Frank-Wolfe with Subsampling Oracle
Supplementary material

Appendix notations. We denote by $\mathbf{E}_t$ the conditional expectation at iteration t, conditioned on all the past and by $\mathbb{E}$ a full expectation. We denote by a tilde the values that come from the deterministic analysis of FW. Denote by $\boldsymbol{r}_t = -\nabla f(\boldsymbol{x}_t)$. For $k \in \mathbb{N}^*$, denote by $[k]$ all integer between 1 and k.

Appendix A. Proof of sub-linear convergence for Randomized Frank-Wolfe

In this section we provide a convergence proof for Algorithm 1. The proof is loosely inspired by that of (Locatello et al., 2017, Appendix B.1), with the obvious difference that the result of the LMO is a random variable in our case.

Theorem 2.1′. *Let f be a function with bounded curvature constant C_f, Algorithm 1 for $\eta \in (0, 1]$, (with step-size chosen by either variants) converges towards a solution of (OPT), satisfying*

$$\mathbb{E}(f(\boldsymbol{x}_T)) - f(\boldsymbol{x}^*) \leq \frac{2(C_f + f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*))}{\eta T + 2} \,. \tag{9}$$

Proof. By definition of the curvature constant, at iteration t we have

$$f(\boldsymbol{x}_t + \gamma(\boldsymbol{s}_t - \boldsymbol{x}_t)) \leq f(\boldsymbol{x}_t) + \gamma\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t\rangle + \frac{\gamma^2}{2} C_f \,. \tag{10}$$

By minimizing with respect to γ on $[0, 1]$ we obtain

$$\gamma_t = \text{clip}_{[0,1]}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t\rangle / C_f \quad , \tag{11}$$

which is the definition of γ_t in the algorithm with Variant 2. Hence, we have

$$f(\boldsymbol{x}_{t+1}) \quad \leq \quad f(\boldsymbol{x}_t) + \min_{\gamma\in[0,1]} \left\{ \gamma\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t\rangle + \frac{\gamma^2}{2} C_f \right\},$$

an inequality which is also valid for Variant 1 since by the line search procedure the objective function at $\boldsymbol{x}_{t+1}$ is always equal or smaller than that of Variant 1. Denote by $h_t = f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$,

$$h_{t+1} \leq h_t + \min_{\gamma\in[0,1]} \left\{ \gamma\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t\rangle + \frac{\gamma^2}{2} C_f \right\}.$$

We write $\widetilde{\boldsymbol{s}}_t$ the FW atom if we had started the FW algorithm at $\boldsymbol{x}_t$, and $\mathbf{E}_t$ the expectation conditionned on all the past until $\boldsymbol{x}_t$, we have

$$\begin{aligned}
\mathbf{E}_t h_{t+1} &\leq h_t + \mathbf{E}_t \min_{\gamma\in[0,1]} \left\{ \gamma\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t\rangle + \frac{\gamma^2}{2} C_f \right\} && (12)\\
&\leq h_t + \mathcal{P}(\boldsymbol{s}_t = \widetilde{\boldsymbol{s}}_t) \min_{\gamma\in[0,1]} \left\{ \gamma\langle \nabla f(\boldsymbol{x}_t), \widetilde{\boldsymbol{s}}_t - \boldsymbol{x}_t\rangle + \frac{\gamma^2}{2} C_f \right\} && (13)\\
&\leq h_t + \eta \min_{\gamma\in[0,1]} \left\{ -\gamma h(\boldsymbol{x}_t) + \frac{\gamma^2}{2} C_f \right\} && (14)\\
&\leq h_t + \eta\big(-\gamma h(\boldsymbol{x}_t) + \frac{\gamma^2}{2} C_f \big) \quad \text{(for any } \gamma \in [0, 1], \text{ by definition of min)} \,, && (15)
\end{aligned}$$

where the second inequality follows from the definition of expectation and the fact that minimum is non-positive since it is zero for $\gamma = 0$. The last inequality is a consequence of uniform sampling as well as it uses that the FW gap is an upper bound on the dual gap, e.g. $\langle -\nabla f(\boldsymbol{x}_t), \widetilde{\boldsymbol{s}}_t - \boldsymbol{x}_t\rangle \geq h(\boldsymbol{x}_t)$.

Induction. From (15) the following is true for any $\gamma \in [0,1]$

$$\mathbf{E}_t(h_{t+1}) \leq h_t(1-\eta\gamma) + \frac{\gamma^2}{2}\eta C_f \ . \tag{16}$$

Taking unconditional expectation and writing $H_t = \mathbb{E}(h_t)$, we get for any $\gamma \in [0,1]$

$$H_{t+1} \leq H_t(1-\eta\gamma) + \frac{\gamma^2}{2}\eta C_f. \tag{17}$$

With $\gamma_t = \frac{2}{\eta t+2} \in [0,1]$, we get by induction

$$H_t \leq 2\frac{C_f+\epsilon_0}{\eta t+2} = \gamma_t(C_f+\epsilon_0), \tag{18}$$

where $\epsilon_0 = f(x_0) - f(x^*)$. Initialization follows the fact that the curvature constant is positive. For $t > 0$, from (17) and the induction hypothesis

$$\begin{aligned}
H_{t+1} &\leq \gamma_t(C_f+\epsilon_0)(1-\eta\gamma_t) + \frac{\gamma_t^2}{2}\eta C_f \\
&\leq \gamma_t(C_f+\epsilon_0)(1-\eta\gamma_t) + \frac{\gamma_t^2}{2}\eta(C_f+\epsilon_0) \\
&\leq \gamma_t(C_f+\epsilon_0)(1-\eta\gamma_t + \frac{\gamma_t}{2}\eta) \\
&\leq (C_f+\epsilon_0)(1-\frac{\gamma_t}{2}\eta)\gamma_t \\
&\leq (C_f+\epsilon_0)\gamma_{t+1}.
\end{aligned}$$

The last inequality comes from the fact that $(1-\frac{\gamma_t}{2}\eta)\gamma_t \leq \gamma_{t+1}$. Indeed, with $\gamma_t = \frac{2}{\eta t+2}$, it is equivalent to

$$\begin{aligned}
(1-\frac{\eta}{\eta t+2})\frac{2}{\eta t+2} &\leq \frac{2}{\eta(t+1)+2} \\
\Leftrightarrow \frac{(\eta t+2)-\eta}{\eta t+2} &\leq \frac{\eta t+2}{\eta(t+1)+2} \\
\Leftrightarrow (\eta t+2-\eta)(\eta(t+1)+2) &\leq (\eta t+2)^2 \\
\Leftrightarrow \eta^2t^2+4\eta t+4-\eta^2 &\leq \eta^2t^2+4\eta t+4.
\end{aligned}$$

The last being true, it concludes the proof. ■

Appendix B. Proof of linear convergence for RAFW

Away curvature and geometric strong convexity. The *away curvature* constant is a modification of the curvature constant described in the previous subsection, in which the FW direction $\boldsymbol{s} - \boldsymbol{x}$ is replaced with an arbitrary direction $\boldsymbol{s} - \boldsymbol{v}$:

$$C_f^A \stackrel{\text{def}}{=} \sup_{\substack{\boldsymbol{x},\boldsymbol{s},\boldsymbol{v}\in\mathcal{M} \\ \gamma\in[0,1] \\ \boldsymbol{y}=\boldsymbol{x}+\gamma(\boldsymbol{s}-\boldsymbol{v})}} \frac{2}{\gamma^2}\big(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \gamma\langle\nabla f(\boldsymbol{x}), \boldsymbol{s}-\boldsymbol{v}\rangle\big) \,.$$

The *geometric strong convexity* constant μ_f depends on both the function and the domain (in contrast to the standard strong convexity definition) and is defined as (see "An Affine Invariant Notion of Strong Convexity" in (Lacoste-Julien & Jaggi, 2015) for more details)

$$\mu_f^A = \inf_{\boldsymbol{x}\in\mathcal{M}} \inf_{\substack{\boldsymbol{x}^*\in\mathcal{M} \\ \langle\nabla f(\boldsymbol{x}), \boldsymbol{x}^*-\boldsymbol{x}\rangle<0}} \frac{2}{\gamma^A(\boldsymbol{x},\boldsymbol{x}^*)^2} B_f(\boldsymbol{x},\boldsymbol{x}^*)$$

where $B_f(\boldsymbol{x},\boldsymbol{x}^*) = f(\boldsymbol{x}^*) - f(\boldsymbol{x}) - \langle\nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x}\rangle$ and $\gamma^A(\boldsymbol{x},\boldsymbol{x}^*)$ the positive step-size quantity:

$$\gamma^A(\boldsymbol{x},\boldsymbol{x}^*) \coloneqq \frac{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x}\rangle}{\langle -\nabla f(\boldsymbol{x}), \boldsymbol{s}_f(\boldsymbol{x}) - \boldsymbol{v}_f(\boldsymbol{x})\rangle}.$$

In particular $\boldsymbol{s}_f(\boldsymbol{x})$ is the Frank Wolfe atom starting from $\boldsymbol{x}$. $\boldsymbol{v}_f(\boldsymbol{x})$ is the away atom when considering all possible expansions of $\boldsymbol{x}$ as a convex combinations of atoms in $\mathcal{A}$. Denote by $\mathcal{S}_{\boldsymbol{x}} \coloneqq \{\mathcal{S} \mid \mathcal{S} \subseteq \mathcal{A}$ such that $\boldsymbol{x}$ is a proper convex combination of all elements in $\mathcal{S}\}$ and by $\boldsymbol{v}_{\mathcal{S}(\boldsymbol{x})} \coloneqq \arg\max_{\boldsymbol{v}\in\mathcal{S}}\langle\nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle$. $\boldsymbol{v}_f(\boldsymbol{x})$ is finally defined by

$$\boldsymbol{v}_f(\boldsymbol{x}) \stackrel{\text{def}}{=} \arg\min_{\{\boldsymbol{v}=\boldsymbol{v}_{\mathcal{S}} \mid \mathcal{S}\in\mathcal{S}_{\boldsymbol{x}}\}} \langle\nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle \,.$$

Following (Lacoste-Julien & Jaggi, 2015, Lemma 9 in Appendix F), the geometric $\tilde{\mu}$-generally-strongly-convex constant is defined as

$$\tilde{\mu}_f = \inf_{\boldsymbol{x}\in\mathcal{M}} \inf_{\substack{\boldsymbol{x}^*\in\chi^* \\ \langle\nabla f(\boldsymbol{x}), \boldsymbol{x}^*-\boldsymbol{x}\rangle<0}} \frac{1}{2\gamma^A(\boldsymbol{x},\boldsymbol{x}^*)^2}\big(f(\boldsymbol{x}^*) - f(\boldsymbol{x}) - 2\langle\nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x}\rangle\big) \,,$$

where χ^* represents the solution set of (OPT).

Notations. In the context of RAFW, $\mathcal{A}$ denotes the finite set of extremes atoms such that $\mathcal{M} = \text{Conv}(\mathcal{A})$. At iteration t, $\mathcal{A}_t$ is a random subset of element of $\mathcal{A} \setminus \mathcal{S}_t$ where $\mathcal{S}_t$ is the current support of the iterate. The Randomized LMO is performed over $\mathcal{V}_t = \mathcal{S}_t \cup \mathcal{A}_t$ so that for Algorithm 2, $\boldsymbol{s}_t \stackrel{\text{def}}{\in} \arg\max_{\boldsymbol{v}\in\mathcal{V}_t}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{v}\rangle$ is the FW atom at iteration t for RAFW.

Note that when $|\mathcal{A} \setminus \mathcal{S}_t| \leq p$, Algorithm 2 does exactly the same as AFW. For the sake of simplicity we will consider that this is not the case. Indeed we would otherwise fall back into the deterministic setting and the proof would just be that of (Lacoste-Julien & Jaggi, 2015).

We use tilde notation for quantities that are specific to the deterministic FW setting. For instance, $\widetilde{\boldsymbol{s}}_t \stackrel{\text{def}}{\in} \arg\max_{\boldsymbol{v}\in\mathcal{A}}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{v}\rangle$ is the FW atom for AFW starting at $\boldsymbol{x}_t$.

Similarly the Away atom is such that $\boldsymbol{v}_t \stackrel{\text{def}}{\in} \arg\min_{v\in\mathcal{S}_t}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{v}\rangle$ and it does not depend on the sub-sampling at iteration t. Here we do not use any tilde because it is a quantity that appears both in AFW and its Randomized counterpart.

In AFW, $\widetilde{g}_t \stackrel{\text{def}}{=} \langle -\nabla f(\boldsymbol{x}_t), \widetilde{\boldsymbol{s}}_t - \boldsymbol{v}_t\rangle = \max_{\boldsymbol{s}\in\mathcal{A}}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{v}_t\rangle$ is an upper-bound of the dual gap, named the *pair-wise dual gap* (Lacoste-Julien & Jaggi, 2015). We consider the corresponding *partial pair-wise dual gap* $\widetilde{g}_t \stackrel{\text{def}}{=} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t\rangle = \max_{\boldsymbol{s}\in\mathcal{V}_t}\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{v}_t\rangle$. It is partial is the sense that the maximum is computed on a subset $\mathcal{V}_t$ of $\mathcal{A}$ which results in the fact that it is not guaranteed anymore to be an upper-bound on the dual-gap.

Structure of the proof. The proof is structured around a main part that uses Lemmas 1 and 3. Lemma 2 is only used to prove Lemma 3.

The main proof follows the scheme of the deterministic one of AFW in (Lacoste-Julien & Jaggi, 2015, Theorem 8). It is divided in three parts. The first part consists in upper bounding $h_t \overset{\text{def}}{=} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$ with $\tilde{g}_t$. It does not depend on the specific construction of the iterates $\boldsymbol{x}_t$ and thus remains the same as that in (Lacoste-Julien & Jaggi, 2015). The second part provides a lower bound on the progress on the algorithm, namely

$$h_{t+1} \leq (1 - \rho_f \big(\frac{g_t}{\widetilde{g}_t}\big)^2) h_t, \tag{19}$$

with $\rho_f = \frac{\mu_f^A}{4C_f^A}$, when it is not doing a *bad drop step* (defined above). As a proxy for this event, we use the binary variable z_t that equals 0 for bad drop steps and 1 otherwise.

The difficulty lies in that we guarantee a geometrical decrease only when $g_t = \widetilde{g}_t$ and $z_t = 1$. Because of the sub-sampling and unlike in the deterministic setting, z_t is a random variable. Lemma 3 provides a lower bound on the probability of interest, $\mathcal{P}(\widetilde{g}_t = g_t \mid z_t = 1)$, for the last part of the main proof.

Finally, the last part of the proof constructs a bound on the number of times we can expect both $z_t = 1$ and $g_t = \widetilde{g}_t$ subject to the constraint that at least half of the iterates satisfy $z_t = 1$. It is done by recurrence.

Appendix B.1. Lemmas

This lemma ensures the chosen direction $\boldsymbol{d}_t$ in RAFW is a good descent direction, and links it with g_t which may be equal to $\widetilde{g}_t$.

Lemma 1. *Let $\boldsymbol{s}_t, \boldsymbol{v}_t$ and $\boldsymbol{d}_t$ be as defined in Algorithm 2. Then for $g_t \overset{\text{def}}{=} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t \rangle$, we have*

$$\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle \geq \frac{1}{2} g_t \geq 0 \,. \tag{20}$$

Proof. The first inequality appeared already in the convergence proof of Lacoste-Julien & Jaggi (2015, Eq. (6)), which we repeat here for completeness. By the definition of $\boldsymbol{d}_t$ we have:

$$\begin{aligned} 2\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle &\geq \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t^A \rangle + \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t^{\text{FW}} \rangle \\ &= \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t \rangle = g_t \end{aligned} \tag{21}$$

We only need to prove that g_t is non-negative. In line 3 of algorithm 2, $\boldsymbol{s}_t$ is the output of LMO performs of the set of atoms $\mathcal{S}_t \cup \mathcal{A}_t \overset{\text{def}}{=} \mathcal{V}_t$,

$$\boldsymbol{s}_t = \arg\max_{\boldsymbol{s} \in \mathcal{V}_t} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle \,,$$

so that we have $\langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t \rangle \geq \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t \rangle$. By definition of g_t, it implies $g_t \geq 0$. ■

Lemma 2 is just a simple combinatorial result needed in Lemma 3. Consider a sequence of m numbers, we lower bound the probability for the maximum of a subset of size greater than p to be equal to the maximum of the sequence.

Lemma 2. *Consider any sequence $(r_i)_{i \in \mathcal{I}}$ in $\mathbb{R}$ with $\mathcal{I} = \{1, \cdots, m\}$, and a subset $\mathcal{I}_p \subseteq \mathcal{I}$ of size p. We have*

$$\mathcal{P}(\max_{i \in \mathcal{I}_p} r_i = \max_{i \in \mathcal{I}} r_i) \geq \frac{p}{m} \,. \tag{22}$$

Proof. Consider $M = \{i \in \mathcal{I} \mid r_i = \max_{j \in \mathcal{I}} r_j\}$. We have $\max_{i \in \mathcal{I}_p} r_i = \max_{i \in \mathcal{I}} r_i$ if and only if at least one element of $\mathcal{I}_p$ belongs to M:

$$\mathcal{P}(\max_{i \in \mathcal{I}_p} r_i = \max_{i \in \mathcal{I}} r_i) = \mathcal{P}(|\mathcal{I}_p \cap M| \geq 1) \,. \tag{23}$$

By definition M has at least one element i_0. Since $\{i_0 \in \mathcal{I}_p\} \subset \{|\mathcal{I}_p \cap M| \geq 1\}$

$$\mathcal{P}(|\mathcal{I}_p \cap M| \geq 1) \geq \mathcal{P}(\{i_0 \in \mathcal{I}_p\}) \,. \tag{24}$$

All subsets are taken uniformly at random, we just have to count the number of subset $\mathcal{I}_p$ of $\mathcal{I}$ of size p with $i_0 \in \mathcal{I}_p$

$$\mathcal{P}(\{i_0 \in \mathcal{I}_p\}) = \frac{\binom{m-1}{p-1}}{\binom{m}{p}} = \frac{p}{m} \tag{25}$$

$$\mathcal{P}(\max_{i \in \mathcal{I}_p} r_i = \max_{i \in \mathcal{I}} r_i) \geq \frac{p}{m} \,. \tag{26}$$

■

In the second part of the main proof we ensure a geometric decrease when both $g_t = \widetilde{g}_t$ and $z_t = 1$, i.e. outside of *bad drop steps*. The following lemma helps quantifying the probability of $g_t = \widetilde{g}_t$ holding when $z_t = 1$.

Lemma 3. *Consider g_t (defined in Lemma 1) to be the partial pair-wise (PW) dual gap of RAFW at iteration t with sub-sampling parameter p on the constrained polytope $\mathcal{M} = \operatorname{conv}(\mathcal{A})$, where $\mathcal{A}$ is a finite set of extremes points of $\mathcal{M}$. $\widetilde{g}_t \overset{def}{=} \max_{\boldsymbol{s} \in \mathcal{A}} \langle -\nabla f(\boldsymbol{x}_t), s - \boldsymbol{v}_t \rangle$ is the pairwise dual gap of AFW starting at $\boldsymbol{x}_t$ on this same polytope. Denote by z_t the binary random variable that equals 0 when the t^{th} iteration of RAFW makes an away step that is a drop step with $\gamma_{max} < 1$ (a bad drop step), and 1 otherwise. Then we have the following bound*

$$\mathcal{P}(g_t = \widetilde{g}_t \mid \boldsymbol{x}_t, z_t = 1) \geq \Big(\frac{p}{|\mathcal{A}|}\Big)^2. \tag{PROB}$$

Proof. Recall that $g_t^A \overset{\text{def}}{=} \langle \boldsymbol{r}_t, \boldsymbol{d}_t^A \rangle$. By definition $\{z_t = 0\} = \{g_t < g_t^A, \gamma_{\max} < 1, \gamma_t^* = \gamma_{\max}\}$, where $\gamma_t^* \overset{\text{def}}{=} \arg\min_{\gamma \in [0, \gamma_{\max}]} f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t^A)$. Its complementary $\{z_t = 1\}$ can thus be expressed as the partition $A_1 \cup A_2 \cup A_3$ where the A_i are defined by

$$A_1 = \{g_t \geq g_t^A\} \quad \text{(performs a FW step)} \tag{27}$$

$$A_2 = \{g_t < g_t^A \,,\, \alpha_{\boldsymbol{v}_t}^{(t)}/(1 - \alpha_{\boldsymbol{v}_t}^{(t)}) \geq 1\} \quad \text{(performs away step with } \gamma_{\max} \geq 1) \tag{28}$$

$$A_3 = \{g_t < g_t^A \,,\, \alpha_{\boldsymbol{v}_t}^{(t)}/(1 - \alpha_{\boldsymbol{v}_t}^{(t)}) < 1 \,,\, \gamma_t^* < \alpha_{\boldsymbol{v}_t}^{(t)}/(1 - \alpha_{\boldsymbol{v}_t}^{(t)})\}. \tag{29}$$

First note that in the case of A_2 and A_3, $\gamma_{\max} = \alpha_{\boldsymbol{v}_t}^{(t)}/(1 - \alpha_{\boldsymbol{v}_t}^{(t)})$. Though the right hand side formulation highlights that it is entirely determined by $\boldsymbol{x}_t$, recalling that $\alpha_{\boldsymbol{v}_t}^{(t)}$ is the mass along the atom $\boldsymbol{v}_t$ in the decomposition of $\boldsymbol{x}_t$ in §3.

From a higher level perspective, these cases are those for which we can guarantee a geometrical decrease of $h_t = f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$ (see second part of main proof). By definition, the A_i are disjoints. A_1 represents a choice of a FW step in RAFW contrary to A_2 and A_3 which stands for an away step choice in RAFW. A_2 is an away step for which there is enough potential mass ($\gamma_{\max} > 1$) to move along the away direction and to ensure sufficient objective decreasing. A_3 encompasses the situations where there is not a lot of mass along the away direction ($\gamma_{\max} < 1$) but which is not a drop step, e.g. the amount of mass is not a limit to the descent.

Our goal is to lower bound $P = \mathcal{P}(g_t = \widetilde{g}_t \mid \boldsymbol{x}_t, z_t = 1)$. The following probabilities will be with respect to the t^{th} sub-sampling only. Notice that g_t^A, $\widetilde{g}_t$ and α_{v_t} are known given $\{\boldsymbol{x}_t, z_t = 1\}$. Using Bayes' rule, and because the A_i are disjoints, we have

$$\begin{aligned} P &= \mathcal{P}(g_t = \widetilde{g}_t \mid \boldsymbol{x}_t, \{z_t = 1\}) \\ &= \frac{\sum_{i=1}^3 \mathcal{P}(g_t = \widetilde{g}_t \mid \boldsymbol{x}_t, A_i)\mathcal{P}(A_i \mid \boldsymbol{x}_t)}{\sum_{i=1}^3 \mathcal{P}(A_i \mid \boldsymbol{x}_t)} \,. \end{aligned} \tag{30}$$

By definition of g_t and $\widetilde{g}_t$, $g_t \leq \widetilde{g}_t$, so that measuring the probability of an event like $\{g_t = \widetilde{g}_t\}$ conditionally on $\{g_t \leq g_t^A\}$ will naturally depend on whether or not, the deterministic condition $\widetilde{g}_t \geq g_t^A$ is satisfied. Hence the following case distinction.

Recall $\mathcal{V}_t = \mathcal{S}_t \cup \mathcal{A}_t$.

Case $\widetilde{g}_t < g_t^A$.

$$P = \frac{\sum_{i=1}^3 \mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_i, \widetilde{g}_t < g_t^A)\mathcal{P}(A_i \mid x_t, \widetilde{g}_t < g_t^A)}{\sum_{i=1}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t < g_t^A)} \,. \tag{31}$$

Recall that $A_1 = \{g_t \geq g_t^A\}$. Since by definition $g_t \leq \widetilde{g}_t$, conditionally on $\{\widetilde{g}_t < g_t^A\}$, the probability of A_1 is zero. Consequently the above reduces to

$$\begin{aligned} P &= \frac{\sum_{i=2}^3 \mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_i, \widetilde{g}_t < g_t^A)\mathcal{P}(A_i \mid x_t, \widetilde{g}_t < g_t^A)}{\sum_{i=2}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t < g_t^A)} \\ &\geq \frac{p}{|\mathcal{A}|}\frac{\sum_{i=2}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t < g_t^A)}{\sum_{i=2}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t \leq g_t^A)} = \frac{p}{|\mathcal{A}|} \,. \end{aligned} \tag{32}$$

Where the last inequality is because for $i = 2, 3$ we have $\mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_i, \widetilde{g}_t < g_t^A) \geq \frac{p}{|\mathcal{A}|}$. Indeed for A_3 (case A_2 is similar) denote

$$\begin{aligned} P_1 &= \mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_3, \widetilde{g}_t < g_t^A) \\ &= \mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle < C_0, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle < C_0, \alpha_{\boldsymbol{v}_t}^{(t)}/(1-\alpha_{\boldsymbol{v}_t}^{(t)}) < 1, \gamma_t^* < \alpha_{\boldsymbol{v}_t}^{(t)}/(1-\alpha_{\boldsymbol{v}_t}^{(t)})) \,. \end{aligned} \tag{33}$$

with $C_0 \stackrel{\text{def}}{=} g_t^A + \langle \boldsymbol{r}_t, \boldsymbol{v}_t\rangle$ and $\boldsymbol{r}_t = -\nabla f(\boldsymbol{x}_t)$ not depending on the t^{th} sub-sampling. Also the event $\{\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle < C_0\}$ is a consequence of $\{\max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle < C_0\}$ so that P_1 simplifies to

$$P_1 = \mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle < C_0, \alpha_{\boldsymbol{v}_t}^{(t)}/(1-\alpha_{\boldsymbol{v}_t}^{(t)}) < 1, \gamma_t^* < \alpha_{\boldsymbol{v}_t}^{(t)}/(1-\alpha_{\boldsymbol{v}_t}^{(t)})) \,. \tag{34}$$

By definition

$$\gamma_t^* \in \operatorname*{arg\,min}_{\gamma\in[0,\frac{\alpha_{\boldsymbol{v}_t}^{(t)}}{1-\alpha_{\boldsymbol{v}_t}^{(t)}}]} f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t^A) \,, \tag{35}$$

so that γ_t^* does not depend on the t^{th} sub-sampling. Finally all the conditioning in the probability of (34) do not depend on this t^{th} sub-sampling. Hence we are in the position of using Lemma 2 for the sequence $(\langle \boldsymbol{r}_t, \boldsymbol{s}\rangle)_{\boldsymbol{s}\in\mathcal{A}}$. Also by definition of $\mathcal{V}_t = \mathcal{S}_t \cup \mathcal{A}_t$, we have $|\mathcal{V}_t| \geq p$ so that we finally get

$$\mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_3, \widetilde{g}_t < g_t^A) \geq \frac{p}{|\mathcal{A}|} \,. \tag{36}$$

This was what was needed to conclude (32).

Case $\widetilde{g}_t \geq g_t^A$. In such a case, P from (30) rewrites as

$$P = \frac{\sum_{i=1}^3 \mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_i, \widetilde{g}_t \geq g_t^A)\mathcal{P}(A_i \mid x_t, \widetilde{g}_t \geq g_t^A)}{\sum_{i=1}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t \geq g_t^A)} \,. \tag{37}$$

Here $\mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_i, \widetilde{g}_t \geq g_t^A) = 0$ for $i = 2, 3$ because A_i implies $g_t < g_t^A$. So that when $\widetilde{g}_t \geq g_t^A$ it is then impossible for g_t to equal $\widetilde{g}_t$.

$$P = \frac{\mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_1, \widetilde{g}_t \geq g_t^A)\mathcal{P}(A_1 \mid x_t, \widetilde{g}_t \geq g_t^A)}{\sum_{i=1}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t \geq g_t^A)} \,.$$

Here also we use, and prove later on (see §below the conclusion of the proof of the Lemma), the lower bound

$$\mathcal{P}(g_t = \widetilde{g}_t \mid x_t, A_1, \widetilde{g}_t \geq g_t^A) \geq \frac{p}{|\mathcal{A}|} \,, \tag{38}$$

that implies

$$P \geq \frac{p}{|\mathcal{A}|} \frac{\mathcal{P}(A_1 \mid x_t, \widetilde{g}_t \geq g_t^A)}{\sum_{i=1}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t \geq g_t^A)} .$$

Because the A_i are disjoint, $\sum_{i=1}^3 \mathcal{P}(A_i \mid x_t, \widetilde{g}_t \geq g_t^A) \leq 1$ we have

$$P \geq \frac{p}{|\mathcal{A}|} \mathcal{P}(A_1 \mid x_t, \widetilde{g}_t \geq g_t^A) .$$

Using a similar lower bound as (38), namely

$$\mathcal{P}(A_1 \mid x_t, \widetilde{g}_t \geq g_t^A) \geq \frac{p}{|\mathcal{A}|} , \tag{39}$$

we finally get

$$P \geq \left(\frac{p}{|\mathcal{A}|}\right)^2 . \tag{40}$$

Since it is hard to precisely count the occurrences of $\{\widetilde{g}_t \geq g_t^A\}$ and $\{\widetilde{g}_t < g_t^A\}$, we use a conservative bound in (40)

$$\mathcal{P}(g_t = \widetilde{g}_t \mid \boldsymbol{x}_t, z_t = 1) \geq \left(\frac{p}{|\mathcal{A}|}\right)^2 . \tag{41}$$

This will of course make our bound on the rate of convergence very conservative.

Justification for (38) and (39).

Lets denote the left hand side of(38) by P_2. By definition of g_t and $\widetilde{g}_t$, with $\boldsymbol{r}_t = -\nabla f(\boldsymbol{x}_t)$, we have:

$$P_2 = \mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s} - \boldsymbol{v}_t\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s} - \boldsymbol{v}_t\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s} - \boldsymbol{v}_t\rangle \geq g_t^A, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s} - \boldsymbol{v}_t\rangle \geq g_t^A) \tag{42}$$

$$= \mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0) , \tag{43}$$

where $C_0 \stackrel{\text{def}}{=} g_t^A + \langle \boldsymbol{r}_t, v_t\rangle$ and $\boldsymbol{r}_t$ does not depend on the random sampling at iteration t. Bayes formula leads to

$$P_2 = \frac{\mathcal{P}(\{\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle\} \cap \{\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0\} \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0)}{\mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0 \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0)} . \tag{44}$$

Conditionally on $\{\max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0\}$, the event $\{\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle\}$ implies$\{\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0\}$ which leads to

$$\begin{aligned} P_2 &= \frac{\mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0)}{\mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0 \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0)} \\ &\geq \mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0) \geq \frac{p}{|\mathcal{A}|} , \end{aligned}$$

where the last inequality is a consequence of applying Lemma 2 on the sequence $(\langle \boldsymbol{r}_t, \boldsymbol{s}\rangle)_{\boldsymbol{s}\in\mathcal{A}}$

Similarly let's denote the left hand side of (39) by P_3. The first inequality is justified because conditionally on $\{\widetilde{g}_t \geq g_t^A\}$, $\{g_t = \widetilde{g}_t\} \subset \{g_t \geq g_t^A\}$ and the last by applying, similarly as for (38), Lemma 2 on the sequence $(\langle \boldsymbol{r}_t, \boldsymbol{s}\rangle)_{\boldsymbol{s}\in\mathcal{A}}$.

$$\begin{aligned} P_3 &= \mathcal{P}(g_t \geq g_t^A \mid \boldsymbol{x}_t, \widetilde{g}_t \geq g_t^A) \\ &\geq \mathcal{P}(g_t = \widetilde{g}_t \mid \boldsymbol{x}_t, \widetilde{g}_t \geq g_t^A), \\ &\geq \mathcal{P}(\max_{\boldsymbol{s}\in\mathcal{V}_t} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \mid \boldsymbol{x}_t, \max_{\boldsymbol{s}\in\mathcal{A}} \langle \boldsymbol{r}_t, \boldsymbol{s}\rangle \geq C_0) \\ &\geq \frac{p}{|\mathcal{A}|} . \end{aligned}$$

■

Appendix B.2. Main proof

Theorem 3.1′. *Consider the set $\mathcal{M} = \operatorname{conv}(\mathcal{A})$, with $\mathcal{A}$ a finite set of extreme atoms, after T iterations of Algorithm 2 (RAFW) we have the following linear convergence rate*

$$\mathbb{E}\big[h(\boldsymbol{x}_{T+1})\big] \leq \big(1 - \eta^2 \rho_f\big)^{\max\{0, \lfloor (T-s)/2 \rfloor\}} h(\boldsymbol{x}_0) \,, \tag{45}$$

with $\rho_f = \frac{\mu_f^A}{4 C_f^A}$, $\eta = \frac{p}{|\mathcal{A}|}$ and $s = |\mathcal{S}_0|$.

Proof. The classical curvature constant used in proofs related to non-Away Frank-Wolfe is

$$C_f \coloneqq \sup_{\substack{\boldsymbol{x},\boldsymbol{s} \in \mathcal{M}, \gamma \in [0,1] \\ \boldsymbol{y} = \boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{v})}} \frac{2}{\gamma^2} \big(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \big) \,. \tag{46}$$

It is tailored for algorithms in which the update is of the form $\boldsymbol{x}_{t+1} = (1-\gamma)\boldsymbol{x}_t + \gamma \boldsymbol{v}_t$, but this is not the shape of all updates in away versions of FW. In (Lacoste-Julien & Jaggi, 2015) they introduced a modification of the above curvature constant that we also use to analyze RAFW. It is defined in (Lacoste-Julien & Jaggi, 2015, equation (26)) as

$$C_f^A \coloneqq \sup_{\substack{\boldsymbol{x},\boldsymbol{s},\boldsymbol{v} \in \mathcal{M}, \gamma \in [0,1] \\ \boldsymbol{y} = \boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{v})}} \frac{2}{\gamma^2} \big(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \gamma \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} - \boldsymbol{v} \rangle \big) \,. \tag{47}$$

It differs from C_f (46) because it allows to move outside of the domain $\mathcal{M}$. We thus require L-lipschitz continuous function on any compact set for that quantity to be upper-bounded. We refer to §**curvature constants** on (Lacoste-Julien & Jaggi, 2015, Appendix D) for thorough details. The first part of the proof reuses the scheme of (Lacoste-Julien & Jaggi, 2015, Theorem 8).

First part. *Upper bounding h_t*: Considering an iterate $\boldsymbol{x}_t$ that is not optimal (e.g. $\boldsymbol{x}_t \neq \boldsymbol{x}^*$), from (Lacoste-Julien & Jaggi, 2015, Eq. (28)), we have

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) = h_t \leq \frac{\widetilde{g}_t^2}{2\mu_f^A} \,, \tag{48}$$

where $\widetilde{g}_t$ is the *pair-wise dual gap* defined by $\widetilde{g}_t = \langle \widetilde{\boldsymbol{s}}_t - \boldsymbol{v}_t, -\nabla f(\boldsymbol{x}_t) \rangle$. $\widetilde{\boldsymbol{s}}_t$ and $\boldsymbol{v}_t$ are respectively the FW atom and the away atom in the classical Away step algorithm (conditionally on $\boldsymbol{x}_t$, the away atom of the randomized variant coincides with the away atom of the non-randomized variant). The result is still valid here as it only uses the definition of the affine invariant version of the strong convexity parameter and does not depend on the way $\boldsymbol{x}_t$ are constructed (see *upper bounding h_t* in (Lacoste-Julien & Jaggi, 2015, Proof for AFW in Theorem 8)).

Note that this implicitly assumes the away atom to be defined, e.g. the support of the iterate $\boldsymbol{x}_t$ never to be zero. This is ensured by the algorithm simply because it always does convex updates.

Second part. *Lower bounding progress $h_t - h_{t+1}$*. Consider $\boldsymbol{x}_t$ a non-optimal iterate. At step t, the update in Algorithm 2 writes $\boldsymbol{x}_{t+1}(\gamma) = \boldsymbol{x}_t + \gamma \boldsymbol{d}_t$. γ is optimized by line-search in the segment $[0, \gamma_{\max}]$. Because in either cases $\boldsymbol{d}_t$ is a difference between two elements of $\mathcal{M}$, from the definition of C_f^A and because of the exact line search, we have

$$f(\boldsymbol{x}_{t+1}) \leq \min_{\gamma \in [0, \gamma_{\max}]} \big(f(\boldsymbol{x}_t) + \gamma \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{\gamma^2}{2} C_f^{\mathcal{A}} \big) \,,$$

so that for any $\gamma \in [0; \gamma_{\max}]$

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) \leq \gamma \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{\gamma^2}{2} C_f^{\mathcal{A}}$$

or again

$$\gamma \frac{g_t}{2} - \frac{\gamma^2}{2} C_f^A \leq f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}), \tag{49}$$

where the last inequality is a consequence of Lemma 1. We write $\gamma_t^B \stackrel{\text{def}}{=} \frac{g_t}{2C_f^A} \geq 0$, the minimizer of the left hand side of (49).

Case $\gamma_{\max} \geq 1$ and $\gamma_t^B \leq \gamma_{\max}$. (49) evaluated on $\gamma = \gamma_t^B$ gives

$$\frac{g_t^2}{4C_f^A} - \frac{g_t^2}{8C_f^A} \leq f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1})$$
$$\implies \left(\frac{g_t}{\widetilde{g}_t}\right)^2 \frac{\widetilde{g}_t^2}{8C_f^A} \leq h_t - h_{t+1}. \tag{50}$$

Indeed, $\boldsymbol{x}_t$ is assumed not to be optimal, so that $\widetilde{g}_t \neq 0$. Combining (50) with (48) gives

$$h_{t+1} \leq h_t - \left(\frac{g_t}{\widetilde{g}_t}\right)^2 \frac{\widetilde{g}_t^2}{8C_f^A} \tag{51}$$
$$\leq h_t - \left(\frac{g_t}{\widetilde{g}_t}\right)^2 \frac{\mu_f^A}{4C_f^A} h_t \tag{52}$$
$$= \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right) h_t \,. \tag{53}$$

Case $\gamma_{\max} \geq 1$ and $\gamma_t^B > \gamma_{\max}$. $\gamma_t^B = \frac{g_t}{2C_f^A}$ implies $g_t \geq 2C_f^A$. (49) transforms into

$$\frac{g_t}{2}\left(\gamma - \frac{\gamma^2}{2}\right) \leq f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1})$$
$$\frac{g_t}{\widetilde{g}_t}\frac{\widetilde{g}_t}{2}\left(\gamma - \frac{\gamma^2}{2}\right) \leq f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \,.$$

Using $\widetilde{g}_t \geq h_t$ and evaluating at $\gamma = 1$, leaves us with

$$h_{t+1} \leq \left(1 - \frac{1}{4}\frac{g_t}{\widetilde{g}_t}\right) h_t. \tag{54}$$

Because $\mu_f^A \leq C_f^A$ (Lacoste-Julien & Jaggi, 2015, Remark 7.) and $\rho_f = \frac{\mu_f^A}{4C_f^A}$, the two previous cases resolve in the following inequality

$$h_{t+1} \leq \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right) h_t \,. \tag{55}$$

Case $\gamma_{\max} < 1$ and $\gamma_t^* < \gamma_{\max}$. By definition

$$\gamma_t^* \quad = \quad \arg\min_{\gamma \in [0, \gamma_{\max}]} f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t) = F(\gamma) \,. \tag{56}$$

f is convex and its minimum on $[0; \gamma_{\max}]$ is not reached at $\gamma_{\max}$. It is then also a minimum on the interval $[0; +\infty]$, and in particular we have

$$\gamma_t^* \quad = \quad \arg\min_{\gamma \in [0, 1]} f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t) = F(\gamma) \,. \tag{57}$$

(49) can then be written with $\gamma \in [0, 1]$ which leads to the previous case result (55).

Case $\gamma_{\max} < 1$ and $\gamma_t^* = \gamma_{\max}$. This corresponds to a particular drop step for which we only guarantee $h_{t+1} \leq h_t$ (exact line-search). We call this case a *bad drop step* (indeed $\gamma_{\max} > 1$ and $\gamma_t^* = \gamma_{\max}$ also corresponds to a drop step, but for which we can prove a bound of the form $h_{t+1} \leq h_t(1 - \rho_f\left(\frac{g_t}{\widetilde{g}_t}\right)^2)$).

We use the binary indicator z_t to distinguish between the step where (55) is guaranteed or not. Denote by $z_t = 0$ when doing a *bad drop step* and $z_t = 1$ otherwise. The second part can be summed-up in

$$h_{t+1} \leq h_t(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2)^{z_t}. \tag{58}$$

Last part. Consider starting RAFW (Algorithm 2) for T iterations at $\boldsymbol{x}_0 \in \operatorname{conv}(\mathcal{V})$, with $s = |\mathcal{S}_0| \geq 0$. We will now prove there are at most $\lfloor \frac{T+s}{2} \rfloor$ drop steps. Let D_T be the number of drop steps after iteration T and F_T the number of FW step adding a new atom until iteration T. By definition, a FW step is not a drop step so that $D_T + F_T \leq T$. Also $|S_T| = |S_0| + |F_T| - |D_T|$, hence $|\mathcal{S}_T| \leq |\mathcal{S}_0| - 2|D_T| + T$ so that $|D_T| \leq \frac{T+s-|\mathcal{S}_T|}{2}$. Finally because $|\mathcal{S}_T| \geq 0$, we have $|D_T| \leq \lfloor \frac{T+s}{2} \rfloor$.

From the first two parts of the main proof, we have that

$$h_T \leq h_0 \prod_{t=0}^{T-1} \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right)^{z_t}, \tag{59}$$

where $(g_t, z_t)_{t \in [0:T-1]}$ are defined along RAFW starting at $\boldsymbol{x}_0$. For $i < j$, we write $\mathbb{E}_{i:j}$ the expectation with respect to all sub-sampling between the i^{th} iteration and the j^{th} iteration included. When taking expectation only over sub-sampling i, we write it $\mathbb{E}_i$.

We will now prove by recurrence on $T \in \mathbb{N}^*$ that

$$\mathbb{E}_{0:T-1}\left(\prod_{t=0}^{T-1} \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right)^{z_t}\right) \leq (1 - \rho_f \eta^2)^{\max\{0, T - \lfloor \frac{T+s}{2} \rfloor\}} = F(T, s) \quad \forall s \in \mathbb{N} \ \forall \boldsymbol{x}_0 \in \mathbb{R}^d \text{ with } |\mathcal{S}_0| = s\,, \tag{60}$$

where $\boldsymbol{x}_0 = \sum_{\boldsymbol{v} \in \mathcal{A}} \alpha_{\boldsymbol{v}}^{(0)} \boldsymbol{v}$ and $\mathcal{S}_0 = \{\boldsymbol{v} \in \mathcal{A} \text{ s.t. } \alpha_{\boldsymbol{v}}^{(0)} > 0\}$.

The rate quantity $\max\{0, T - \lfloor \frac{T+s}{2} \rfloor\}$ represents the number of steps (between iteration 0 and $T-1$) in which $z_t = 1$, e.g. the steps in which there is a possibility of having geometrical decrease. Note that the geometrical decrease happens only when $g_t = \widetilde{g}_t$.

The key insight in the global bound is to recall (from section 3) that if the support is a singleton, i.e. $|\mathcal{S}_t| = 1$, RAFW does a FW step hence $z_t = 1$. We consequently distinguish whether or not the first iterate has an initial support of size 1. We then use the recurrence property starting the algorithm at $\boldsymbol{x}_1$ and running $T-1$ iterations.

Initialization. We will now prove the recurrence property (60) for $T = 1$. If $s \geq 2$, $\max\{0, T - \lfloor \frac{T+s}{2} \rfloor\} = 0$ and (60) is true because $\left(1 - \rho_f \left(\frac{g_0}{\widetilde{g}_0}\right)^2\right) \leq 1$. If $s = 1$, this implies that the first step needs to be a Frank-Wolfe step. We necessarily have $z_0 = 1$ and so

$$\begin{aligned}
\mathbb{E}_0\left(\left(1 - \rho_f \left(\frac{g_0}{\widetilde{g}_0}\right)^2\right)^{z_0}\right) &= \mathbb{E}_0\left(\left(1 - \rho_f \left(\frac{g_0}{\widetilde{g}_0}\right)^2\right) \mid z_0 = 1\right) && (61)\\
&\leq 1 - \rho_f \mathcal{P}(g_0 = \widetilde{g}_0 \mid z_0 = 1) && (62)\\
&\leq 1 - \rho_f \eta^2 \leq 1 \leq F(1,1)\,, && (63)
\end{aligned}$$

with $\eta = \frac{p}{|\mathcal{A}|}$ where F is defined in (60) and where the last inequality follows from (PROB) in Lemma 3.

Recurrence. Consider the property (60) when running $T-1$ iteration. By the tower property of conditional expectations

$$\mathbb{E}_{0:T-1}\left(\prod_{t=0}^{T-1} \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right)^{z_t}\right) = \mathbb{E}_{0:T-1}\left[\left(1 - \rho_f \left(\frac{g_0}{\widetilde{g}_0}\right)^2\right)^{z_0} \mathbb{E}_{1:T-1}\left(\prod_{t=1}^{T-1} \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right)^{z_t}\right)\right]. \tag{64}$$

We can apply the recurrence property with $T-1$ iterations and starting point $\boldsymbol{x}_1$ on $\mathbb{E}_{1:T-1}(\prod_{t=1}^{T-1} \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right)^{z_t})$ so that

$$\mathbb{E}_{0:T-1}\left(\prod_{t=0}^{T-1} \left(1 - \rho_f \left(\frac{g_t}{\widetilde{g}_t}\right)^2\right)^{z_t}\right) \leq \mathbb{E}_0\left[\left(1 - \rho_f \left(\frac{g_0}{\widetilde{g}_0}\right)^2\right)^{z_0} F(T-1, |\mathcal{S}_1|)\right]\,, \tag{65}$$

where $|\mathcal{S}_1|$, the support of $\boldsymbol{x}_1$, depends on z_0. Indeed $z_0 = 0$ implies a drop step and as such it decreases the support of the iterate. Thus we have to distinguish the case according to the size of the support of $\boldsymbol{x}_0$.

Case $|\mathcal{S}_0| = 1$. With $\boldsymbol{x}_0 = 0$, RAFW starts with a FW step and as such $z_0 = 1$ as well as $2 \geq |\mathcal{S}_1| \geq 1$ so that

$$\mathbb{E}_{0:T-1}\Big(\prod_{t=0}^{T-1}\big(1-\rho_f\big(\frac{g_t}{\widetilde{g}_t}\big)^2\big)^{z_t}\Big) = \mathbb{E}_0\big[\big(1-\rho_f\big(\frac{g_0}{\widetilde{g}_0}\big)^2\big) \mid z_0 = 1\big]F(T-1,|\mathcal{S}_1|) \tag{66}$$

$$\leq (1-\rho_f\eta^2)F(T-1,2) \leq F(T,1)\,, \tag{67}$$

by applying (PROB) in Lemma 3. The last equality concludes the heredity in that case.

Case $|\mathcal{S}_0| \geq 2$. Here it is possible for z_0 to equal 0 or 1. If $z_0 = 1$, then $|\mathcal{S}_1| \leq |\mathcal{S}_0| + 1$, while if $z_0 = 0$, it implies a drop step, we have $|\mathcal{S}_1| = |\mathcal{S}_0| - 1$. If we decompose the expectation according to the value of z_0 we obtain

$$\mathbb{E}_{0:T-1}\Big(\prod_{t=0}^{T-1}\big(1-\rho_f\big(\frac{g_t}{\widetilde{g}_t}\big)^2\big)^{z_t}\Big) \leq \mathcal{P}(z_0=1)\mathbb{E}_0\big[\big(1-\rho_f\big(\frac{g_0}{\widetilde{g}_0}\big)^2\big) \mid z_0 = 1\big]F(T-1,|\mathcal{S}_1|) \tag{68}$$

$$+\mathcal{P}(z_0=0)F(T-1,|\mathcal{S}_0|-1) \tag{69}$$

$$\leq \mathcal{P}(z_0=1)\big(1-\rho_f\eta^2\big)F(T-1,|\mathcal{S}_0|+1) + \mathcal{P}(z_0=0)F(T-1,|\mathcal{S}_0|-1) \tag{70}$$

$$\leq \mathcal{P}(z_0=1)\big(1-\rho_f\eta^2\big)F(T-1,s+1) + \mathcal{P}(z_0=0)F(T-1,s-1)\,. \tag{71}$$

We used the fact that $F(T,|\mathcal{S}_1|) \leq F(T-1,|\mathcal{S}_0|+1)$. Since we do not have access to the values of $\mathcal{P}(z_0=0)$ and $\mathcal{P}(z_0=1)$, we bound it in the following manner

$$\mathbb{E}_{0:T-1}\Big(\prod_{t=0}^{T-1}\big(1-\rho_f\big(\frac{g_t}{\widetilde{g}_t}\big)^2\big)^{z_t}\Big) \leq \max\big((1-\rho_f\eta^2)F(T-1,s+1), F(T-1,s-1)\big) \leq F(T,s)\,, \tag{72}$$

where the last inequality is just about writing the definition of F. It concludes the heredity result.

Conclusion: Starting RAFW at $\boldsymbol{x}_0$, after T iterations, we have

$$h_T \leq h_0 \prod_{t=0}^{T-1}\big(1-\rho_f\big(\frac{g_t}{\widetilde{g}_t}\big)^2\big)^{z_t}\,. \tag{73}$$

Applying (60) we get

$$\mathbb{E}_{0:T-1}(h_T) \leq h_0(1-\rho_f\eta^2)^{\max\{0,T-\lfloor\frac{T+s}{2}\rfloor\}}$$

$$\leq h_0(1-\rho_f\eta^2)^{\max\{0,\lfloor\frac{T-s}{2}\rfloor\}}\,. \tag{74}$$

■

Generalized strongly convex.

Theorem 3.2′. *Suppose f has bounded smoothness constant C_f^A and is $\tilde{\mu}$-generally-strongly convex. Consider the set $\mathcal{M} = conv(\mathcal{A})$, with $\mathcal{A}$ a finite set of extreme atoms. Then after T iterations of Algorithm 2, with $s = |\mathcal{S}_0|$ and a p parameter of sub-sampling, we have*

$$\mathbb{E}\big[h(\boldsymbol{x}_{T+1})\big] \leq \big(1-\eta^2\tilde{\rho}_f\big)^{\max\{0,\lfloor\frac{T-s}{2}\rfloor\}}h(\boldsymbol{x}_0)\,, \tag{75}$$

with $\tilde{\rho}_f = \frac{\tilde{\mu}}{4C_f^A}$ and $\eta = \frac{p}{|\mathcal{A}|}$.

Proof. The conclusion of proof of (Lacoste-Julien & Jaggi, 2015, Th. 11) is that we have similarly as equation (48) by:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) = h_t \leq \frac{g_t^2}{2\tilde{\mu}_f}\,, \tag{76}$$

where $\tilde{\mu}_f > 0$ is a similar measure of the affine invariant strong convexity constant but for generalized strongly convex function.

We can thus write the twin of equation (58)

$$h_{t+1} \leq h_t\big(1 - \tilde{\rho}_f\big(\frac{g_t}{\widetilde{g}_t}\big)^2\big)^{z_t} \ , \tag{77}$$

with $\tilde{\rho}_f = \frac{\tilde{\mu}_f}{4C_f^A}$. The rest of the proof follows is the same as that of Theorem 3.1. ■

Appendix C. Technical issues of previous work

In this section we highlight some technical issues present in previous work.

Appendix C.1. Randomized Frank-Wolfe in Frandi et al. (2016)

Frandi et al. (2016) present a Randomized FW algorithm for the case of the ℓ_1 ball in $\mathbb{R}^d$. Denote by $\mathcal{A} = \{\pm \boldsymbol{e}_i \ \forall i \in [d]\}$, where $\boldsymbol{e}_i$ is the canonical basis (i.e., the vector that is zero everywhere except on the i-th coordinate, where it equals one) the extremes atoms of the ℓ_1 ball. Up to the iterative explicit implementation of the *residuals*, (Frandi et al., 2016, Algorithm 2) with the sampling size $p \in [n]$ and our RFW (Algorithm 1) are equivalent for the following choice of $\mathcal{A}_t$ in RFW

$$\mathcal{A}_t = \{\pm \boldsymbol{e}_i \ \ \forall i \in \mathcal{I}_p\} \ , \quad \text{where } \mathcal{I}_p \text{ is random subset of } [d] \text{ of size } p. \tag{78}$$

Convergence result. In this case, (Frandi et al., 2016, Proposition 2) gives the following convergence bound in expectation after t iterations:

$$\mathbb{E}(f(\boldsymbol{x}_t)) - f(\boldsymbol{x}^*) \leq \frac{4C_f}{t+2} \ . \tag{79}$$

First, it is rather surprising that, unlike in our Theorem 2.1, the sub-sampling size p does not appear in the convergence bound. A closer inspection at their Lemma 2 reveals some errors in their proof. For the remainder of this section we will use the notation in (Frandi et al., 2016).

The point of interest. The proof of their Proposition 2 starts with the following inequality derived from the curvature constant:

$$f(\alpha_\lambda^{(k+1)}) \leq f(\alpha^{(k)}) + \lambda \big(u^{(k)} - \alpha^{(k)}\big)^T \nabla f(\alpha^{(k)}) + \lambda^2 C_f \ . \tag{80}$$

Then it is claimed that the following equation, Eq. (24) in their paper, is a direct consequence "after some algebraic manipulations"

$$\mathbb{E}_{\mathcal{S}^{(k)}}\big[f(\alpha_\lambda^{(k+1)})\big] \leq f(\alpha^{(k)}) + \lambda \mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)} - \alpha^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] + \lambda^2 C_f \ . \tag{81}$$

which is not clear unless $u^{(k)}$ is independent of the sampling set, something that is not verified given that it is chosen *precisely from* the sampling set.

Technical details. λ being positive, for Eq. (81) to be true, we should necessarily have the following

$$\mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)} - \alpha^{(k)}\big)^T \nabla f(\alpha^{(k)})\big] \leq \mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)} - \alpha^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] \ . \tag{82}$$

$\alpha^{(k)}$ as well as $\nabla f(\alpha^{(k)})$ are deterministic with respect to the $\mathcal{S}^{(k)}$ sampling set so the previous equation is equivalent to

$$\mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \nabla f(\alpha^{(k)})\big] - \big(\alpha^{(k)}\big)^T \nabla f(\alpha^{(k)}) \leq \mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] - \big(\alpha^{(k)}\big)^T \mathbb{E}_{\mathcal{S}^{(k)}}\big[\widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] \tag{83}$$

Since the sub-sampling of $\mathcal{S}^{(k)}$ is uniform and by definition of $\widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})$ in (Frandi et al., 2016, equation (14)) we have $\mathbb{E}_{\mathcal{S}^{(k)}}\big[\widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] = \nabla f(\alpha^{(k)})$. Then (82) is equivalent to

$$\mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \nabla f(\alpha^{(k)})\big] \quad \leq \quad \mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] \ . \tag{84}$$

Also by definition in (Frandi et al., 2016, equation (22)), $u^{(k)}$ the FW atom has its support on $\mathcal{S}^{(k)}$ as well as from (Frandi et al., 2016, equation (6)) we have that $\big(u^{(k)}\big)^T \nabla f(\alpha^{(k)}) < 0$. So that $\big(u^{(k)}\big)^T \nabla f(\alpha^{(k)}) = \big(u^{(k)}\big)^T \nabla_{\mathcal{S}^{(k)}} f(\alpha^{(k)})$ and finally (82) is equivalent to

$$\frac{|\mathcal{S}^{(k)}|}{p} \mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] \quad \leq \quad \mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] \ , \tag{85}$$

this last inequality being false in general because $\frac{|\mathcal{S}^{(k)}|}{p} < 1$ and $\mathbb{E}_{\mathcal{S}^{(k)}}\big[\big(u^{(k)}\big)^T \widetilde{\nabla}_{\mathcal{S}^{(k)}} f(\alpha^{(k)})\big] \leq 0$.