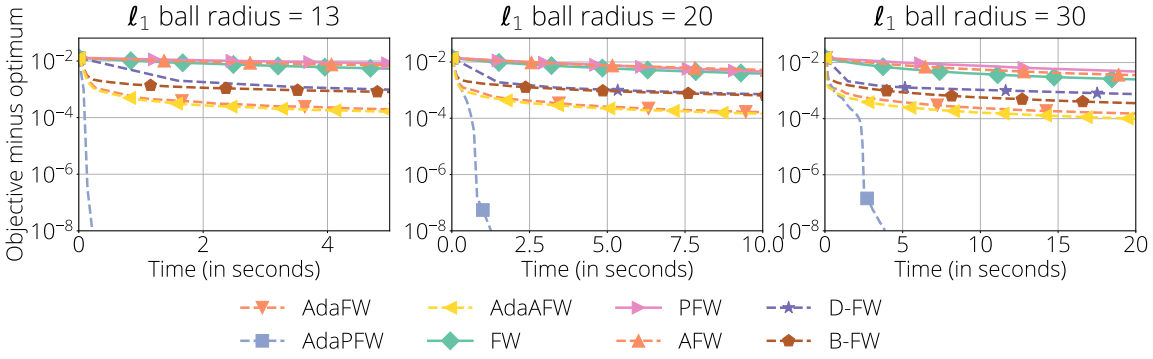

Figure 2: **Comparison of different FW variants**. Problem is ℓ_1-regularized logistic regression and dataset is Madelon in the first, RCV1 in the second figure.

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:

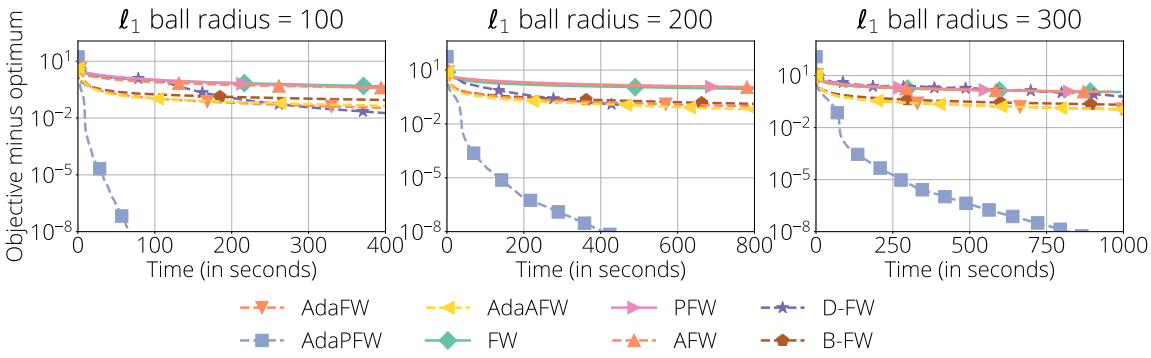

Figure 3: **Comparison of different FW variants**. Problem is ℓ_1-regularized logistic regression and dataset is RCV1.

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 [9] 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

$$\underset{\|\boldsymbol{X}\|_* \leq \beta}{\arg\min} \frac{1}{n} \sum_{(i,j)\in\mathcal{I}}^{n} L_\xi(\boldsymbol{A}_{i,j} - \boldsymbol{X}_{i,j}) \,, \tag{159}$$

where H_1 is the Huber loss, defined as

$$L_\xi(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \xi, \\ \xi(|a| - \frac{1}{2}\xi), & \text{otherwise}\,. \end{cases} \tag{160}$$

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).

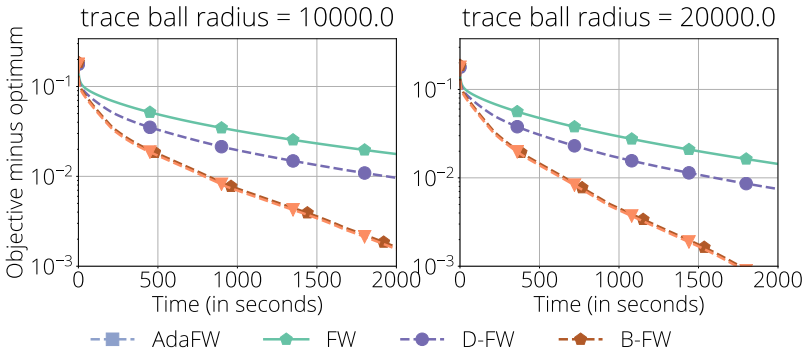

Figure 4: **Comparison of different FW variants**. Comparison of FW variants on the Movielens 1M dataset.

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