Self-concordant analysis of Frank-Wolfe algorithms - Supplementary Material-

Pavel Dvurechensky [1] **Petr Ostroukhov** [2] **Kamil Safin** [2] **Shimrit Shtern** [3] **Mathias Staudigl** [4]

Outline

The supplementary material of this paper is organized as follows.

- Appendix A contains further details on self-concordant (SC) functions.
- Appendix B is devoted to proof of Theorem 3.1 for Variant 1 of Algorithm 2. Since this proof relies on some standard estimates on self-concordant functions, we include those auxiliary estimates as well.
- Appendix C is organized around the convergence proof of Variant 2 of Algorithm 2, which is Theorem 3.4 in the main text. We also give some guidelines how the parameters and initial values of the backtracking subroutine are chosen.
- Appendix D contains the linear convergence proof under the availability of the restricted local linear minimization oracle (LLOO)
- Appendix E collects detailed evaluations of the numerical performances of the Algorithms constructed in this paper in the context of the Portfolio optimization and the Poisson inverse problem.
- Appendix F outlines the construction of the LLOO for simplex constraints, following (Garber & Hazan, 2016).

A. Proofs of Section 2

We first introduce a classical result on SC functions, showing its affine invariance.

Lemma A.1 (Nesterov (2018), Thm. 5.1.2)**.** *Let $f \in \mathcal{F}_M$ and $\mathcal{A}(x) = Ax + b : \mathbb{R}^n \to \mathbb{R}^p$ a linear operator. Then $\tilde{f} \triangleq f \circ \mathcal{A} \in \mathcal{F}_M$.*

When we apply Frank-Wolfe (FW) to the minimization of a function $f \in \mathcal{F}_M$, the search direction at position x is determined by the target state $s(x) = s$ defined in (4). If $A : \tilde{\mathcal{X}} \to \mathcal{X}$ is a surjective linear re-parametrization of the domain $\mathcal{X}$, then the new optimization problem $\min_{\tilde{\mathcal{X}}} \tilde{f}(\tilde{x}) = f(A\tilde{x})$ is still within the frame of problem (P). Furthermore, the updates produced by FW are not affected by this re-parametrization since

$$\langle \nabla \tilde{f}(\tilde{x}), \hat{s} \rangle = \langle \nabla f(A\tilde{x}), A\hat{s} \rangle = \langle \nabla f(x), s \rangle$$

for $x = A\tilde{x} \in \mathcal{X}, s = A\hat{s} \in \mathcal{X}$.

Proposition A.2. *Suppose there exists $x \in \operatorname{dom} f \cap \mathcal{X}$ such that $\|\nabla f(x)\|_x^* \leq \frac{2}{M}$. Then (P) admits a unique solution.*

Proof. For all $x, y \in \operatorname{dom} f$ we know that

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{4}{M^2}\omega\left(\frac{M}{2}\|y - x\|_x^2\right) \\ &\leq f(x) - \|\nabla f(x)\|_x^* \cdot \|y - x\|_x \\ &+ \frac{4}{M^2}\omega\left(\frac{M}{2}\|y - x\|_x^2\right) \\ &= f(x) + \left(\frac{2}{M} - \|\nabla f(x)\|_x^*\right)\|y - x\|_x \\ &- \frac{4}{M^2}\ln\left(1 + \frac{M}{2}\|y - x\|_x\right). \end{aligned}$$

Define the level set $\mathcal{L}_f(\alpha) \triangleq \{x | f(x) \leq \alpha\}$ and pick $y \in \mathcal{L}_f(f(x))$. For such a point, we get

$$\begin{aligned} &\frac{4}{M^2}\ln\left(1 + \frac{M}{2}\|y - x\|_x\right) \\ &\qquad \geq \left(\frac{2}{M} - \|\nabla f(x)\|_x^*\right)\|y - x\|_x. \end{aligned}$$

Consider the function $t \mapsto \varphi(t) \triangleq \frac{\ln(1+t)}{t}$ for $t > 0$. For $t > 0$ it is true that $\varphi(t) < 1$, and so we need that $\|\nabla f(x)\|_x^* \leq \frac{2}{M}$. Since $\lim_{t\to\infty} \varphi(t) = 0$, it follows that $\mathcal{L}_f(f(x))$ is bounded. By the Weierstrass theorem, existence of a solution follows (see e.g. (Bertsekas, 1999)). If $x^* \in \operatorname{dom} f \cap \mathcal{X}$ is a solution, we know that

$$f(x) \geq f(x^*) + \frac{4}{M^2}\omega(\frac{M}{2}\|x - x^*\|_{x^*}).$$

Hence, if x would be any alternative solution, we immediately conclude that $x = x^*$. ■

B. Proofs of convergence of Variant 1 of Algorithm 2

This supplementary material contains all results needed to establish the convergence of Version 1 of Algorithm 2. We start with some basic estimates helping to proof the main result about this numerical scheme.

B.1. Preliminary Results

We recall the basic inequalities for SC functions.

$$f(\tilde{x}) \geq f(x) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{4}{M^2}\omega\left(\mathsf{d}(x, \tilde{x})\right) \quad (1)$$

$$f(\tilde{x}) \leq f(x) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{4}{M^2}\omega_*\left(\mathsf{d}(x, \tilde{x})\right) \quad (2)$$

We need a preliminary error bound around the unique solution.

Lemma B.1. *For all $x \in \operatorname{dom} f$ we have:*

$$\frac{4}{M^2}\omega\left(\frac{M}{2}\|x - x^*\|_{x^*}\right) \leq f(x) - f(x^*).$$

Proof. If $x \in \operatorname{dom} f \cap \mathcal{X}$, then eq. (1) shows

$$\begin{aligned} f(x) &\geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{4}{M^2}\omega\left(\mathsf{d}(x^*, x)\right) \\ &\geq f(x^*) + \frac{4}{M^2}\omega(\mathsf{d}(x^*, x)). \end{aligned}$$

■

We next prove a restricted strong convexity property of SC functions.

Lemma B.2. *For all $x \in \mathcal{S}(x^0)$ we have*

$$f(x) - f(x^*) \geq \frac{\sigma_f}{6}\|x - x^*\|_2^2. \quad (3)$$

Proof. Lemma B.1 gives $f(x) - f(x^*) \geq \frac{4}{M^2}\omega(\mathsf{d}(x^*, x))$. Observe that for all $t \in [0, 1]$

$$\omega(t) = t - \ln(1 + t) = \sum_{j=2}^{\infty} \frac{(-1)^j t^j}{j} \geq \frac{t^2}{2} - \frac{t^3}{3} \geq \frac{t^2}{6}.$$

Coupled with the fact that $x^* \in \mathcal{S}(x^0)$ and the hypothesis that $x \in \mathcal{S}(x^0)$, we see that

$$f(x) - f^* \geq \frac{1}{6}\|x - x^*\|_{x^*}^2 \geq \frac{\sigma_f}{6}\|x - x^*\|_2^2.$$

■

Also, we need the next classical fact for SC functions.

Lemma B.3. *Let $\mathcal{W}(x, r) = \{x' \in \mathbb{R}^n | \frac{M}{2}\|x' - x\|_x < r\}$ denote the Dikin ellipsoid with radius r around x. For all $x \in \operatorname{dom} f$ we have $\mathcal{W}(x, 1) \subset \operatorname{dom} f$.*

Proof. See Nesterov (2018). ■

B.2. Estimates for the Algorithm

For $x \in \operatorname{dom} f$, define the target vector

$$s(x) = \operatorname*{argmin}_{x \in \mathcal{X}} \langle \nabla f(x), x \rangle, \quad (4)$$

and

$$\mathsf{Gap}(x) = \langle \nabla f(x), x - s(x) \rangle. \quad (5)$$

Moreover, for all $x \in \operatorname{dom} f$, let us define

$$\mathsf{e}(x) \triangleq \mathsf{d}(x, s(x)) = \frac{M}{2}\|s(x) - x\|_x. \quad (6)$$

Given $x \in \mathcal{X}$ and $t > 0$, set $x_t^+ \triangleq x + t(s(x) - x)$. Assume that $\mathsf{e}(x) \neq 0$. By construction,

$$\mathsf{d}(x, x_t^+) = \frac{tM}{2}\|s(x) - x\|_x = t\mathsf{e}(x) < 1,$$

iff $t < 1/\mathsf{e}(x)$. Choosing $t \in (0, 1/\mathsf{e}(x))$, we conclude from (2)

$$\begin{aligned} f(x_t^+) &\leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{4}{M^2}\omega_*(t\mathsf{e}(x)) \\ &\leq f(x) - t\,\mathsf{Gap}(x) + \frac{4}{M^2}\omega_*(t\mathsf{e}(x)) \end{aligned}$$

This reveals the interesting observation that for minimizing an SC-function, we can search for a step size α_k which minimizes the model function

$$\eta_x(t) \triangleq t\,\mathsf{Gap}(x) - \frac{4}{M^2}\omega_*(t\mathsf{e}(x)), \quad (7)$$

defined for $t \in (0, 1/\mathsf{e}(x))$

Proposition B.4. *For $x \in \operatorname{dom} f \cap \mathcal{X}$, the function $t \mapsto \eta_x(t)$ defined in (7) is concave and uniquely maximized at the value*

$$\mathsf{t}(x) \triangleq \frac{\mathsf{Gap}(x)}{\mathsf{e}(x)(\mathsf{Gap}(x) + \frac{4}{M^2}\mathsf{e}(x))} \equiv \frac{\gamma(x)}{\mathsf{e}(x)}. \quad (8)$$

If $\alpha(x) \triangleq \min\{1, \mathsf{t}(x)\}$ is used as a step-size in Variant 1 of Algorithm 2, and define $\Delta(x) \triangleq \eta_x(\alpha(x))$, then

$$f(x + \alpha(x)(s(x) - x)) \leq f(x) - \Delta(x). \quad (9)$$

Proof. For $x \in \operatorname{dom} f \cap \mathcal{X}$, define

$$\eta_x(t) \triangleq t\,\mathsf{Gap}(x) - \frac{4}{M^2}\omega_*(t\mathsf{e}(x)). \quad (10)$$

We easily compute $\eta_x''(t) = \frac{4}{M^2}\frac{\mathsf{e}(x)}{(1 - t\mathsf{e}(x))^2} > 0$. Hence, the function is concave and uniquely maximized at

$$\mathsf{t}(x) \triangleq \frac{\mathsf{Gap}(x)}{\mathsf{e}(x)(\mathsf{Gap}(x) + \frac{4}{M^2}\mathsf{e}(x))} \equiv \frac{\gamma(x)}{\mathsf{e}(x)}. \quad (11)$$

Furthermore, one can easily check that $\eta_x(0) = 0$, and $\eta_x(\mathtt{t}(x)) = \frac{4}{M^2}\omega\left(\frac{M^2}{4}\frac{\mathsf{Gap}(x)}{\mathsf{e}(x)}\right) > 0$, whenever $\mathsf{e}(x) > 0$. Hence, it follows that

$$\eta_x(t) > 0 \quad \forall t \in (0, \mathtt{t}(x)]. \tag{12}$$

■

We now construct the step size sequence $(\alpha_k)_{k\geq 0}$ by setting $\alpha_k = \min\{1, \mathtt{t}(x^k)\}$ for all $k \geq 0$. Convexity of $\mathcal{X}$ and the fact that $\alpha_k \mathsf{e}(x^k) < 1$ guarantees that $(x^k)_{k\geq 0} \subset \operatorname{dom} f \cap \mathcal{X}$. For the feasibility, we use Lemma B.3. Thus, at each iteration, we reduce the objective function value by at least the quantity $\Delta_k \equiv \eta_{x^k}(\alpha_k)$, so that $f(x^{k+1}) \leq f(x^k) - \Delta_k < f(x^k)$.

Proposition B.5. *The following assertions hold for Variant 1 of Algorithm 2:*

(a) $\left(f(x^k)\right)_{k\geq 0}$ *is non-increasing;*

(b) $\sum_{k\geq 0}\Delta_k < \infty$*, and hence the sequence* $(\Delta_k)_{k\geq 0}$ *converges to 0;*

(c) For all $K \geq 1$ *we have* $\min_{0\leq k<K}\Delta_k \leq \frac{1}{K}(f(x^0) - f^*)$.

Proof. This proposition can be deduced from Proposition 5.2 in (Dvurechensky et al., 2019). We give a proof for the sake of being self-contained. Evaluating eq. (9) along the iterate sequence, and calling $\Delta_k = \Delta(x^k)$, we get for all $k \geq 0$,

$$f(x^{k+1}) - f(x^k) \leq -\Delta_k.$$

Telescoping this expression shows that for all $K \geq 1$,

$$f(x^K) - f(x^0) \leq -\sum_{k=0}^{K-1}\Delta_k.$$

Since $\Delta_k > 0$, the sequence $\left(f(x^k)\right)_{k\geq 0}$ is monotonically decreasing. We conclude that for all $K \geq 1$,

$$\sum_{k=0}^{K-1}\Delta_k \leq f(x^0) - f(x^K) \leq f(x^0) - f^* \tag{13}$$

and therefore,

$$\min_{1\leq k\leq K}\Delta_k \leq \frac{1}{K}(f(x^0) - f^*). \tag{14}$$

Hence, $\lim_{k\to\infty}\Delta_k = 0$. ■

We can bound the sequence $(\mathsf{e}^k)_{k\geq 0}$, defined as $\mathsf{e}^k \triangleq \mathsf{e}(x^k)$, as

$$\frac{M\sqrt{\sigma_f}}{2}\|s^k - x^k\|_2 \leq \mathsf{e}^k \leq \frac{M\sqrt{L_{\nabla f}}}{2}\|s^k - x^k\|_2. \tag{15}$$

In order to derive convergence rates, we need to lower bound the per-iteration decrease in the objective function. A detailed analysis of the sequence $(\Delta_k)_{k\geq 0}$ reveals an explicit lower bound on the per-iteration decrease which relates the gap function to the sequence $(\Delta_k)_{k\geq 0}$.

Lemma B.6. *For all* $k \geq 0$ *we have*

$$\Delta_k \geq \min\{\mathtt{a}\,\mathsf{Gap}(x^k), \mathtt{b}\,\mathsf{Gap}(x^k)^2\}, \tag{16}$$

where $\mathtt{a} \triangleq \min\left\{\frac{1}{2}, \frac{2(1-\ln(2))}{M\sqrt{L_{\nabla f}}\operatorname{diam}(\mathcal{X})}\right\}$ *and* $\mathtt{b} \triangleq \frac{1-\ln(2)}{L_{\nabla f}\operatorname{diam}(\mathcal{X})^2}$.

Proof. Let us start with an iteration k at which $\alpha_k = \mathtt{t}(x^k)$. In this case, we make progress according to

$$\eta_{x^k}(\mathtt{t}(x^k)) = \frac{\mathsf{Gap}(x^k)}{\mathsf{e}(x^k)}\gamma(x^k) + \frac{4}{M^2}\gamma(x^k) + \frac{4}{M^2}\ln\left(\frac{(4/M^2)\mathsf{e}(x^k)}{\mathsf{Gap}(x^k) + (4/M^2)\mathsf{e}(x^k)}\right).$$

Define $y := \frac{(4/M^2)\mathsf{e}(x^k)}{\mathsf{Gap}(x^k)}$. Rewriting the above display in terms of this new variable, we arrive, after some elementary algebra, at the expression

$$\eta_{x^k}(t(x^k)) = \frac{\mathsf{Gap}(x^k)}{\mathsf{e}(x^k)}\left[1 + y\ln\left(\frac{y}{1+y}\right)\right].$$

Consider the function $\phi : (0,\infty) \to (0,\infty)$, given by $\phi(t) \triangleq 1 + t\ln\left(\frac{t}{1+t}\right)$. When $t \in (0,1)$, since

$$\begin{aligned}\phi'(t) &= \ln\left(\frac{t}{1+t}\right) + t\frac{1+t}{t}\left(\frac{1}{1+t} - \frac{t}{(1+t)^2}\right)\\ &= \ln\left(\frac{t}{1+t}\right) + 1 - \frac{t}{1+t}\\ &= \ln\left(1 - \frac{1}{1+t}\right) + \frac{1}{1+t} < 0,\end{aligned}$$

we conclude that $\phi(t)$ is decreasing for $t \in (0,1)$. Hence, $\phi(t) \geq \phi(1) = 1 - \ln 2$, for all $t \in (0,1)$. On the other hand, if $t \geq 1$,

$$\begin{aligned}\frac{d}{dt}\left(\frac{\phi(t)}{1/t}\right) &= \frac{d}{dt}(t\phi(t))\\ &= 1 + 2t\ln\left(\frac{t}{1+t}\right) + \frac{t}{1+t} \geq 0.\end{aligned}$$

Hence, $t \mapsto \frac{\phi(t)}{1/t}$ is an increasing function for $t \geq 1$, and thus $\phi(t) \geq \frac{1-\ln 2}{t}$, for all $t \geq 1$. We conclude that

$$\eta_{x^k}(\mathtt{t}(x^k)) \geq \frac{\mathsf{Gap}(x^k)}{\mathsf{e}(x^k)}(1 - \ln(2))\min\left\{1, \frac{\mathsf{Gap}(x^k)}{(4/M^2)\mathsf{e}(x^k)}\right\}.$$

Now consider an iteration k in which $\alpha_k = 1$. The per-iteration decrease of the objective function is explicitly given by

$$\eta_{x^k}(1) = \left[\mathsf{Gap}(x^k) + \frac{4}{M^2}\mathsf{e}(x^k)\right] + \frac{4}{M^2}\ln(1-\mathsf{e}(x^k)) = \mathsf{Gap}(x^k)\left[1 + y + \frac{y}{\mathsf{e}(x^k)}\ln(1-\mathsf{e}(x^k))\right].$$

Since $\alpha_k = 1$, it is true that $\mathsf{e}(x^k) < \gamma(x^k) < 1$, and therefore $\frac{1}{\mathsf{e}(x^k)}\ln(1-\mathsf{e}(x^k)) > \frac{1}{\gamma(x^k)}\ln(1-\gamma(x^k))$. Finally, using the identity $1 + y = \frac{1}{\gamma(x^k)}$, we arrive at the lower bound

$$\eta_{x^k}(1) \geq \mathsf{Gap}(x^k)\left[1 + y + y(1+y)\ln\left(\frac{y}{1+y}\right)\right] \geq \frac{\mathsf{Gap}(x^k)}{2}.$$

Summarizing all these computations, we see that for all $k \geq 0$, the per-iteration decrease is at least

$$\Delta_k \geq \min\left\{\frac{\mathsf{Gap}(x^k)}{2}, \frac{(1-\ln(2))\,\mathsf{Gap}(x^k)}{\mathsf{e}(x^k)}, \frac{(1-\ln(2))\,\mathsf{Gap}(x^k)^2}{(4/M^2)\mathsf{e}(x^k)^2}\right\}.$$

From eq. (15), we deduce that $\mathsf{e}(x) \leq \frac{M\sqrt{L_{\nabla f}}}{2}\operatorname{diam}(\mathcal{X})$. Hence, after setting $\mathtt{a} \triangleq \min\left\{\frac{1}{2}, \frac{2(1-\ln(2))}{M\sqrt{L_{\nabla f}}\operatorname{diam}(\mathcal{X})}\right\}$ and $\mathtt{b} \triangleq \frac{1-\ln(2)}{L_{\nabla f}\operatorname{diam}(\mathcal{X})^2}$, we see that

$$\Delta_k \geq \min\{\mathtt{a}\,\mathsf{Gap}(x^k), \mathtt{b}\,\mathsf{Gap}(x^k)^2\}.$$

■

B.3. Proof of Theorem 3.6

With the help of the lower bound in Lemma B.6, we are now able to establish the $\mathcal{O}(k^{-1})$ convergence rate in terms of the approximation error $h_k \triangleq f(x^k) - f^*$.

By convexity, we have $\mathsf{Gap}(x^k) \geq h_k$. Therefore, the lower bound for Δ_k can be estimated as $\Delta_k \geq \min\{\mathtt{a}h_k, \mathtt{b}h_k^2\}$, which implies

$$h_{k+1} \leq h_k - \min\{\mathtt{a}h_k, \mathtt{b}h_k^2\} \qquad \forall k \geq 0. \tag{17}$$

Given this recursion, we can identify two phases characterizing the process $(h_k)_{k\geq 0}$. In Phase I, the approximation error is at least $\mathtt{a}/\mathtt{b}$, and in Phase II the approximation error falls below this value.

For fixed initial condition $x^0 \in \operatorname{dom} f \cap \mathcal{X}$, we can subdivide the time domains according to Phases I and II as

$$\mathcal{K}_1(x^0) \triangleq \{k \geq 0 | h_k > \frac{\mathtt{a}}{\mathtt{b}}\}, \text{ (Phase I)}$$
$$\mathcal{K}_2(x^0) \triangleq \{k \geq 0 | h_k \leq \frac{\mathtt{a}}{\mathtt{b}}\}, \text{ (Phase II)}.$$

Since $(h_k)_k$ is monotonically decreasing and bounded from below by the positive constant $\mathtt{a}/\mathtt{b}$ on Phase I, the set $\mathcal{K}_1(x^0)$ is at most finite. Let us set

$$T_1(x^0) \triangleq \inf\{k \geq 0 | h_k \leq \frac{\mathtt{a}}{\mathtt{b}}\}, \tag{18}$$

the first time at which the process (h_k) enters Phase II. To get a worst-case estimate on this quantity, assume that $0 \in \mathcal{K}_1(x^0)$, so that $\mathcal{K}_1(x^0) = \{0, 1, \ldots, T_1(x^0) - 1\}$. Then, for all $k = 1, \ldots, T_1(x^0) - 1$ we have $\frac{\mathtt{a}}{\mathtt{b}} < h_k \leq h_{k-1} - \min\{\mathtt{a}h_{k-1}, \mathtt{b}h_{k-1}^2\} = h_{k-1} - \mathtt{a}h_{k-1}$. Note that $\mathtt{a} \leq 1/2$, so we make progressions like a geometric series. Hence, $h_k \leq (1-\mathtt{a})^k h_0$ for all $k = 0, \ldots, T_1(x^0) - 1$. By definition $h_{T_1(x^0)-1} > \frac{\mathtt{a}}{\mathtt{b}}$, so we get $\frac{\mathtt{a}}{\mathtt{b}} \leq h_0(1-\mathtt{a})^{T_1(x^0)-1}$ iff $(T_1(x^0) - 1)\ln(1-\mathtt{a}) \geq \ln\left(\frac{\mathtt{a}}{h_0\mathtt{b}}\right)$. Hence,

$$T_1(x^0) \leq \left\lceil \frac{\ln\left(\frac{\mathtt{a}}{h_0\mathtt{b}}\right)}{\ln(1-\mathtt{a})}\right\rceil \leq \left\lceil \frac{1}{\mathtt{a}}\ln\left(\frac{h_0\mathtt{b}}{\mathtt{a}}\right)\right\rceil. \tag{19}$$

After these number of iterations, the process will enter Phase II, at which $h_k \leq \frac{\mathtt{a}}{\mathtt{b}}$ holds. Therefore, $h_k \geq h_{k+1} + \mathtt{b}h_k^2$, or equivalently,

$$\frac{1}{h_{k+1}} \geq \frac{1}{h_k} + \mathtt{b}\frac{h_k}{h_{k+1}} \geq \frac{1}{h_k} + \mathtt{b}.$$

Pick $N > T_1(x^0)$ an arbitrary integer. Summing this relation for $k = T_1(x^0)$ up to $k = N-1$, we arrive at

$$\frac{1}{h_N} \geq \frac{1}{h_{T_1(x^0)}} + \mathtt{b}(N - T_1(x^0) + 1).$$

By definition $h_{T_1(x^0)} \leq \frac{\mathtt{a}}{\mathtt{b}}$, so that for all $N > T_1(x^0)$, we see

$$\frac{1}{h_N} \geq \frac{\mathtt{b}}{\mathtt{a}} + \mathtt{b}(N - T_1(x^0) + 1).$$

Consequently,

$$h_N \leq \frac{1}{\frac{\mathtt{b}}{\mathtt{a}} + \mathtt{b}(N - T_1(x^0) + 1)} \leq \frac{1}{\mathtt{b}(N - T_1(x^0) + 1)} = \frac{L_{\nabla f}\operatorname{diam}(\mathcal{X})^2}{(1-\ln(2))(N - T_1(x^0) + 1)}.$$

Define the stopping time $N_\varepsilon(x^0) \triangleq \inf\{k \geq 0 | h_k \leq \varepsilon\}$. Then, by definition, it is true that $h_{N_\varepsilon(x^0)-1} > \varepsilon$, and consequently, evaluating the bound for h_N at $N = N_\varepsilon(x^0) - 1$, we obtain the relation

$$\varepsilon \leq \frac{L_{\nabla f}\operatorname{diam}(\mathcal{X})^2}{(1-\ln(2))(N_\varepsilon(x^0) - T_1(x^0))}.$$

Combining with the estimate (19), and solving the previous relation of $N_\varepsilon(x^0)$ gives us

$$N_\varepsilon(x^0) \leq \left\lceil \frac{1}{\mathtt{a}}\ln\left(\frac{h_0\mathtt{b}}{\mathtt{a}}\right)\right\rceil + \frac{L_{\nabla f}\operatorname{diam}(\mathcal{X})^2}{(1-\ln(2))\varepsilon}. \tag{20}$$

C. Proofs for Variant 2 of Algorithm 2

In this section we describe them main steps in the convergence analysis of Variant 2 of Algorithm 2. In order to ensure that the evaluation of the function $\texttt{step}(f, v, x, g, \mathcal{L})$ needs only finitely many iterations, we need to establish a conceptual global descent lemma. Such a descent property is established in the next Lemma, which corresponds to Lemma 3.2 in the main text.

Lemma C.1. *Assume that $x^k \in \mathcal{S}(x^0)$ for all $k \geq 0$. For all $t \in [0, \gamma_k]$, it holds true that $x^k + t(s^k - x^k) \in \mathcal{S}(x^k)$, and*

$$\|\nabla f(x^k + t(s^k - x^k)) - \nabla f(x^k)\| \leq L_{\nabla f} t \|s^k - x^k\|_2.$$

Proof. The descent property $x^k + t(s^k - x^k) \in \mathcal{S}(x^k)$ for $t \in [0, \gamma_k]$ follows directly from the definition of γ_k. By the mean-value theorem, for all $\sigma > 0$ such that $x^k + t(s^k - x^k) \in \mathcal{S}(x^k)$, we have

$$\begin{aligned}
&\|\nabla f(x^k + t(s^k - x^k)) - \nabla f(x^k)\|_2 \\
&= \|\int_0^t \nabla^2 f(x^k + \tau(s^k - x^k))\, d\tau \cdot (s^k - x^k)\|_2 \\
&\leq \int_0^t \|\nabla^2 f(x^k + \tau(s^k - x^k))(s^k - x^k)\|_2\, d\tau \\
&\leq L_{\nabla f} t \|s^k - x^k\|_2.
\end{aligned}$$

■

This implies a localized version of the descent Lemma, which reads as

$$\begin{aligned}
&f(x^k + t(s^k - x^k)) - f(x^k) \\
&-\langle \nabla f(x^k), t(s^k - x^k)\rangle \leq \tfrac{L_{\nabla f} t^2}{2} \|s^k - x^k\|_2
\end{aligned} \tag{21}$$

for all $t \in [0, \gamma_k]$. Introducing the quadratic model

$$Q(x^k, t, \mu) \triangleq f(x^k) - t\,\mathsf{Gap}(x^k) + \frac{t^2 \mu}{2} \|s(x^k) - x^k\|_2^2, \tag{22}$$

this reads as

$$f(x^k + t(s^k - x^k)) \leq Q(x^k, t, L_{\nabla f}). \tag{23}$$

C.1. Initial parameters

The backtracking subroutine, Algorithm 3, needs to know initial values for the Lipschitz estimate $\mathcal{L}_{-1}$. In Pedregosa et al. (2020), it is recommended to use the following heuristic: Choose $\varepsilon = 10^{-3}$, or any other positive numbers smaller than this. Then set

$$\mathcal{L}_{-1} = \frac{\|\nabla f(x^0) - \nabla f(x^0 + \varepsilon(s^0 - x^0))\|}{\varepsilon \|s^0 - x^0\|}$$

The function $\texttt{step}$ depends on hyperparameters γ_u and γ_d. It is recommended to use $\gamma_d = 0.9$ and $\gamma_u = 2$. This method also needs an initial choice for the Lipschitz parameter μ between $\gamma_d \mathcal{L}_{k-1}$ and $\mathcal{L}_{k-1}$. A choice that is reported to work well is

$$\mu = \mathrm{Clip}_{[\gamma_d \mathcal{L}_{k-1}, \mathcal{L}_{k-1}]} \left(\frac{\mathsf{Gap}(x^k)^2}{2(f(x^k) - f(x^{k-1}))\|s^k - x^k\|_2^2} \right).$$

C.2. Overhead of the backtracking

Evaluation of the sufficient decrease condition in Algorithm 3 requires two extra evaluations of the objective function. If the condition is verified, then it is only evaluated at the current and next iterate. Following Nesterov (2013) we have the following estimate on the number of necessary function evaluations during a single execution of the backtracking procedure.

Proposition C.2. *Let N_k be the number of function evaluations of the sufficient decrease condition up to iteration k. Then*

$$\begin{aligned}
N_k \leq &(k+1)\left(1 - \frac{\ln(\gamma_d)}{\ln(\gamma_u)}\right) \\
&+ \frac{1}{\ln(\gamma_u)} \max\{0, \ln\left(\frac{\gamma_u L_{\nabla f}}{\mathcal{L}_{-1}}\right)\}
\end{aligned}$$

Proof. Call $m_k \geq 0$ the number of function evaluations needed in executing Algorithm 3 at stage k. Since the algorithm multiples the current Lipschitz parameter $\mathcal{L}_{k-1}$ by $\gamma_u > 1$ every time that the sufficient decrease condition is not satisfied, we know that $\mathcal{L}_k \geq \gamma_d \mathcal{L}_{k-1} \gamma_u^{m_k - 1}$. Hence,

$$m_k \leq 1 + \ln\left(\frac{\mathcal{L}_k}{\mathcal{L}_{k-1}}\right) \frac{1}{\ln(\gamma_u)} - \frac{\ln(\gamma_d)}{\ln(\gamma_u)}.$$

Since $N_k = \sum_{i=0}^k m_i$, we conclude

$$N_k \leq (k+1)\left(1 - \frac{\ln(\gamma_d)}{\ln(\gamma_u)}\right) + \frac{1}{\ln(\gamma_u)} \ln\left(\frac{\mathcal{L}_k}{\mathcal{L}_{-1}}\right).$$

By definition of the Lipschitz parameters, we see that $\mathcal{L}_k \leq \max\{\gamma_u L_{\nabla f}, \mathcal{L}_{-1}\}$. Hence, we can bound $\ln\left(\frac{\mathcal{L}_k}{\mathcal{L}_{-1}}\right) \leq \max\{0, \ln\left(\frac{\gamma_u L_{\nabla f}}{\mathcal{L}_{-1}}\right)\}$. ■

Proposition C.2 implies that most of the backtracking subroutines terminate already after a single evaluation of the objective function gradient. Indeed, if we choose hyperparameters as $\gamma_d = 0.9$ and $\gamma_u = 2$, then $1 - \frac{\ln(\gamma_d)}{\ln(\gamma_u)} \leq 1.16$ and so, asymptotically, no more than 16% of the iterates will result in more than one gradient evaluation.

C.3. Proof of Theorem 3.3

The proof of Theorem 3.4 needs the next auxiliary result which we establish first.

Lemma C.3. *We have for all* $t \in [0,1]$

$$f(x^{k+1}) \leq f(x^k) - t\,\mathsf{Gap}(x^k) + \frac{t^2\mathcal{L}_k}{2}\|s^k - x^k\|^2.$$

Proof. Consider the following quadratic optimization problem

$$\min_{t\in[0,1]} \{-t\,\mathsf{Gap}(x^k) + \frac{\mathcal{L}_k t^2}{2}\|s^k - x^k\|^2\}.$$

This has the unique solution

$$\alpha_k = \tau_k(\mathcal{L}_k) = \min\left\{1, \frac{\mathsf{Gap}(x^k)}{\mathcal{L}_k\|s^k - x^k\|^2}\right\}.$$

It therefore follows,

$$\begin{aligned}-\alpha_k\,\mathsf{Gap}(x^k) + \frac{\alpha_k^2\mathcal{L}_k}{2}\|s^k - x^k\|^2 \\ \leq -t\,\mathsf{Gap}(x^k) + \frac{t^2\mathcal{L}_k}{2}\|s^k - x^k\|^2.\end{aligned}$$

By definition of the backtracking procedure, Algorithm 3, we conclude

$$\begin{aligned}f(x^{k+1}) &= f(x^k + \alpha_k(s^k - x^k)) \leq Q(x^k, \alpha_k, \mathcal{L}_k) \\ &= f(x^k) - \alpha_k\,\mathsf{Gap}(x^k) + \frac{\alpha_k^2\mathcal{L}_k}{2}\|s^k - x^k\|^2 \\ &\leq f(x^k) - t\,\mathsf{Gap}(x^k) + \frac{t^2\mathcal{L}_k}{2}\|s^k - x^k\|^2\end{aligned}$$

for all $t \in [0,1]$. ∎

Proof of Theorem 3.3. Define the Fenchel conjugate

$$f^*(u) \triangleq \sup_{z\in\mathrm{dom}(f)} \{\langle z, u\rangle - f(z)\}. \tag{24}$$

Since f is proper, closed and convex, so is the Fenchel conjugate f^*. Moreover, since f is smooth and convex on $\mathrm{dom}\, f$, we know that $f^*(u) = \langle \nabla f(z^*(u)), u\rangle - f(z^*(u))$, where $z^*(u)$ is the unique solution to the equation $\nabla f(z^*(u)) = u$. By definition, we have $f^*(\nabla f(x^k)) \geq \langle \nabla f(x^k), x^k\rangle - f(x^k)$, and by convexity, we know that $\langle \nabla f(x^k), u\rangle - f(u) \leq \langle \nabla f(x^k), x^k\rangle - f(x^k)$ for all $u \in \mathrm{dom}\, f$. We conclude, that

$$f^*(\nabla f(x^k)) \leq \langle \nabla f(x^k), x^k\rangle - f(x^k).$$

Hence, actually equality must hold between both sides, i.e.

$$f^*(\nabla f(x^k)) = \langle \nabla f(x^k), x^k\rangle - f(x^k). \tag{25}$$

Define the support function $H_{\mathcal{X}}(c) \triangleq \sup_{x\in\mathcal{X}}\langle c, x\rangle$, and

$$\psi(z) \triangleq -f^*(z) - H_{\mathcal{X}}(-z) \quad \forall z \in \mathrm{dom}\, f^*. \tag{26}$$

We obtain the following series of equivalences:

$$\begin{aligned}\mathsf{Gap}(x^k) &= \langle \nabla f(x^k), x^k - s^k\rangle \\ &= \langle \nabla f(x^k), x^k\rangle + \langle -\nabla f(x^k), s^k\rangle \\ &= \langle \nabla f(x^k), x^k\rangle + H_{\mathcal{X}}(-\nabla f(x^k)) \\ &= f^*(\nabla f(x^k)) + f(x^k) + H_{\mathcal{X}}(-\nabla f(x^k)) \\ &= f(x^k) - \psi(\nabla f(x^k)).\end{aligned}$$

We note that ψ is concave, and a dual objective function to f. Indeed, by the Fenchel-Young inequality, we know that $f(x) + f^*(y) \geq \langle y, x\rangle$ for all $x \in \mathrm{dom}\, f$ and $y \in \mathrm{dom}\, f^*$. From this inequality, we readily deduce that

$$\min_{x\in\mathcal{X}} f(x) \geq -f^*(y) - H_{\mathcal{X}}(-y).$$

Therefore,

$$f^* \triangleq \min_{x\in\mathcal{X}} f(x) = \max_{y\in\mathrm{dom}\, f^*} \psi(y) \triangleq \psi^*. \tag{27}$$

We know that for all $t \in [0,1]$,

$$f(x^{k+1}) \leq f(x^k) - t\,\mathsf{Gap}(x^k) + \frac{t^2\mathcal{L}_k}{2}\|s^k - x^k\|_2^2$$

Let us introduce the auxiliary sequence $y^0 = \nabla f(x^0)$, and $y^{k+1} = (1-\xi_k)y^k + \xi_k\nabla f(x^k)$ where $\xi_k \triangleq \frac{2}{k+3}$. We observe that

$$\begin{aligned}f(x^k) - \psi(y^k) &= f(x^k) - f^* + \psi^* - \psi(y^k) \\ &\geq f(x^k) - f^*.\end{aligned}$$

Hence, defining the approximation error $h_k \triangleq f(x^k) - f^*$, we see $f(x^k) - \psi(y^k) \geq h_k$ for all $k \geq 0$. Moreover, since ψ is concave, we know that

$$\psi(y^{k+1}) \geq (1-\xi_k)\psi(y^k) + \xi_k\psi(\nabla f(x^k)).$$

Consequently,

$$\begin{aligned}h_{k+1} &\leq f(x^{k+1}) - \psi(y^{k+1}) \\ &\leq f(x^k) - \xi_k\,\mathsf{Gap}(x^k) + \frac{\xi_k^2\mathcal{L}_k}{2}\|s^k - x^k\|_2^2 \\ &\quad - (1-\xi_k)\psi(y^k) - \xi_k\psi(\nabla f(x^k)) \\ &= (1-\xi_k)[f(x^k) - \psi(y^k)] + \frac{\xi_k^2\mathcal{L}_k}{2}\|s^k - x^k\|_2^2 \\ &\leq (1-\xi_k)[f(x^k) - \psi(y^k)] + \frac{\xi_k^2\mathcal{L}_k}{2}\,\mathrm{diam}(\mathcal{X})^2.\end{aligned}$$

Define $A_k \triangleq \frac{1}{2}(k+1)(k+2)$ for $k \geq 0$. For this specification, it is easy to check that

$$A_{k+1}(1-\xi_k) = A_k, \text{ and} \tag{28}$$

$$A_{k+1}\frac{\xi_k^2}{2} \leq 1. \tag{29}$$

Hence,

$$\begin{aligned} A_{k+1}[f(x^{k+1}) - \psi(y^{k+1})] &\leq A_{k+1}(1-\xi_k)[f(x^k) - \psi(y^k)] \\ &\quad + A_{k+1}\frac{\xi_k^2}{2}\mathcal{L}_k \operatorname{diam}(\mathcal{X})^2 \\ &\leq A_k[f(x^k) - \psi(y^k)] \\ &\quad + \mathcal{L}_k \operatorname{diam}(\mathcal{X})^2. \end{aligned}$$

Summing from $i = 0, \ldots, k-1$, and calling

$$\bar{\mathcal{L}}_k \triangleq \frac{1}{k}\sum_{i=0}^{k-1} \mathcal{L}_i,$$

this implies

$$\begin{aligned} f(x^k) - \psi(y^k) &\leq \frac{1}{A_k}[f(x^0) - \psi(y^0)] + \frac{k \operatorname{diam}(\mathcal{X})^2}{2A_k}\bar{\mathcal{L}}_k \\ &= \frac{2}{(k+1)(k+2)}[f(x^0) - \psi(y^0)] + \frac{k \operatorname{diam}(\mathcal{X})^2}{(k+1)(k+2)}\bar{\mathcal{L}}_k \\ &= \frac{2\,\mathsf{Gap}(x^0)}{(k+1)(k+2)} + \frac{k \operatorname{diam}(\mathcal{X})^2}{(k+1)(k+2)}\bar{\mathcal{L}}_k \end{aligned}$$

Since $\mathcal{L}_k \leq L_{\nabla f}$, we get $h_k = \mathcal{O}(k^{-1})$. ■

D. Proof of Theorem 4.2

Let us define $\mathcal{P}(x^0) \triangleq \{x \in \mathcal{X} : f(x) \leq f^* + \mathsf{Gap}(x^0)\}$. We prove this theorem by induction. For $k = 0$, we have the given initial condition $x^0 \in \operatorname{dom} f \cap \mathcal{X}$. Since $x^0 \in \mathcal{P}(x^0)$ trivially, we know from Lemma B.2 that

$$h_0 \geq \frac{\sigma_f}{6}\|x^0 - x^*\|_2^2.$$

Set $r_0 \geq \sqrt{\frac{6h_0}{\sigma_f}}$, this implies $x^* \in B(x^0, r_0)$. Since $s^0 = \mathcal{A}(x^0, r_0, \nabla f(x^0))$, the definition of the Local Linear minimization oracle (LLOO) tells us

$$\begin{aligned} f(x^*) - f(x^0) &\geq \langle \nabla f(x^0), x^* - x^0\rangle \\ &\geq \langle \nabla f(x^0), s^0 - x^0\rangle \geq -\mathsf{Gap}(x^0), \end{aligned}$$

where the first inequality is a consequence of the convexity of f, whereas the last inequality follows by definition of the dual gap function. Therefore, we have $h_0 \leq \mathsf{Gap}(x^0)$. Set $r_0 = \sqrt{\frac{6\,\mathsf{Gap}(x^0)}{\sigma_f}}$, and let $\alpha_0 = \min\left\{\frac{\mathsf{Gap}(x^0)}{\frac{4}{M^2}(\mathsf{e}^0)^2}, 1\right\}\frac{1}{\mathsf{e}^0+1} < 1$. Note that this choice of α_0 guarantees that $\alpha_0\mathsf{e}^0 \leq \frac{\mathsf{e}^0}{\mathsf{e}^0+1} < 1$. Hence, doing the update $x^1 = x^0 + \alpha_0(s^0 - x^0)$ we know from convexity of $\mathcal{X}$ and Lemma B.3 that $x^1 \in \operatorname{dom} f \cap \mathcal{X}$.

Apply inequality (2) to conclude

$$f(x^1) \leq f(x^0) + \alpha_0\langle \nabla f(x^0), s^0 - x^0\rangle + \frac{4}{M^2}\omega_*(\alpha_0\mathsf{e}^0).$$

Since, $\omega_*(t) = -t - \ln(1-t)$, it follows

$$\omega_*(t) = \sum_{j=2}^{\infty}\frac{t^j}{j} \leq \frac{t^2}{2}\sum_{j=0}^{\infty} t^j = \frac{t^2}{2(1-t)} \quad \forall t \in [0,1).$$

Since $\alpha_0\mathsf{e}^0 \in [0,1)$, for all $k \geq 0$, we therefore arrive at the estimate

$$h_1 \leq h_0 + \alpha_0(f(x^*) - f(x^0)) + \frac{4}{M^2}\frac{\alpha_0^2(\mathsf{e}^0)^2}{2(1-\alpha_0\mathsf{e}^0)}.$$

By the above said, we know that $1 - \alpha_0\mathsf{e}^0 \geq \frac{1}{1+\mathsf{e}^0}$, and therefore,

$$\begin{aligned} h_1 &\leq (1-\alpha_0)h_0 + \frac{2\alpha_0^2}{M^2}(\mathsf{e}^0)^2(1+\mathsf{e}^0) \\ &\leq (1-\alpha_0)h_0 + \frac{2\alpha_0^2}{M^2}(\mathsf{e}^0)^2(1+\mathsf{e}^0) \\ &\leq (1-\alpha_0)\,\mathsf{Gap}(x^0) + \frac{2\alpha_0^2}{M^2}(\mathsf{e}^0)^2(1+\mathsf{e}^0) \end{aligned}$$

Plugging in the chosen value of α_0 we obtain that

$$\begin{aligned} h_1 &\leq \mathsf{Gap}(x^0) + \alpha_0(-\mathsf{Gap}(x^0) + \frac{2\alpha_0}{M^2}(\mathsf{e}^0)^2(1+\mathsf{e}^0)) \\ &\leq \mathsf{Gap}(x^0)(1 - \frac{\alpha_0}{2}) \end{aligned}$$

Notice that by the definition of the LLOO we have that

$$\begin{aligned} \mathsf{e}^0 &\leq \sqrt{L_{\nabla f}}\frac{M}{2}\|x^0 - s^0\| \\ &\leq \frac{\sqrt{L_{\nabla f}}M}{2}\min\{\rho r_0, \operatorname{diam}(\mathcal{X})\} \end{aligned}$$

which implies that

$$\frac{\mathsf{Gap}(x^0)}{\frac{4}{M^2}(\mathsf{e}^0)^2} \geq \frac{\mathsf{Gap}(x^0)}{L_{\nabla f}\rho^2 r_0^2} \geq \frac{\sigma_f}{6L_{\nabla f}\rho^2}$$

where the last inequality follows from the definition of r_0.

Thus, we have that

$$\alpha_0 \geq \min\{\frac{\sigma_f}{6L_{\nabla f}\rho^2}, 1\}\frac{1}{1 + \sqrt{L_{\nabla f}}M\operatorname{diam}(\mathcal{X})/2} \equiv \bar{\alpha},$$

which implies that

$$\begin{aligned} h_1 &\leq \mathsf{Gap}(x^0)(1 - \frac{\alpha_0}{2}) \\ &\leq \mathsf{Gap}(x^0)\exp(-\frac{\alpha_0}{2}) \\ &= \mathsf{Gap}(x^0)c_1. \end{aligned}$$

This verifies the claim for $k = 0$. Now proceed inductively. Suppose that

$$h_k \leq \mathsf{Gap}(x^0)c_k,\ c_k \triangleq \exp\left(-\frac{1}{2}\sum_{i=0}^{k-1}\alpha_i\right). \tag{30}$$

Then $x^k \in \mathcal{P}(x^0)$, and Lemma B.2 tells us that

$$\|x^k - x^*\|_2^2 \leq \frac{6h_k}{\sigma_f} \leq \frac{6\,\mathsf{Gap}(x^0)}{\sigma_f} c_k = r_0^2 c_k \equiv r_k^2.$$

Hence, $x^* \in \mathbb{B}(x^k, r_k)$. Proceeding as for $k = 0$, let us again make the educated guess that we can take a step size $\alpha_k < 1/c_k$. By the same argument as before, we obtain sufficient decrease

$$h_{k+1} \leq h_k + \alpha_k \langle \nabla f(x^k), s^k - x^k \rangle + \frac{4}{M^2} \frac{\alpha_k^2 (\mathsf{e}^k)^2}{2(1 - \alpha_k \mathsf{e}^k)}.$$

Since $s^k = \mathcal{A}(x^k, r_k, \nabla f(x^k))$, we know by definition of the LLOO that $\langle \nabla f(x^k), x^* - x^k \rangle \geq \langle \nabla f(x^k), s^k - x^k \rangle$ and $\|s^k - x^k\|_2 \leq \rho r_k$. Consequently, setting $\alpha_k = \min\{\frac{c_k \mathsf{Gap}(x^0)}{\frac{4}{M^2}(\mathsf{e}^k)^2}, 1\} \frac{1}{(1+\mathsf{e}^k)}$ we obtain that

$$\begin{aligned} h_{k+1} &\leq (1 - \alpha_k) h_k + \frac{2}{M^2} \frac{\alpha_k^2 (\mathsf{e}^k)^2}{1 - \alpha_k \mathsf{e}^k} \\ &\leq (1 - \alpha_k)\, \mathsf{Gap}(x^0) c_k + \frac{2}{M^2} \alpha_k^2 (\mathsf{e}^k)^2 (1 + \mathsf{e}^k) \\ &\leq \mathsf{Gap}(x^0) c_k (1 - \frac{\alpha_k}{2}) \end{aligned}$$

where the second inequality follow from the fact that

$$\begin{aligned} \alpha_k \mathsf{e}^k &\leq \frac{\mathsf{e}^k}{1 + \mathsf{e}^k} < 1, \text{ and} \\ 1 - \alpha_k \mathsf{e}^k &\geq \frac{1}{1 + \mathsf{e}^k}. \end{aligned}$$

as well as the upper bound on h_k obtained by the induction step. Finally by the definition of the LLOO we have that

$$\begin{aligned} \mathsf{e}^k &\leq \sqrt{L_{\nabla f}} \frac{M}{2} \|x^k - s^k\| \\ &\leq \frac{\sqrt{L_{\nabla f}} M}{2} \min\{\rho r_k, \operatorname{diam}(\mathcal{X})\}, \end{aligned}$$

which implies that

$$\frac{M^2\, \mathsf{Gap}(x^0) c_k}{4(\mathsf{e}^k)^2} \geq \frac{\mathsf{Gap}(x^0) c_k}{L_{\nabla f} \rho^2 r_k^2} \geq \frac{\sigma_f}{6 L \rho^2}$$

where the last inequality follows from the definition of r_k. Thus, we have that

$$\alpha_k \geq \min\{\frac{\sigma_f}{6 L_{\nabla f} \rho^2}, 1\} \frac{1}{1 + \sqrt{L_{\nabla f}} \frac{M \operatorname{diam}(\mathcal{X})}{2}} \equiv \bar{\alpha},$$

which implies that

$$\begin{aligned} h_{k+1} &\leq \mathsf{Gap}(x^0) c_k (1 - \frac{\alpha_k}{2}) \\ &\leq \mathsf{Gap}(x^0) c_k \exp\left(-\frac{\alpha_k}{2}\right) \\ &= \mathsf{Gap}(x^0) c_{k+1}. \end{aligned}$$

Since $\alpha_k \geq \bar{\alpha}$ for all k, we thus have shown that

$$h_k \leq \mathsf{Gap}(x^0) \exp(-k\bar{\alpha}/2). \tag{31}$$

E. Numerical Experiments

We give extensive information about the numerical experiments we have conducted in order to test the performance of all three algorithms developed in this paper. In the numerical experiments we tested the performance of Variant 1 (V1) and Variant 2 (V2) of Algorithm 2, and compared them with the performance of Frank-Wolfe with standard step-size of $\frac{2}{k+2}$ (Standard), and step-size determined by exact line-search (Line-S.). As a further benchmark, the self-concordant Proximal-Newton (PN) of Tran-Dinh et al. (2015), as implemented in the SCOPT package[1], is included. For the portfolio optimization problem, Algorithm 4 is also implemented. All codes are written in Python 3, with packages for scientific computing NumPy 1.18.1 and SciPy 1.4.1. The experiments were conducted on a PC with Intel Core i5-7500 3.4GHzs, with a total of 16GB RAM.

In both experiments the Frank-Wolfe based methods have been terminated after 50,000 iterations. Because of its higher computational complexity, we decided to stop PN after 1,000 iterations. Each algorithm was terminated early if the optimality gap in a given iteration was lower than $1e-10$. LLOO was only implemented for the portfolio selection problem, using the local linear oracle given in (Garber & Hazan, 2016) for the simplex, as described in Appendix F in the supplementary materials.

E.1. Results on the Portfolio Optimization problem

For the Portfolio Optimization problem we used synthetic data, as in Section 6.4 in (Sun & Tran-Dinh, 2018). The details of the data generating process are as follows. We generate matrix R with given price ratios as: $R := ones(n, p) + N(0, 0.1)$, which allows the closing price to vary about 10% between two consecutive periods. We used different sizes of matrix R: $(n, p) = (1000, 800), (1000, 1200)$, and $(1000, 1500)$ with 4 samples for each size. Hence, there are totally 12 datasets. The detailed results for 9 out of these 12 datasets is reported in Figure E.1.

E.2. Results on the Poisson Inverse problem

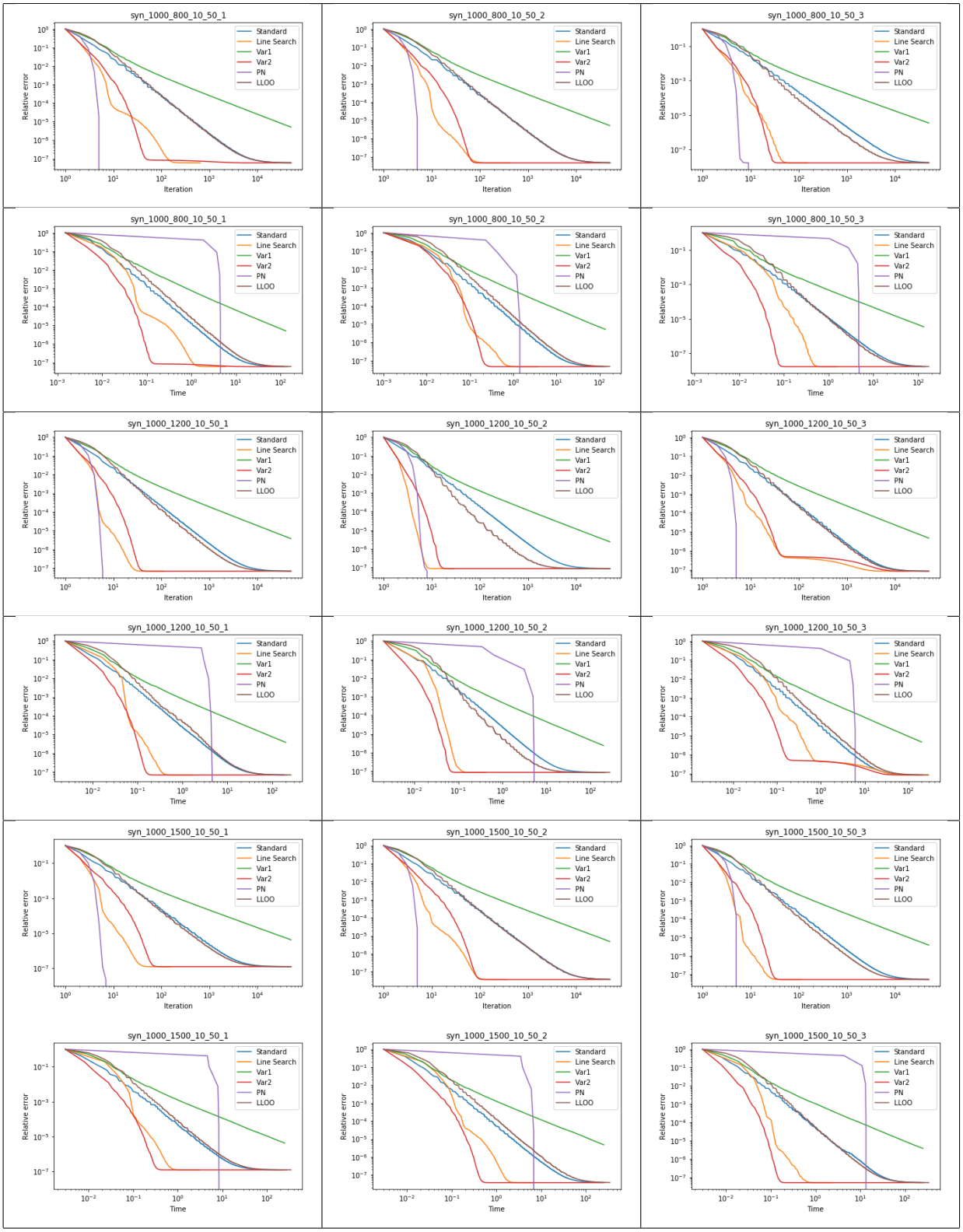
For the Poisson inverse problem we used the datasets a1a-a9a from the LIBSVM library (Chang & Lin, 2011). The results on each individual data set are displayed in Figure E.2.

F. Constructing an LLOO for Simplex Constraints

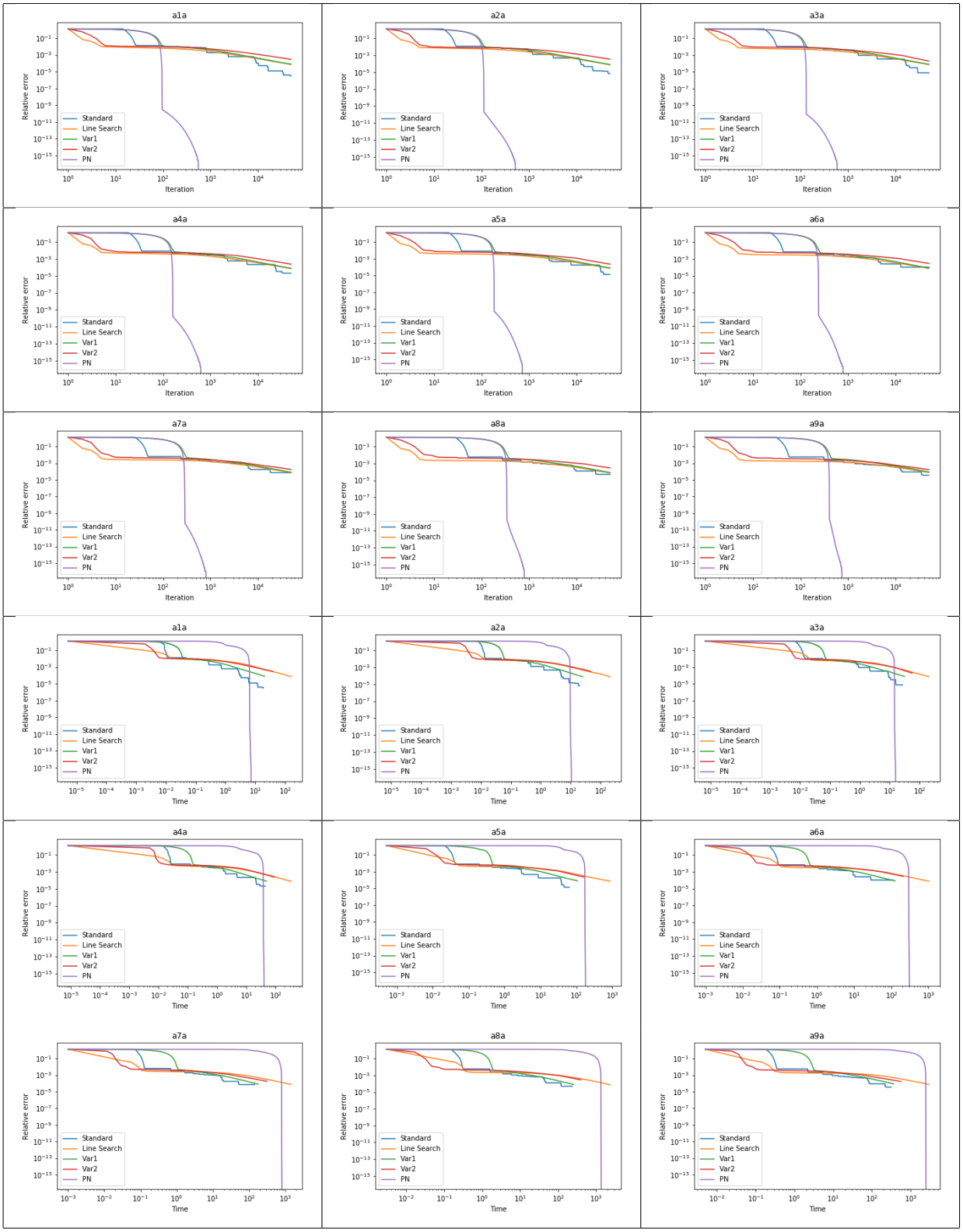
In this section we outline a construction of an LLOO in case where the polytope is the unit simplex $\mathcal{X} \equiv \Delta_n = \{x \in$

[1] https://www.epfl.ch/labs/lions/technology/scopt/

SC analysis of FW algorithms



SC analysis of FW algorithms



$\mathbb{R}^n | x \geq 0, \sum_{i=1}^n x_i = 1\}$. This construction is given in (Garber & Hazan, 2016).

Lemma F.1. *Given a point $x \in \Delta_n$, a radius $r > 0$, and a linear objective $c \in \mathbb{R}^n$, consider the optimization problem*

$$\min\{\langle c, y\rangle | \ \|x - y\|_1 \leq d\} \tag{32}$$

for some $d > 0$. Let us denote by p^* an optimal solution to problem (32) when we set $d = \sqrt{n}r$. Then p^* is the output of an LLOO with parameter $\rho = \sqrt{n}$ for $\mathcal{X} = \Delta_n$. That is, for all $y \in \Delta_n \cap B(x, r)$

$$\langle p^*, c\rangle \leq \langle y, c\rangle \text{ and } \|x - p^*\| \leq \sqrt{n}r. \tag{33}$$

Algorithm 1 implements an LLOO for the unit simplex. In this algorithm we use the Kronecker delta $\delta_{i,j} = 1$ if $j = i$ and 0 otherwise.

Algorithm 1 LLOO for the simplex

Input: $x \in \Delta_n$, radius $r > 0$, cost vector $c \in \mathbb{R}^n$.
Set $d = \sqrt{n}r$
$m = \min\{d/2, 1\}$
$i^* = \operatorname{argmin}_{1 \leq i \leq n} c_i$
$p_+ = m[\delta_{i^*,1}; \ldots; \delta_{i^*,n}]^\top$.
$p_- = \mathbf{0}_n \in \mathbb{R}^n$
Let $i_1, \ldots, i_n$ be a permutation over $\{1, \ldots, n\}$ such that $c_{i_1} \geq c_{i_2} \geq \ldots \geq c_{i_n}$.
Set $k = \min\{\ell | \sum_{j=1}^{\ell} x_{i_j} \geq m\}$
For all $1 \leq j \leq k-1$ set $(p_-)_{i_j} = x_{i_j}$
$p_{i_k} = m - \sum_{j=1}^{k-1} x_{i_j}$
Update $p = x + p_+ - p_-$.

References

Bertsekas, D. *Nonlinear Programming*. Athena Scientific, 1999.

Dvurechensky, P., Staudigl, M., and Uribe, C. A. Generalized self-concordant Hessian-Barrier algorithms. *arXiv preprint arXiv:1911.01522*, 2019.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), 2011.

Garber, D. and Hazan, E. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 2013, 140(1), 125-161.

Nesterov, Y. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer International Publishing, 2018a.

Pedregosa, Fabian and Negiar, Geoffrey and Askari, Armin and Jaggi, Martin Linearly Convergent Frank-Wolfe with Backtracking Line-Search *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Polyak, B. T. *Introduction to Optimization*. Optimization Software, 1987.

Tran-Dinh, Q., Kyrillidis, A., and Cevher, V. Composite self-concordant minimization. *The Journal of Machine Learning Research*, 16(1):371–416, 2015.

Sun, T. and Tran-Dinh, Q. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 2018.