Generalized Self-Concordant Analysis of Frank-Wolfe algorithms*

Pavel Dvurechensky[1], Kamil Safin[2], Shimrit Shtern[3], and Mathias Staudigl[4]

[1]Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany, (Pavel.Dvurechensky@wias-berlin.de)
[2]Moscow Institute of Physics and Technology, Dolgoprudny, Russia, (kamil.safin@phystech.edu)
[3]Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel, (shimrits@technion.ac.il)
[4]Department of Data Science and Knowledge Engineering, Maastricht University, P.O. Box 616, NL–6200 MD Maastricht, The Netherlands, (m.staudigl@maastrichtuniversity.nl)

October 2, 2020

Abstract

Projection-free optimization via different variants of the Frank-Wolfe (FW) method has become one of the cornerstones in large scale optimization for machine learning and computational statistics. Numerous applications within these fields involve the minimization of functions with self-concordance like properties. Such generalized self-concordant (GSC) functions do not necessarily feature a Lipschitz continuous gradient, nor are they strongly convex. Indeed, in a number of applications, e.g. inverse covariance estimation or distance-weighted discrimination problems in support vector machines, the loss is given by a GSC function having unbounded curvature, implying absence of theoretical guarantees for the existing FW methods. This paper closes this apparent gap in the literature by developing provably convergent FW algorithms with standard $O(1/k)$ convergence rate guarantees. If the problem formulation allows the efficient construction of a local linear minimization oracle, we develop a FW method with linear convergence rate.

1 Introduction

Statistical analysis using generalized self-concordant (GSC) functions as a loss function is gaining increasing attention in the machine learning community [**?** **?** **?** **?**]. Beyond machine learning, GSC loss functions are also used in image analysis [**?**] and quantum state tomography [**?**]. This class of loss functions allows to obtain faster statistical rates similar to least-squares [**?**]. At the same time, the minimization of empirical risk in this setting is a challenging optimization problem in high dimensions. Thus, without knowledge of specific structure, interior point, or other polynomial time methods, are unappealing. Moreover, large-scale optimization models in machine learning often depend on noisy data and thus precise high-accuracy solutions are not really needed or obtainable. All these features make simple optimization algorithms with low implementation

*This paper is a significant extension of the conference version [**?**] presented at the 37th International Conference on Machine Learning (ICML2020)

Algorithm 1: `FW-Standard`

Input: $x^0 \in \operatorname{dom} f \cap \mathcal{X}$ initial state; $\varepsilon > 0$ tolerance level
for $k = 1, \ldots$ **do**
  if $\mathsf{Gap}(x^k) > \varepsilon$ **then**
    Obtain $s^k = s(x^k)$
    Set $x^{k+1} = x^k + \alpha_k(s^k - x^k)$ for some $\alpha_k \in [0, 1]$.
  end if
end for

costs the preferred methods of choice. In this paper we focus on projection-free methods which rely on the availability of a Linear Minimization Oracle (LMO). Such algorithms are known as Conditional Gradient (CG) or Frank-Wolfe (FW) methods, and can be traced back to [**?** **?**]. For a given convex compact set $\mathcal{X} \subset \mathbb{R}^n$, and a convex objective function f, the FW method aims to solve the smooth convex optimization problem

$$\min_{x \in \mathcal{X}} f(x), \tag{P}$$

by sequential calls of a LMO returning at point x the target vector

$$s(x) \in \arg\min_{d \in \mathcal{X}} \langle \nabla f(x), d \rangle. \tag{1.1}$$

The selection $s(x)$ is determined via some pre-defined tie-breaking rule, whose specific form is of no importance for our work. Computing this target state is the only computational bottleneck of the method. FW is therefore tailored to domains $\mathcal{X}$ over which one can easily minimize linear functions, but where orthogonal projections would be hard to compute. Progress of the algorithm is monitored via a merit function. The standard merit function in optimization problems with a convex structure is the dual gap function

$$\mathsf{Gap}(x) \triangleq \max_{s \in \mathcal{X}} \langle \nabla f(x), x - s \rangle, \tag{1.2}$$

known also as the *Frank-Wolfe gap* in this context. It is easy to see that $\mathsf{Gap}(x) \geq 0$ for all $x \in \mathcal{X}$, with equality if and only if x is a solution to (**??**). Since f is convex on $\mathcal{X}$, it is easy to see that any point in $(\mathsf{Gap})^{-1}(0)$ is an exact solution (global minimum) to problem (**??**). The vanilla FW, whose pseudo-code is given in Algorithm **??**, reduces the dual gap function by sequentially solving linear minimization subproblems to obtain the target points $s(x)$.

As usual, the performance of a method depends heavily on the design of efficient step-size policies. Two popular options for selecting α_k are either to set $\alpha_k = \frac{2}{k+2}$, or else to compute α_k via an exact line-search

$$\alpha_k = \operatorname*{argmin}_{t \in [0,1]} f((1-t)x^k + ts^k). \tag{1.3}$$

Under either of these step-size policies, it is well known that Algorithm **??** exhibits an $O(1/k)$ rate of convergence for solving (**??**) in case where f is convex and either possess a Lipschitz continuous gradient, or a bounded curvature constant. The latter is a slight weakening of the classical Lipschitz gradient assumption, and is the key quantity in the modern analysis of FW algorithms initiated in [**?**]. Indeed, [**?**] defines the *curvature constant*

$$\kappa_f \triangleq \sup_{x,s \in \mathcal{X}, \gamma \in [0,1]} \frac{2}{\gamma^2} D_f(x + \gamma(s - x), x), \tag{1.4}$$

involving the Bregman function of the objective function f (cf. eq. (**??**)). Assuming that $\kappa_f < \infty$, [**?**] derived a $O(1)\frac{\kappa_f \operatorname{diam}(\mathcal{X})}{\varepsilon}$ sublinear rate of convergence to reach an ε-optimal solution. This iteration complexity result was shown by [**?**] to be optimal, even when f is strongly convex. This is quite surprising, since gradient methods are known to display linear convergence on strongly convex minimization problems [**?**]. Departing from here, recent research has focused on improving the convergence guarantees provided by the vanilla FW.

Linearly convergent FW methods [**?**] obtained linear convergence rates in well conditioned problems under the a-priori assumption that the solution lies in the relative interior of the feasible set, and the rate of convergence explicitly depends on the distance of the solution from the boundary (see also [**?** **?**]). If no a-priori information on the location of the solution is available, there are essentially two known twists of the vanilla FW to boost the convergence rates. One twist is to modify the search directions via *corrective* or *away* search directions [**?** **?** **?** **?** **?**]. These approaches however require more fine grained oracles, which are in general not available for GSC functions. First, Away-step FW needs a vertex oracle. However, vertices need not be in the domain of a GSC function (i.e. the origin in Example **??**). Even worse, [**?**] show that Away-step FW can even deteriorate the convergence properties of the method when minimizing GSC functions. Second, Away-steps and Pairwise FW have been often considered as impractical, since their implementation requires an exact line search. [**?**] resolved this by developing backtracking variants of Away-step and Pairwise-FW. These ideas are crucial for our development of the here proposed backtracking FW-variant (Algorithm **??**, `Backtrack-GSC`). The alternative twist is to change the design of the LMO [**?** **?** **?**]. In particular, the work [**?**] has been fundamental to the construction of our linearly convergent variant (Algorithm **??**, `FW-LLOO`).

FW for ill-conditioned functions FW methods are very often applied to smooth minimization problems which are *well-conditioned*: The function f is strongly convex over the feasible set, and its gradient is Lipschitz continuous. In this paper we are interested in functions which are possibly *ill-conditioned*: f is neither assumed to be globally strongly convex, nor to posses a Lipschitz continuous gradient over the feasible set. The development of FW-methods for such ill-conditioned problems has received quite some attention recently. [**?**] requires the gradient of the objective function to be Hölder continuous. Implicitly it is assumed that $\mathcal{X} \subseteq \operatorname{dom} f$, an assumption we do not make, and is also not satisfied in important applications (e.g. $0 \in \mathcal{X}$, but $0 \notin \operatorname{dom} f$ in the Covariance Estimation problem in Section **??**). Specialized to solving a quadratic Poisson inverse problem in phase retrieval, [**?**] provided a globally convergent FW method using the convex reformulation based on the PhaseLift approach [**?**]. They constructed a provably convergent FW variant using a new step size policy derived from estimate sequence techniques [**?** **?**] in order to match the proof technique of [**?**]. In this paper we develop a unified approach for FW-methods for minimizing generalized self-concordant functions - a class of functions including the convex reformulation of [**?**].

The main difficulties one faces in minimizing functions with self-concordance like properties can be easily illustrated with a basic, in some sense minimal, example:

Example 1.1. Consider the function $f(x, y) = -\ln(x) - \ln(y)$ where $x, y > 0$ satisfy $x + y = 1$. This function is the standard self-concordant barrier for the positive orthant (the log-barrier) and thus $(2, 3)$-generalized self-concordant (see Definition **??**). Its Bregman divergence is easily calculated as

$$D_f(u, v) = \sum_{i=1}^{2} \left[-\ln\left(\frac{u_i}{v_i}\right) + \frac{u_i}{v_i} - 1 \right] \quad u = (u_1, u_2), v = (v_1, v_2).$$

Neither the function f nor its gradient is Lipschitz continuous over the set of interest. In particular the curvature constant is unbounded, i.e $\kappa_f = \infty$. Moreover, if we start from $u^0 = (1/4, 3/4)$ and apply the standard $2/(k+2)$-step size policy, then $\alpha_0 = 1$, which leads to $u^1 = s(u^0) = (1, 0) \notin \operatorname{dom} f$. Clearly, the standard method fails.

The logarithm is one of the canonical members of (generalized) self-concordant functions, and thus the above example is quite representative for the class of optimization problems of interest in this paper. It is therefore clear that the standard analysis of [**?**], and all subsequent investigations relying on estimates of the Lipschitz constant of the gradient or the curvature, cannot be applied straightforwardly to the problem of minimizing a GSC function via Frank-Wolfe methods. The class of GSC functions is a significant extension of the classical self-concordant (SC) functions, well known from the theory of interior point methods [**?**]. Motivated by the new analysis of the logistic regression problem in [**?**], GSC functions have been characterized in [**?**], and keep on receiving a lot of attention in statistical learning and data science recently [**?** **?** **?**], because of their universal appearance as loss functions in generalized linear models. Besides applications in statistics, generalized self-concordant functions are of some importance in scientific computing. [**?**] construct self-concordant barriers for a class of polytopes arising naturally in combinatorial optimization. [**?**] show that the well-known matrix balancing problem minimizes a GSC function. We believe that our results are going to be useful in such problems as well.

1.1 Main Contributions and outline of the paper

In this paper we demonstrate that FW indeed works when minimizing a GSC function over a compact convex set. Section **??** contains necessary definitions and properties for the class of GSC functions in a self-contained way. Section **??** constructs new adaptive step-size policies ensuring global convergence and standard $O(1/k)$ sublinear convergence rates. The derivation of these step-size schemes fully relies on basic properties of GSC functions, and their analysis provides some hints on how to achieve acceleration of standard FW. Leveraging upon the Local Linear Optimization Oracle (LLOO) constructed in [**?**], we establish in Section **??** the first linearly convergent FW method for minimizing GSC functions. In a previous conference version [**?**] we considered only SC functions, which are a subclass of GSC functions. Section **??** reports results from extensive numerical experiments of the proposed algorithms and their comparison with the baselines.

Notation Given a proper, closed, and convex function $f : \mathbb{R}^n \to (-\infty, \infty]$, we denote by $\operatorname{dom} f \triangleq \{x \in \mathbb{R}^n | f(x) < \infty\}$ the (effective) domain of f. For a set X, we define the indicator function $\delta_X(x) = \infty$ if $x \notin X$, and $\delta_X(x) = 0$ otherwise. We use $\mathbf{C}^k(\operatorname{dom} f)$ to denote the class of functions $f : \mathbb{R}^n \to (-\infty, \infty]$ which are k-times continuously differentiable on their effective domain. We denote by ∇f the gradient map, and $\nabla^2 f$ the Hessian map.

Let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of nonnegative, and positive real numbers, respectively. We use $\mathbb{S}^n \triangleq \{x \in \mathbb{R}^{n\times n} | x^\top = x\}$ the set of symmetric matrices, and $\mathbb{S}^n_+, \mathbb{S}^n_{++}$ to denote the set of symmetric positive semi-definite and positive definite matrices, respectively. Given $Q \in \mathbb{S}^n_{++}$ we define the weighted inner product $\langle u, v\rangle_Q \triangleq \langle Qu, v\rangle$ for $u, v \in \mathbb{R}^n$, and the corresponding norm $\|u\|_Q \triangleq \sqrt{\langle u, u\rangle_Q}$. The associated dual norm is $\|v\|^*_Q \triangleq \sqrt{\langle v, v\rangle_{Q^{-1}}}$.

For $Q \in \mathbb{S}^n$, we let $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$ denote the smallest and largest, respectively, eigenvalue of the matrix Q.

2 Generalized self-concordant minimization

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a three-times continuously differentiable function on $\operatorname{dom}\varphi$. Recall that φ is convex if and only if $\varphi''(t) \geq 0$ for all $t \in \operatorname{dom}\varphi$.

Definition 2.1 ([**?**])**.** Let $\varphi \in \mathbf{C}^3(\operatorname{dom}\phi)$ be a convex function with $\operatorname{dom}\varphi$ open. Given $\nu > 0$ and $M_\varphi > 0$ some constants, we call φ (M_φ, ν) *generalized self-concordant* (GSC) if

$$|\varphi'''(t)| \leq M_\phi \varphi''(t)^{\frac{\nu}{2}} \qquad \forall t \in \operatorname{dom}\varphi. \tag{2.1}$$

If $\varphi(t) = \frac{a}{2}t^2 + bt + c$ for any constant $a \geq 0$ we get a $(0, \nu)$-generalized self-concordant function. Hence, any convex quadratic function is GSC for any $\nu > 0$. Standard one-dimensional examples are summarized in Table **??**.

Function name	Form of $\varphi(t)$	ν	M_φ	$\operatorname{dom}\varphi$	Lipschitz smooth
Burg entropy	$-\ln(t)$	3	2	$(0,\infty)$	No
Boltzmann-Shannon entropy	$t\ln(t) + \delta_{(0,\infty)}(t)$	4	1	$(0,\infty)$	No
Logistic	$\ln(1+e^{-t})$	2	1	$(-\infty,\infty)$	Yes
Exponential	e^{-t}	2	1	$(-\infty,\infty)$	Yes
Negative Power	$t^{-q}, q > 0$	$\frac{2(q+3)}{q+2}$	$\frac{q+2}{\sqrt[q+2]{q(q+1)}}$	$(0,\infty)$	No
Positive Power	$t^q + \delta_{(0,\infty)}(t), q \in (1,2)$	$\frac{2(3-q)}{2-q}$	$\frac{2-q}{\sqrt[2-q]{q(q-1)}}$	$(0,\infty)$	No
Arcsine distribution	$\frac{1}{\sqrt{1-t^2}}$	$\frac{14}{5}$	< 3.25	$(-1,1)$	No

Table 1: Examples of univariate GSC functions.

This definition generalizes to multivariate functions by requiring GSC along every straight line. Specifically, let $f : \mathbb{R}^n \to \mathbb{R}$ be a closed convex, lower semi-continuous function with effective domain $\operatorname{dom} f$ which is an open nonempty subset of $\mathbb{R}^n$. For $x \in \operatorname{dom} f$ and $u, v \in \mathbb{R}^n$, define the real-valued function $\varphi(t) := \langle \nabla^2 f(x+tv)u, u\rangle$. For $t \in \operatorname{dom}\varphi$, one sees that $\phi'(t) = \langle D^3 f(x+tv)[v]u, u\rangle$, where $D^3 f(x)[v]$ denotes the third-derivative tensor at (x, v), viewed as a bilinear mapping $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

Definition 2.2 ([**?**])**.** A closed convex function $f \in \mathbf{C}^3(\operatorname{dom} f)$, with $\operatorname{dom} f$ open, is called (M_f, ν) generalized self-concordant of the order $\nu > 0$ and constant $M_f \geq 0$ if for all $x \in \operatorname{dom} f$

$$|\langle D^3 f(x)[v]u, u\rangle| \leq M_f \|u\|_x^2 \|v\|_x^{\nu-2} \|v\|_2^{3-\nu} \qquad \forall u, v \in \mathbb{R}^n. \tag{2.2}$$

We denote this class of functions as $\mathcal{F}_{M_f,\nu}$.

In the extreme case $\nu = 2$ we recover the definition $|\langle D^3 f(x)[v]u, u\rangle| \leq M_f \|u\|_x^2 \|v\|_2$, which is the generalized self-concordance definition in [**?**]. If $\nu = 3$ and $u = v$ the definition becomes $|\langle D^3 f(x)[u]u, u\rangle| \leq M_f \|u\|_x^3$, which is the standard self-concordance definition due to [**?**]. Extending classical results on self-concordant function, it is easy to see that generalized self-concordance is invariant under affine transformations and re-parametrizations of the domain.[1] Affine invariance

[1]Section **??** in Appendix **??** collects relevant properties of GSC functions.

allows us to cover a very broad class of composite convex optimization of the form

$$\min_{x\in\mathcal{X}}\{f(x) \triangleq g(\mathbf{E}x) + \langle q, x\rangle\}. \tag{2.3}$$

We assume that $g : \mathbb{R}^m \to (-\infty, \infty]$ belongs to the class $\mathcal{F}_{M_g,\nu}$, $\mathbf{E} \in \mathbb{R}^{m\times n}$ is a given matrix and $q \in \mathbb{R}^n$ a given vector. This formulation covers empirical risk-minimization with a norm-regularization, among many others.

Example 2.1. To illustrate this claim, consider the regularized optimization problem $f(x) = \phi(x) + \lambda\|x\|_1$, where ϕ is a generalized self-concordant loss function, arising in generalized linear models (see, e.g., [**?** **?**]), or in the distance-weighted discrimination problem in support vector machines (see Section **??**). Minimizing this function f over a convex compact set $\mathcal{X}$, means that there exists a constant $R > 0$ such that $\|x\|_1 \le R$ for all $x \in \mathcal{X}$. Therefore, we can reformulate the composite minimization problem (**??**) as the generalized self-concordant minimization problem

$$\min\{\phi(x) + \lambda y | x \in \mathcal{X}, \|x\|_1 \le R, y \in [0, R]\}.$$

For $\lambda > 0$ fixed, the function $(x, y) \mapsto \phi(x) + \lambda y$ is clearly GSC as well.

2.1 Important Estimates

The Hessian of a function $f \in \mathcal{F}_{M_f,\nu}$ defines a semi-norm $\|u\|_x \triangleq \sqrt{\langle u, u\rangle_{\nabla^2 f(x)}}$ for all $x \in \operatorname{dom} f$, with dual norm $\|d\|_x^* \triangleq \sup_{d\in\mathbb{R}^n}\{2\langle d, a\rangle - \|d\|_x^2\}$. Note that if $\nabla^2 f(x) \in \mathbb{S}^n_{++}$ then $\|\cdot\|_x$ is a true norm, and $\|d\|_x^* = \sqrt{\langle d, d\rangle_{[\nabla^2 f(x)]^{-1}}}$. The local norm is an important ingredient in our development, since it gives us a tool to measure the distance from the boundary of $\operatorname{dom}(f)$. Given $\nu \ge 2$ and $f \in \mathcal{F}_{M_f,\nu}$, we define the distance-like function

$$\mathsf{d}_\nu(x, y) \triangleq \begin{cases} M_f\|y - x\|_2 & \text{if } \nu = 2, \\ \frac{\nu-2}{2}M_f\|y - x\|_2^{3-\nu} \cdot \|y - x\|_x^{\nu-2} & \text{if } \nu > 2. \end{cases} \tag{2.4}$$

Define the *Dikin Ellipsoid*

$$\mathcal{W}(x; r) \triangleq \{y \in \mathbb{R}^n : \mathsf{d}_\nu(x, y) < r\} \quad \forall (x, r) \in \operatorname{dom} f \times \mathbb{R}. \tag{2.5}$$

Lemma 2.3 ([**?**], Prop. 7)**.** *Let $f \in \mathcal{F}_{M_f,\nu}$ with $\nu > 2$. We have $\mathcal{W}(x; 1) \subset \operatorname{dom} f$ for all $x \in \operatorname{dom} f$.*

The inclusion $\mathcal{W}(x; 1) \subset \operatorname{dom} f$ is a generalization of a well-known classical property of SC functions [**?**]. It generalizes only if $\nu > 2$. This is intuitive, since for $\nu = 2$, the distance function d_2 effectively boils down to the euclidean distance, and thus is not adaptive to the local geometry. Our algorithmic scheme will take as inputs functions $f \in \mathcal{F}_{M_f,\nu}$ with $\nu \ge 2$. This covers the important case of standard self-concordant functions ($\nu = 3$), as well as exponential and power functions featuring GSC parameters $\nu \in (2, 3)$ (cf. Table **??**). However, our method also works well for generalized self-concordant function of order $\nu > 3$. This complete range of parameters cannot be analyzed by the Newton method developed in [**?**].

We define the *Bregman divergence* associated to $f \in \mathcal{F}_{M_f,\nu}$ as

$$D_f(x, y) \triangleq f(x) - f(y) - \langle \nabla f(y), x - y\rangle \quad \text{for } x, y \in \operatorname{dom} f. \tag{2.6}$$

Since this divergence will be a crucial quantity of interest in measuring the per-iteration progress of our method, it is instrumental to have bounds on the function values.

Lemma 2.4 ([**?**], Prop. 10)**.** *Let $x \in \operatorname{dom} f$ for $f \in \mathcal{F}_{M_f,\nu}$ and $\nu \geq 2$. Then*

$$\omega_\nu(-\mathsf{d}_\nu(x,y))\|y-x\|_x^2 \leq D_f(y,x) \leq \omega_\nu(\mathsf{d}_\nu(x,y))\|y-x\|_x^2, \tag{2.7}$$

where, if $\nu > 2$, the right-hand side inequality of (**??**) *holds if and only if $\mathsf{d}_\nu(x,y) < 1$. $\omega_\nu(\cdot)$ is defined as*

$$\omega_\nu(t) \triangleq \begin{cases} \frac{1}{t^2}(e^t - t - 1) & \text{if } \nu = 2, \\ \frac{-t-\ln(1-t)}{t^2} & \text{if } \nu = 3, \\ \frac{(1-t)\ln(1-t)+t}{t^2} & \text{if } \nu = 4, \\ \left(\frac{\nu-2}{4-\nu}\right)\frac{1}{t}\left[\frac{\nu-2}{2(3-\nu)t}((1-t)^{\frac{2(3-\nu)}{2-\nu}} - 1) - 1\right] & \text{otherwise.} \end{cases} \tag{2.8}$$

The function $\omega_\nu(\cdot)$ is strictly convex and one can check that $\omega_\nu(t) \geq 0$ for all $t \in \operatorname{dom}(\omega_\nu)$.

3 FW works for generalized self-concordant functions

In this section we describe two provably convergent modifications of the vanilla FW scheme (Algorithm **??**, `FW-standard`) displaying sublinear convergence rates.

3.1 Preparations

Throughout the paper, we assume the following to be true:

Assumption 1. The solution set $\mathcal{X}^*$ of (**??**) is nonempty. The function f in (**??**) belongs to the class $\mathcal{F}_{M_f,\nu}$ with known parameters M_f and $\nu \in [2,4]$. $\mathcal{X}$ is convex compact and the LMO search direction (**??**) can be computed efficiently and accurately.

For $x \in \operatorname{dom} f$, define the target vector $s(x)$ as in (**??**), and the Frank-Wolfe gap $\mathsf{Gap}(x)$ as in (**??**). Moreover, let us define

$$\mathsf{e}(x) \triangleq \|s(x)-x\|_x \text{ and } \beta(x) \triangleq \|s(x)-x\|_2 \quad \forall x \in \operatorname{dom} f \tag{3.1}$$

We can unravel (**??**) to get the two bounds:

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \omega_\nu(-\mathsf{d}_\nu(x,y))\|y-x\|_x^2, \text{ and} \tag{3.2}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \omega_\nu(\mathsf{d}_\nu(x,y))\|y-x\|_x^2. \tag{3.3}$$

If $x = x^* \in \mathcal{X}^* \subset \operatorname{dom} f \cap \mathcal{X}$ is a solution to (**??**), the necessary and sufficient optimality condition reads as

$$\langle \nabla f(x^*), x - x^*\rangle \geq 0 \qquad \forall x \in \mathcal{X}. \tag{3.4}$$

3.2 A Frank-Wolfe method with analytical step-size

Our first FW method (Algorithm **??**, `FW-GSC`) for minimizing GSC functions employs a novel adaptive step-size rule, derived from estimates provided by general properties of GSC functions. An attractive feature of this new step size policy is that it is available in analytical form, making it potentially very attractive for numerical optimization. Indeed, the analytical step size rule allows us to do away with any globalization strategy (e.g. line search). This has significant practical impact when the evaluation of the function is expensive.

Algorithm 2: `FW-GSC`

Input: $x^0 \in \operatorname{dom} f \cap \mathcal{X}$ initial state, $\varepsilon > 0$ error tolerance, and $f \in \mathcal{F}_{M,\nu}(\operatorname{dom} f)$.
for $k = 0, \ldots$ **do**
  if $\mathsf{Gap}(x^k) > \varepsilon$ **then**
    Obtain $s^k = s(x^k)$ from (**??**)
    Obtain $\alpha_k = \alpha_\nu(x^k)$ from (**??**)
    Set $x^{k+1} = x^k + \alpha_k(s^k - x^k)$
  end if
end for

Given $x \in \mathcal{X}$, set $x_t^+ \triangleq x + t(s(x) - x)$, and assume that $\mathsf{e}(x) \neq 0$. Moving from the current position x to the point x_t^+, we know that $\mathsf{d}_\nu(x, x_t^+) = t\delta_\nu(x)$, where

$$\delta_\nu(x) \triangleq \begin{cases} M_f\beta(x) & \text{if } \nu = 2, \\ \frac{\nu-2}{2} M_f\beta(x)^{3-\nu}\mathsf{e}(x)^{\nu-2} & \text{if } \nu > 2. \end{cases} \tag{3.5}$$

Choosing $t \in (0, 1/\delta_\nu(x))$, we conclude from eq. (**??**)

$$\begin{aligned} f(x_t^+) &\leq f(x) + \langle \nabla f(x), x_t^+ - x\rangle + \omega_\nu(\mathsf{d}_\nu(x, x_t^+))\|x_t^+ - x\|_x^2 \\ &\leq f(x) + \langle \nabla f(x), x_t^+ - x\rangle + \omega_\nu(t\delta_\nu(x))t^2\mathsf{e}(x)^2 \\ &\leq f(x) - t\,\mathsf{Gap}(x) + \omega_\nu(t\delta_\nu(x))t^2\mathsf{e}(x)^2 \end{aligned}$$

For $x \in \operatorname{dom} f \cap \mathcal{X}$, define $\eta_{x,\nu} : \mathbb{R}_+ \to \mathbb{R} \cup \{-\infty, \infty\}$ by

$$\eta_{x,\nu}(t) \triangleq \mathsf{Gap}(x)\left[t - \omega_\nu(t\delta_\nu(x))t^2\frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}\right]. \tag{3.6}$$

Note that $\eta_{x,\nu}(t)$ is strictly concave on $\operatorname{dom}(\eta_{x,\nu}) \subseteq [0, 1/\delta_\nu(x)]$. This leads to the per-iteration change in the objective function value as

$$f(x_t^+) - f(x) \leq -\eta_{x,\nu}(t) \qquad \forall t \in (0, 1/\delta_\nu(x)).$$

Since $\eta_{x,\nu}(t) > 0$ for $t \in (0, 1/\delta_\nu(x))$, we are ensured that we make progress in reducing the objective function value when choosing a step size within the indicated range. To optimize this per-iteration decrease, we search for a value t such that the per-iteration decrease is as big as possible. Hence, we aim to find t which solves the concave maximization problem

$$\sup_{t \geq 0} \eta_{x,\nu}(t). \tag{3.7}$$

Call $\mathtt{t}_\nu(x)$ a solution of this program. Since we have to stay within the feasible set, we cannot simply use the number $\mathtt{t}_\nu(x)$ as our step size as it might lead to an infeasible point. Therefore, we propose the truncated step-size

$$\alpha_\nu(x) \triangleq \min\{1, \mathtt{t}_\nu(x)\} \quad \forall x \in \operatorname{dom} f, \tag{3.8}$$

and make the following observation, depending on the GSC-parameter $\nu \geq 2$:

(a) If $\nu > 2$ we use the Dikin ellipsoid condition $\delta_\nu(x)\alpha_\nu(x) < 1$, by requiring that $\mathtt{t}_\nu(x) < 1/\delta_\nu(x)$. Surprisingly, it turns out that this is always the case for $\nu \in (2, 4]$.

(b) If $\nu = 2$, we know that $\mathrm{dom}(f) = \mathbb{R}^n$, and consequently it suffices to guarantee $\alpha_2(x) \leq 1$.

We emphasize that the basic step-size rule is derived by identifying a suitable local majorizing model $f(x) - \eta_{x,\nu}(t)$. Minimization with respect to t aligns the model as close as possible to the effective change in the objective function value. Thus, the majorizing model can be seen as a worst-case model for the objective function, as it ignores all fine details of the true objective function and uses only growth estimates of GSC functions. Therefore, similar to [**?** **?**], the derived adaptive step size policy can be regarded as an optimal choice in the analytic worst-case sense.

3.3 A Backtracking FW variant

`FW-GSC` comes with several drawbacks: First, it relies on the minimization of a majorizing worst-case model. This overestimation strategy leads to a worst-case performance estimate, relying on various state-dependent quantities, such as the local norm $\mathsf{e}(x^k)$. Evaluating these objects requires the evaluation of the matrix-vector product between the Hessian $\nabla^2 f(x^k)$, and the FW search direction $s(x^k) - x^k$.[2] In order to circumvent these potentially computationally intensive tasks, we develop in this section a backtracking variant of `FW-GSC`, in the spirit of [**?**]. We also note that, such variants of accelerated gradient methods are known already for a long time [**?** **?** **?** **?**].

Consider the quadratic model

$$Q(x, t, \mu) \triangleq f(x) - t\,\mathsf{Gap}(x) + \frac{t^2\mu}{2}\|s(x) - x\|_2^2 = f(x) - t\,\mathsf{Gap}(x) + \frac{t^2\mu}{2}\beta(x)^2, \tag{3.9}$$

where $x \in \mathcal{X}$ is the current position of the algorithm, and t, μ are positive parameters. From the complexity analysis of `FW-GSC` we know that there exists a range of step-size parameters $t > 0$ that guarantee decrease in the objective function value. Denote by $\mathcal{S}(x) \triangleq \{x' \in \mathcal{X} | f(x') \leq f(x)\}$, and set $\gamma_k \triangleq \sup\{t > 0 | x^k + t(s^k - x^k) \in \mathcal{S}(x^k)\}$ as well as $L_k \triangleq \max_{x \in \mathcal{S}(x^k)} \nabla^2 f(x^k)$. Then, for all $t \in [0, \gamma_k]$, it holds true that $f(x^k + t(s^k - x^k)) \leq f(x^k)$. Therefore, by the mean-value-theorem

$$\|\nabla f(x^k + t(s^k - x^k)) - \nabla f(x^k)\| \leq L_k t \|s^k - x^k\|_2 \quad \forall t \in (0, \gamma_k).$$

Hence, we get the sufficient decrease inequality

$$f(x^k + t(s^k - x^k)) - f(x^k) \leq -t\,\mathsf{Gap}(x^k) + \frac{L_k t^2}{2}\|s^k - x^k\|_2^2 = -t\,\mathsf{Gap}(x^k) + \frac{L_k t^2}{2}\beta(x^k)^2,$$

which reads in terms of the quadratic model (**??**)

$$f(x^k + t(s^k - x^k) \leq Q(x^k, t, L_k) \qquad \forall t \in (0, \gamma_k). \tag{3.10}$$

The idea behind the backtracking procedure is to dispense with the computation of the local Lipschitz estimate L_k, and replace it with the backtracking procedure $\mathtt{step}(f, v^k, x^k, \mathsf{Gap}(x^k), \mathcal{L}_{k-1})$ (Algorithm **??**) as an inner-loop within Algorithm **??** (`Backtrack-FW`). In particular, the implementation of `Backtrack-FW` does not require the evaluation of the Hessian matrix $\nabla^2 f(x^k)$.

[2] In fact, evaluating the local norm requires the Hessian matrix $\nabla^2 f(x)$, and thus Algorithm `FW-GSC` is actually second-order method.

Algorithm 3: `BackTrackFW-GSC`

Input: $x^0 \in \operatorname{dom} f \cap \mathcal{X}$ initial state, $f \in \mathcal{F}_{M,\nu}(\operatorname{dom} f)$
for $k = 1, \ldots$ **do**
  if $\mathsf{Gap}(x^k) > \varepsilon$ **then**
    Obtain $s^k = s(x^k)$ and set $v^k = s^k - x^k$
    Set $(\alpha_k, \mathcal{L}_k) = \texttt{step}(f, v^k, x^k, \mathsf{Gap}(x^k), \mathcal{L}_{k-1})$
    Set $x^{k+1} = x^k + \alpha_k(s^k - x^k)$
  end if
end for

Algorithm 4: Function $\texttt{step}(f, v, x, g, \mathcal{L})$

Choose $\gamma_u > 1, \gamma_d < 1$
Choose $\mu \in [\gamma_d \mathcal{L}, \mathcal{L}]$
$\alpha = \min\{\frac{g}{\mu \|v\|_2^2}, 1\}$
if $f(x + \alpha v) > Q(x, \alpha, \mu)$ **then**
  $\mu \leftarrow \gamma_u \mu$
  $\alpha \leftarrow \min\{\frac{g}{\mu \|v\|_2^2}, 1\}$
end if
Return α, μ

4 Complexity analysis

4.1 Complexity Analysis of `FW-GSC`

Based on the preliminary analysis of Section **??**, our strategy to determine the step-size policy is to first compute $\mathtt{t}_\nu(x)$, and then clip the value accordingly. A technical analysis of the optimization problem (**??**), delegated to Appendix **??**, yields the following explicit values for $\mathtt{t}_\nu(x)$.

Theorem 4.1. *Given $f \in \mathcal{F}_{M_f,\nu}$. Then the unique solution to program* (**??**) *is given by*

$$\mathtt{t}_\nu(x) = \begin{cases} \frac{1}{\delta_2(x)} \ln\left(1 + \frac{\mathsf{Gap}(x)\delta_2(x)}{\mathsf{e}(x)^2}\right) & \text{if } \nu = 2, \\ \frac{1}{\delta_\nu(x)}\left[1 - \left(1 + \frac{\delta_\nu(x)\,\mathsf{Gap}(x)}{\mathsf{e}(x)^2}\frac{4-\nu}{\nu-2}\right)^{-\frac{\nu-2}{4-\nu}}\right] & \text{if } \nu \in (2,3) \cup (3,4), \\ \frac{\mathsf{Gap}(x)}{\delta_3(x)\,\mathsf{Gap}(x) + \mathsf{e}(x)^2} & \text{if } \nu = 3, \\ \frac{1}{\delta_4(x)}\left[1 - \exp\left(-\frac{\delta_4(x)\,\mathsf{Gap}(x)}{\mathsf{e}(x)^2}\right)\right] & \text{if } \nu = 4 \end{cases} \tag{4.1}$$

where $\delta_\nu(x)$ is defined in eq. (**??**).

Next we show that `FW-GSC` is well-defined using the step size policy (**??**).

Proposition 4.2. *Let $\{x^k\}_{k\geq 0}$ be generated by `FW-GSC` with step size policy $\{\alpha_\nu(x^k)\}_{k\geq 0}$ defined in* (**??**). *Then $x^k \in \mathcal{X}$ for all $k \geq 0$.*

Proof. If $\nu = 2$ then since $\alpha_2(x^k) \leq 1$, feasibility follows immediately from convexity of $\mathcal{X}$.
If $\nu \in (2, 4]$, we remark that whenever $x^k \in \mathcal{X}$, we deduce from (**??**) that $\mathtt{t}_\nu(x^k)\delta_\nu(x^k) < 1$. If $\mathtt{t}_\nu(x^k) > 1$, then $\alpha_\nu(x^k)\delta_\nu(x^k) = \delta_\nu(x^k) < \mathtt{t}_\nu(x^k)\delta_\nu(x^k) < 1$. ∎

To assess the iteration complexity of FW-GSC, we simplify the notation a bit by setting $\alpha_k \equiv \alpha_\nu(x^k)$ and $\Delta^k \equiv \eta_{x^k,\nu}(\alpha_\nu(x^k))$. Along the sequence $\{x^k\}_{k\geq 0}$ we have $\mathsf{d}_\nu(x^k, x^{k+1}) = \alpha_k \delta_\nu(x^k) < 1$, and we reduce the objective function value by at least the quantity $\Delta^k > 0$. Whence,

$$f(x^{k+1}) \leq f(x^k) - \Delta^k < f(x^k), \tag{4.2}$$

so that $f(x^k) \leq f(x^0)$, or equivalently, $\{x^k\}_{k\geq 0} \subset \mathcal{S}(x^0) \triangleq \{x \in \operatorname{dom} f \cap \mathcal{X} | f(x) \leq f(x^0)\}$. It is clear that $x^* \in \mathcal{S}(x^0)$.

Lemma 4.3. *The set $\mathcal{S}(x^0)$ is compact.*

Proof. By the optimality condition (**??**) and the bound (**??**), we have that for all $x \in \mathcal{S}(x^0)$

$$f(x^0) \geq f(x) \geq f(x^*) + \omega_\nu(-\mathsf{d}_\nu(x^*, x))\|x - x^*\|^2_{x^*}.$$

This means that $\omega_\nu(-\mathsf{d}_\nu(x^*, x))\|x - x^*\|^2_{x^*} \leq f(x^0) - f^*$. Therefore,

$$\mathcal{S}(x^0) \subseteq \{x \in \operatorname{dom} f | \omega_\nu(-\mathsf{d}_\nu(x^*, x))\|x - x^*\|^2_{x^*} \leq f(x^0) - f^*\} \subset \operatorname{dom} f$$

Clearly, the sandwiched set is closed. Since $\mathcal{X}$ is compact, the claim follows. ∎

Accordingly, the numbers

$$L_{\nabla f} \triangleq \max_{x\in\mathcal{S}(x^0)} \lambda_{\max}(\nabla^2 f(x)), \text{ and } \sigma_f \triangleq \min_{x\in\mathcal{S}(x^0)} \lambda_{\min}(\nabla^2 f(x)),$$

are well defined and finite. Furthermore, since the level set $\mathcal{S}(x^0)$ is compact, we know $\nabla^2 f(x) \succ 0$ for all $x \in \mathcal{S}(x^0)$ [**?** , Prop. 3a], and hence $\sigma_f > 0$. By [**?** , Thm. 2.1.11], for any $x \in \mathcal{S}(x^0)$ it holds that

$$f(x) - f(x^*) \geq \frac{\sigma_f}{2}\|x - x^*\|_2^2. \tag{4.3}$$

Proposition **??** below shows asymptotic convergence to a solution along subsequences. We omit the standard proof, as it follows from [**?**].

Proposition 4.4. *The following assertions hold for FW-GSC:*

(a) $\{f(x^k)\}_{k\geq 0}$ *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* $\min_{0\leq k<K} \Delta^k \leq \frac{1}{K}(f(x^0) - f^*)$.

In order to assess the iteration complexity of the method we need a lower bound on the sequence $\{\Delta^k\}_{k\geq 0}$. We start with a bound at iterations satisfying $\mathtt{t}_\nu(x^k) > 1$.

Lemma 4.5. *If* $\mathtt{t}_\nu(x^k) > 1$, *we have* $\Delta^k \geq \frac{1}{2}\,\mathsf{Gap}(x^k)$.

Proof. See Appendix **??**. ∎

Next, we turn to iterates for which $\mathtt{t}_\nu(x^k) \leq 1$. In this case, the per-iteration progress reads as $\Delta^k = \eta_{x^k,\nu}(\mathtt{t}_\nu(x^k))$, and enjoys the following lower bound:

Lemma 4.6. *If* $\mathtt{t}_\nu(x^k) \le 1$, *we have*

$$\Delta_k \ge \tilde{\Delta}^k \triangleq \begin{cases} \frac{2\ln 2-1}{\operatorname{diam}(\mathcal{X})} \min\left\{\frac{\mathsf{Gap}(x^k)}{M_f \operatorname{diam}(\mathcal{X})}, \frac{\mathsf{Gap}(x^k)^2}{\operatorname{diam}(\mathcal{X}) L_{\nabla f}}\right\} & \text{if } \nu = 2, \\ \frac{\tilde{\gamma}_\nu}{\operatorname{diam}(\mathcal{X})} \min\left\{\frac{\mathsf{Gap}(x^k)}{\left(\frac{\nu}{2}-1\right) M_f L_{\nabla f}^{(\nu-2)/2}}, \frac{-1}{\mathtt{b}} \frac{\mathsf{Gap}(x^k)^2}{L_{\nabla f} \operatorname{diam}(\mathcal{X})}\right\} & \text{if } \nu \in (2,3) \cup (3,4), \\ \frac{1-\ln 2}{\frac{M_f}{2}\sqrt{L_{\nabla f}} \operatorname{diam}(\mathcal{X})} \min\{\mathsf{Gap}(x^k), \frac{M_f \mathsf{Gap}(x^k)^2}{\sqrt{L_{\nabla f}} \operatorname{diam}(\mathcal{X})}\} & \text{if } \nu = 3, \\ \frac{\exp(-1)}{M_f L_{\nabla f}} \min\{\mathsf{Gap}(x^k), \frac{M_f \mathsf{Gap}(x^k)^2}{\operatorname{diam}(\mathcal{X})}\} & \text{if } \nu = 4 \end{cases} \tag{4.4}$$

where $\tilde{\gamma}_\nu \triangleq 1 + \frac{4-\nu}{2(3-\nu)}\left(1 - 2^{2(3-\nu)/(4-\nu)}\right)$ *and* $\mathtt{b} \triangleq \frac{2-\nu}{4-\nu}$.

Proof. See Appendix **??**. ∎

Remark 4.1. It can be checked that $\lim_{\nu \to 3} \tilde{\gamma}_\nu = 1 - \ln(2)$, so that the lower bound $\tilde{\Delta}^k$ is continuous in the parameter range $\nu \in (2, 4)$.

Combining Lemma **??** with Lemma **??** and estimates summarized in Appendix **??**, we get the next fundamental relation.

Proposition 4.7. *Let* $\{x^k\}_{k \ge 0}$ *be generated by* `FW-GSC`. *Then, for all* $k \ge 0$ *we have*

$$\Delta^k \ge \min\{\mathtt{c}_1(\nu)\, \mathsf{Gap}(x^k), \mathtt{c}_2(\nu)\, \mathsf{Gap}(x^k)^2\}$$

where

$$\mathtt{c}_1(\nu) \triangleq \begin{cases} \min\left\{\frac{1}{2}, \frac{2\ln(2)-1}{M_f \operatorname{diam}(\mathcal{X})^2}\right\} & \text{if } \nu = 2, \\ \min\left\{\frac{1}{2}, \frac{\tilde{\gamma}_\nu}{\operatorname{diam}(\mathcal{X})(\nu/2-1) M_f L_{\nabla f}^{(\nu-2)/2}}\right\} & \text{if } \nu \in (2,3) \cup (3,4), \\ \min\left\{\frac{1}{2}, \frac{1-\ln 2}{\frac{M_f}{2}\sqrt{L_{\nabla f}} \operatorname{diam}(\mathcal{X})}\right\} & \text{if } \nu = 3, \\ \min\left\{\frac{1}{2}, \frac{\exp(-1)}{M_f L_{\nabla f}}\right\} & \text{if } \nu = 4 \end{cases} \tag{4.5}$$

and

$$\mathtt{c}_2(\nu) \triangleq \begin{cases} \frac{2\ln(2)-1}{L_{\nabla f} \operatorname{diam}(\mathcal{X})^2} & \text{if } \nu = 2, \\ \frac{-1}{\mathtt{b}} \frac{\tilde{\gamma}_\nu}{\operatorname{diam}(\mathcal{X})^2 L_{\nabla f}} & \text{if } \nu \in (2,3) \cup (3,4), \\ \frac{1-\ln 2}{\frac{M_f}{2} L_{\nabla f} \operatorname{diam}(\mathcal{X})^2} & \text{if } \nu = 3, \\ \frac{\exp(-1)}{M_f L_{\nabla f} \operatorname{diam}(\mathcal{X})} & \text{if } \nu = 4 \end{cases} \tag{4.6}$$

Proof. We only illustrate the lower bound for the case $\nu = 2$. All other claims can be verified in exactly the same way. From Lemma **??** we know that $\Delta^k \ge \frac{1}{2}\mathsf{Gap}(x^k)$ whenever $\mathtt{t}_2(x^k) > 1$. If $\mathtt{t}_2(x^k) \le 1$, then $\Delta^k \ge \frac{2\ln 2-1}{\operatorname{diam}(\mathcal{X})} \min\left\{\frac{\mathsf{Gap}(x^k)}{M_f \operatorname{diam}(\mathcal{X})}, \frac{\mathsf{Gap}(x^k)^2}{\operatorname{diam}(\mathcal{X}) L_{\nabla f}}\right\}$. Consequently,

$$\Delta^k \ge \min\left\{\min\left\{\frac{1}{2}, \frac{2\ln(2)-1}{M_f \operatorname{diam}(\mathcal{X})^2}\right\} \mathsf{Gap}(x^k), \frac{2\ln(2)-1}{\operatorname{diam}(\mathcal{X})^2 L_{\nabla f}} \mathsf{Gap}(x^k)^2\right\}.$$

∎

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

Theorem 4.8. *Let $\{x^k\}_{k\geq 0}$ be generated by* `FW-GSC`*. For $x^0 \in \mathcal{X} \cap \operatorname{dom} f$, define $N_\varepsilon(x^0) \triangleq \inf\{k \geq 0 | h_k \leq \varepsilon\}$. Then, for all $\varepsilon > 0$,*

$$N_\varepsilon(x^0) \leq \frac{\ln\left(\frac{\mathsf{c}_1(\nu)}{h_0 \mathsf{c}_2(\nu)}\right)}{\ln(1-\mathsf{c}_1(\nu))} + \frac{1}{\mathsf{c}_2(\nu)\varepsilon}. \tag{4.7}$$

Proof. To simplify the notation, let us set $\mathsf{c}_1 \equiv \mathsf{c}_1(\nu)$ and $\mathsf{c}_2 \equiv \mathsf{c}_2(\nu)$. By convexity, we have $\mathsf{Gap}(x^k) \geq h_k$. Therefore, Proposition **??** shows that $\Delta^k \geq \min\{\mathsf{c}_1 h_k, \mathsf{c}_2 h_k^2\}$. This implies

$$h_{k+1} \leq h_k - \min\{\mathsf{c}_1 h_k, \mathsf{c}_2 h_k^2\} \qquad \forall k \geq 0.$$

From this inequality we see that h_k is decreasing and, hence, there are two phases of convergence:

Phase I. $\mathsf{c}_1 h_k < \mathsf{c}_2 h_k^2$, which is equivalent to $h_k > \frac{\mathsf{c}_1}{\mathsf{c}_2}$.

Phase II. $\mathsf{c}_1 h_k \geq \mathsf{c}_2 h_k^2$, which is equivalent to $h_k \leq \frac{\mathsf{c}_1}{\mathsf{c}_2}$.

For fixed initial condition $x^0 \in \operatorname{dom} f \cap \mathcal{X}$, we can thus subdivide the time domain into the set $\mathcal{K}_1(x^0) \triangleq \{k \geq 0 | h_k > \frac{\mathsf{c}_1}{\mathsf{c}_2}\}$ (Phase I) and $\mathcal{K}_2(x^0) \triangleq \{k \geq 0 | h_k \leq \frac{\mathsf{c}_1}{\mathsf{c}_2}\}$ (Phase II). Since $\{h_k\}_{k\in\mathcal{K}_1(x^0)}$ is decreasing and bounded from below by the positive constant $\mathsf{c}_1/\mathsf{c}_2$, the set $\mathcal{K}_1(x^0)$ is at most finite. Let us set

$$T_1(x^0) \triangleq \inf\{k \geq 0 | h_k \leq \frac{\mathsf{c}_1}{\mathsf{c}_2}\}, \tag{4.8}$$

the first time at which the process $\{h_k\}_k$ enters Phase II. To get a worst-case estimate on this quantity, we assume without loss of generality 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 $\frac{\mathsf{c}_1}{\mathsf{c}_2} < h_k \leq h_{k-1} - \min\{\mathsf{c}_1 h_{k-1}, \mathsf{c}_2 h_{k-1}^2\} = h_{k-1} - \mathsf{c}_1 h_{k-1}$. Note that $\mathsf{c}_1 \leq 1/2$, so we make progressions like a geometric series, i.e. we have linear convergence in this phase. Hence, $h_k \leq (1-\mathsf{c}_1)^k h_0$ for all $k = 0, \ldots, T_1(x^0) - 1$. By definition $h_{T_1(x^0)-1} > \frac{\mathsf{c}_1}{\mathsf{c}_2}$, so we get $\frac{\mathsf{c}_1}{\mathsf{c}_2} \leq h_0(1-\mathsf{c}_1)^{T_1(x^0)-1}$ iff $(T_1(x^0) - 1)\ln(1-\mathsf{c}_1) \geq \ln\left(\frac{\mathsf{c}_1}{h_0\mathsf{c}_2}\right)$. Hence,

$$T_1(x^0) \leq \left\lceil \frac{\ln\left(\frac{\mathsf{c}_1}{h_0\mathsf{c}_2}\right)}{\ln(1-\mathsf{c}_1)} \right\rceil. \tag{4.9}$$

After these number of iterations, the process will enter Phase II, at which $h_k \leq \frac{\mathsf{c}_1}{\mathsf{c}_2}$ holds. Therefore, $h_k \geq h_{k+1} + \mathsf{c}_2 h_k^2$, or equivalently,

$$\frac{1}{h_{k+1}} \geq \frac{1}{h_k} + \mathsf{c}_2 \frac{h_k}{h_{k+1}} \geq \frac{1}{h_k} + \mathsf{c}_2. \tag{4.10}$$

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

$$\frac{1}{h_N} \geq \frac{1}{h_{T_1(x^0)}} + \mathsf{c}_2(N - T_1(x^0) + 1).$$

By definition $h_{T_1(x^0)} \leq \frac{\mathsf{c}_1}{\mathsf{c}_2}$, so that for all $N > T_1(x^0)$, we see

$$\frac{1}{h_N} \geq \frac{\mathsf{c}_2}{\mathsf{c}_1} + \mathsf{c}_2(N - T_1(x^0) + 1).$$

Consequently,

$$h_N \leq \frac{1}{\frac{c_2}{c_1} + c_2(N - T_1(x^0) + 1)} \leq \frac{1}{c_2(N - T_1(x^0) + 1)}. \tag{4.11}$$

By definition of the stopping time $N_\varepsilon(x^0)$, it is true that $h_{N_\varepsilon(x^0)-1} > \varepsilon$. Consequently, evaluating (**??**) at $N = N_\varepsilon(x^0) - 1$, we obtain

$$\varepsilon \leq \frac{1}{c_2(N_\varepsilon(x^0) - T_1(x^0))} \Leftrightarrow N_\varepsilon(x^0) \leq T_1(x^0) + \frac{1}{c_2\varepsilon}.$$

Combining this upper bound with (**??**) shows the claim. ∎

4.2 Complexity Analysis of `BacktrackFW-GSC`

Before assessing the iteration complexity of `Backtrack-FW`, we give a (standard) estimate on the number of calls of the backtracking subroutine, in the spirit of [**?**].

Proposition 4.9. *Let N_k be the number of function evaluations of the sufficient decrease condition* (**??**) *up to iteration k. Then*

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

Proof. The proof is analogous to [**?** , Lemma 4] and [**?**] and thus omitted. ∎

In practice, a good choice for the parameters is $\gamma_d = 0.9$ and $\gamma_u = 2$, which would roughly result into 16 % of the iterates with more than a single function evaluation.

Theorem 4.10. *Let $\{x^k\}_{k\geq 0}$ be generated by `Backtrack-FW`. Then, for all $x^0 \in \mathcal{X} \cap \operatorname{dom} f$, we have*

$$h_k \leq \frac{2\,\mathsf{Gap}(x^0)}{(k+1)(k+2)} + \frac{k\,\operatorname{diam}(\mathcal{X})^2}{(k+1)(k+2)} L_{\nabla f}. \tag{4.12}$$

Proof. Define the Fenchel conjugate

$$f^*(u) = \sup_{z \in \operatorname{dom} f} \{\langle z, u\rangle - f(z)\}. \tag{4.13}$$

Since f is proper and closed convex, f^* is well-defined and also proper, closed and convex. By Fermat's rule, if $u^*(x) \in \mathbb{R}^n$ satisfies $\nabla f(u^*(x)) - x = 0$, then f^* is well defined at x. In particular, this shows that $\operatorname{dom}(f^*) \triangleq \{z \in \mathbb{R}^n | (\exists u \in \mathbb{R}^n) : \nabla f(u) - z = 0\}$.

Moreover, if $z^*(u)$ solves the equation $\nabla f(z^*(u)) = u$, then we have $f^*(u) = \langle u, \nabla f(z^*(u))\rangle - f(z^*(u))$. In particular, we see that for $x \in \operatorname{dom} f$, we have $\nabla f(x) \in \operatorname{dom} f^*$, which implies $f^*(\nabla f(x^k)) \geq \langle \nabla f(x^k), x^k\rangle - f(x^k)$ for all $k \geq 0$. On the other hand, convexity shows that for all $u \in \operatorname{dom} f$, $\langle \nabla f(x^k), x^k\rangle - f(x^k) \geq \langle \nabla f(x^k), u\rangle - f(u)$. Whence,

$$\langle \nabla f(x^k), x^k\rangle - f(x^k) = \sup_{u \in \operatorname{dom} f} \{\langle u, \nabla f(x^k)\rangle - f(u)\} = f^*(\nabla f(x^k)). \tag{4.14}$$

Define the support function $H_{\mathcal{X}}(u) \triangleq \sup\{\langle u, x\rangle | x \in \mathcal{X}\}$ for all $u \in \mathbb{R}^n$. Since $\mathcal{X}$ is convex compact, the support function is convex (sub-additive) and finite everywhere, i.e. $\operatorname{dom} H_{\mathcal{X}} = \mathbb{R}^n$. Its Fenchel

conjugate is the indicator function $\delta_{\mathcal{X}}(x) \triangleq H^*_{\mathcal{X}}(x)$. Using these concepts, we can rewrite the primal optimization problem (**??**) as the non-smooth composite convex optimization problem

$$\min_{x\in\mathbb{R}^n}\{f(x) + \delta_{\mathcal{X}}(x)\}.$$

The dual maximization problem is

$$\psi^* \triangleq \max_{z\in\mathbb{R}^n}\{\psi(z) \triangleq -f^*(z) - H_{\mathcal{X}}(-z)\}. \tag{4.15}$$

Strong duality states that $f^* = \psi^*$. We further see $\mathsf{Gap}(x^k) = \langle \nabla f(x^k), x^k - s^k\rangle = \langle \nabla f(x^k), x^k\rangle + H_{\mathcal{X}}(-\nabla f(x^k))$, which, when coupled with (**??**), yields

$$\mathsf{Gap}(x^k) = f(x^k) - \psi(\nabla f(x^k)). \tag{4.16}$$

Lemma **??** tells us that for all $t \in [0,1]$,

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

Let us introduce the auxiliary "dual averaging" 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 obtain the primal-dual gap estimate

$$f(x^k) - \psi(y^k) = f(x^k) - f^* + \psi^* - \psi(y^k) \ge f(x^k) - f^* = h_k.$$

Since ψ is concave, we see that

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

Consequently, in conjunction with (**??**),

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

Define $A_k \triangleq \frac{1}{2}(k+1)(k+2)$ for $k \ge 0$. For this specification, it is easy to check that $A_{k+1}(1-\xi_k) = A_k$, and $A_{k+1}\frac{\xi_k^2}{2} \le 1$. Hence,

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

Summing from $i = 0,\dots,k-1$, and calling $\bar{\mathcal{L}}_k \triangleq \frac{1}{k}\sum_{i=0}^{k-1}\mathcal{L}_i$, this implies

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

Since $\mathcal{L}_k \le L_{\nabla f}$, it follows $\bar{\mathcal{L}}_k \le L_{\nabla f}$, and hence, when combined with (**??**), we conclude

$$h_k \le \frac{2\,\mathsf{Gap}(x^0)}{(k+1)(k+2)} + \frac{k\operatorname{diam}(\mathcal{X})^2}{(k+1)(k+2)}L_{\nabla f}.$$

■

5 A linearly convergent variant of Frank-Wolfe for generalized self-concordant functions

In this section we show how the local linear minimization oracle of [**?**] can be adapted to accelerate the convergence of FW-methods for minimizing GSC functions. In particular, we work out an analytic step-size criterion which guarantees linear convergence towards the unique solution of (**??**). The construction is a non-trivial modification of [**?**], as it exploits the local descent properties of GSC functions. In particular, we neither assume global Lipschitz continuity, nor strong convexity of the objective function. Instead, we assume that the feasible set admits the explicit representation

$$\mathcal{X} \triangleq \{x \in \mathbb{R}^n | Ax = a, Bx \leq b\}, \tag{5.1}$$

where $A, B \in \mathbb{R}^{m\times n}$ and $a, b \in \mathbb{R}^m$. Let $\mathbb{B}(x, r) \triangleq \{y \in \mathbb{R}^n | \ \|y - x\|_2 \leq r\}$ denote the closed ℓ_2-ball with radius r and center x.

Definition 5.1 ([**?**], Def. 2.5)**.** A procedure $\mathcal{A}(x, r, c)$, where $x \in \mathcal{X}, r > 0, c \in \mathbb{R}^n$, is a LLOO with parameter $\rho \geq 1$ for the polytope $\mathcal{X}$ if $\mathcal{A}(x, r, c)$ returns a point $u(x, r, c) = u \in \mathcal{X}$ such that

$$\forall y \in \mathbb{B}(x, r) \cap \mathcal{X} : \langle c, y\rangle \geq \langle c, u\rangle, \text{ and } \|x - u\|_2 \leq \rho r. \tag{5.2}$$

We refer to [**?**] for illustrative examples for oracles $\mathcal{A}(x, r, c)$. In particular, [**?**] provide an explicit construction of the LLOO for a simplex and for general polytopes. Let us define the value function of the LLOO by

$$\Gamma(x, r) \triangleq \langle \nabla f(x), x - u(x, r, \nabla f(x))\rangle = \max_{s\in\mathbb{B}(x,r)\cap\mathcal{X}} \langle \nabla f(x), x - s\rangle. \tag{5.3}$$

It is clear that $\Gamma(x, r) \leq \mathsf{Gap}(x)$ for all $x \in \mathcal{X} \cap \operatorname{dom} f$. We further redefine the local norm as

$$\mathsf{e}(x) \triangleq \|u(x, r, \nabla f(x)) - x\|_x \qquad \forall x \in \operatorname{dom} f.$$

To measure step-lengths, we also redefine the function $\delta_\nu(x)$ to

$$\delta_\nu(x) \triangleq \begin{cases} M_f\|u(x, r, \nabla f(x)) - x\|_2 & \text{if } \nu = 2, \\ \frac{\nu-2}{2} M_f\|u(x, r, \nabla f(x)) - x\|_2^{3-\nu}\|u(x, r, \nabla f(x)) - x\|_x^{\nu-2} & \text{if } \nu > 2. \end{cases}, \tag{5.4}$$

Our point of departure is the upper estimate for GSC functions (**??**), which in the present situation reads as

$$\begin{aligned} f\left(x + t(u(x, r, \nabla f(x)) - x)\right) &\leq f(x) + t\langle \nabla f(x), u(x, r, \nabla f(x)) - x\rangle + \omega_\nu(t\delta_\nu(x))t^2\mathsf{e}(x)^2 \\ &= f(x) - t\Gamma(x, r) + \omega_\nu(t\delta_\nu(x))t^2\mathsf{e}(x)^2 \\ &\leq f(x) - t\,\mathsf{Gap}(x) + \omega_\nu(t\delta_\nu(x))t^2\mathsf{e}(x)^2 \\ &= f(x) - \frac{t}{2}\,\mathsf{Gap}(x) - \frac{\mathsf{Gap}(x)}{2}\left(t - \omega_\nu(t\delta_\nu(x))t^2\frac{2\mathsf{e}(x)^2}{\mathsf{Gap}(x)}\right). \end{aligned}$$

Phrasing this in terms of the approximation error $h(x) = f(x) - f(x^*)$, we get first

$$-h(x) = f(x^*) - f(x) \geq \langle \nabla f(x), x^* - x\rangle \geq -\mathsf{Gap}(x),$$

Algorithm 5: `FW-LLOO`

Input: $\mathcal{A}(x,r,c)$-LLOO with parameter $\rho \geq 1$ for polytope $\mathcal{X}$, $f \in \mathcal{F}_{M_f,\nu}(\operatorname{dom} f)$. $\sigma_f > 0$ convexity parameter.
$x^0 \in \operatorname{dom} f \cap \mathcal{X}$, and let $h_0 = f(x^0) - f^*$, and $c_0 = 1$.
for k=0 **do**
  Set $r_0 = \sqrt{\frac{2\,\mathsf{Gap}(x^0)}{\sigma_f}}$.
  Obtain $u^0 = u(x^0, r_0, \nabla f(x^0))$ as in (**??**).
  Set $\alpha_0 = \alpha_\nu(x^0)$ as in (**??**).
  Update $x^1 = x^0 + \alpha_0(u^0 - x^0)$
end for
for $k = 1, \ldots$ **do**
  if $\mathsf{Gap}(x^k) > \varepsilon$ **then**
    Set $c_k = \exp\left(-\frac{1}{2}\sum_{i=0}^{k-1}\alpha_i\right)$
    Set $r_k = r_0 c_k$.
    Obtain $u^k = u(x^k, r_k, \nabla f(x^k))$ as in (**??**).
    Set $\alpha_k = \alpha_\nu(x^k)$ as in (**??**).
    Set $x^{k+1} = x^k + \alpha_k(u^k - x^k)$
  end if
end for

using convexity of the objective function f, and second

$$\begin{aligned} h\left(x + t(u(x,r,\nabla f(x)) - x)\right) &\leq h(x) - \frac{t}{2}\mathsf{Gap}(x) - \frac{\mathsf{Gap}(x)}{2}\left(t - \omega_\nu(t\delta_\nu(x))t^2\frac{2\mathsf{e}(x)^2}{\mathsf{Gap}(x)}\right) \\ &\leq \left(1 - \frac{t}{2}\right)\mathsf{Gap}(x) - \frac{\mathsf{Gap}(x)}{2}\left(t - \omega_\nu(t\delta_\nu(x))t^2\frac{2\mathsf{e}(x)^2}{\mathsf{Gap}(x)}\right). \end{aligned} \tag{5.5}$$

Similar to `FW-GSC`, our goal is to choose the stepsize t such that the expression in the brackets on the right-hand-side above is non-negative. To that end, let us introduce the function

$$\psi_\nu(t) \triangleq t - \xi\omega_\nu(t\delta)t^2 \qquad t \in [0, 1/\delta),$$

where $\xi, \delta \geq 0$ are free parameters. Observe that setting $\delta = \delta_\nu(x)$ and $\xi = \frac{2\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, we obtain the expression in the last brackets of (**??**). Note that this function is used already in the complexity analysis of `FW-GSC`, and thoroughly discussed in Appendix **??**. In particular, $t \mapsto \psi_\nu(t)$ is concave unimodal with $\psi_\nu(0) = 0$, increasing on the interval $[0, t_\nu^*)$ and decreasing on $[t_\nu^*, \infty)$, where the cut-off value t_ν^* is defined in eq. (**??**). For the readers' convenience we reprint its definition here:

$$t_\nu^* \triangleq \begin{cases} \frac{1}{\delta}\ln\left(1 + \frac{\delta}{\xi}\right) & \text{if } \nu = 2, \\ \frac{1}{\delta}\left[1 - \left(1 + \frac{\delta}{\xi}\frac{4-\nu}{\nu-2}\right)^{-\frac{\nu-2}{4-\nu}}\right] & \text{if } \nu \in (2,3) \cup (3,4), \\ \frac{1}{\delta+\xi} & \text{if } \nu = 3, \\ \frac{1}{\delta}\left[1 - \exp\left(-\frac{\delta}{\xi}\right)\right] & \text{if } \nu = 4. \end{cases} \tag{5.6}$$

Setting $\delta = \delta_\nu(x)$ and $\xi = \frac{2\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, an efficient choice of the step size is

$$\alpha_\nu(x) = \min\{1, \frac{1}{\delta_\nu(x)}, \tau_\nu(x)\}, \tag{5.7}$$

where $\tau_\nu(x)$ is the unique positive root of the function $\psi_\nu(t)$, using the parameters $\delta = \delta_\nu(x)$ and $\xi = \frac{2\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$. Indeed, this guarantees decrease in the objective function value, feasibility of the iterates and linear convergence rates, once a suitable lower bond on $\alpha_\nu(x)$ has been identified. To work out a practical lower bound, observe that $\tau_\nu(x) \geq t_\nu^*$ implies

$$\alpha_\nu(x) \geq \min\{1, \frac{1}{\delta_\nu(x)}, t_\nu^*\} = \min\{1, t_\nu^*\}, \tag{5.8}$$

using the fact that $\delta_\nu(x) t_\nu^* < 1$.

Given the sequence $\{x^k\}_k$, we define the associated sequence $\{t_\nu^k\}_k$, in which t_ν^k is (**??**) evaluated at the parameters $\delta = \delta_\nu(x^k) \equiv \delta_k$ and $\xi = \frac{2\mathsf{e}(x^k)^2}{\mathsf{Gap}(x^k)c_k} \equiv \xi_k$. We thus have $\alpha_k = \alpha_\nu(x^k) \geq \min\{1, t_\nu^k\}$ for all $k \geq 0$.

Theorem 5.2. *Let $\{x^k\}_{k\geq 0}$ be generated by* `LLOO-FW`*. Then, for all $k \geq 0$, we have $x^* \in \mathbb{B}(x^k, r_k)$ and*

$$h_k \leq \mathsf{Gap}(x^0) \exp\left(-\frac{1}{2}\sum_{i=0}^{k-1} \alpha_i\right) \tag{5.9}$$

Proof. Let us define $\mathcal{P}(x^0) \triangleq \left\{x \in \mathcal{X} : f(x) \leq f^* + \mathsf{Gap}(x^0)\right\}$. We proceed by induction. For $k = 0$, we have the given initial condition $x^0 \in \operatorname{dom} f \cap \mathcal{X}$. Trivially, $x^0 \in \mathcal{P}(x^0)$, so that (**??**) gives

$$f(x^0) - f(x^*) = h_0 \geq \frac{\sigma_f}{2}\|x^0 - x^*\|_2^2. \tag{5.10}$$

Denote by $u^0 \equiv u(x^0, r_0, \nabla f(x^0))$ the solution of a single call of procedure $\mathcal{A}(x^0, r_0, \nabla f(x^0))$. Set $\delta_0 \equiv \delta_\nu(x^0)$. Since $r_0 = \sqrt{\frac{2\,\mathsf{Gap}(x^0)}{\sigma_f}} \geq \sqrt{\frac{2h_0}{\sigma_f}}$, (**??**) implies that $x^* \in \mathbb{B}(x^0, r_0)$. By (**??**), we have

$$h(x^0 + t(u^0 - x^0)) \leq (1 - t/2)\,\mathsf{Gap}(x^0) - \frac{\mathsf{Gap}(x^0)}{2}\left(t - \omega_\nu(t\delta_0)t^2\frac{2\mathsf{e}(x^0)^2}{\mathsf{Gap}(x^0)}\right)$$

for $t \in (0, \min\{1, 1/\delta_0\})$. By the choice of the stepsize α_0, the second term in brackets above is positive, so we get for $t = \alpha_0$ and $x_1 = x^0 + \alpha_0(u^0 - x^0)$ the inequality

$$h_1 = h(x^1) \leq (1 - \alpha_0/2)\,\mathsf{Gap}(x^0) \leq \mathsf{Gap}(x^0)\exp(-\alpha_0/2).$$

Now assume that for some $k \geq 1$ it holds

$$h_k \leq \mathsf{Gap}(x^0)c_k,\ c_k \triangleq \exp\left(-\frac{1}{2}\sum_{i=0}^{k-1}\alpha_i\right). \tag{5.11}$$

Then $x^k \in \mathcal{P}(x^0)$ and (**??**) leads to

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

Hence, $x^* \in \mathbb{B}(x^k, r_k)$ and using the definition of the LLOO returning the target vector $u^k = u(x^k, r_k, \nabla f(x^k))$, eq. (**??**) tells us

$$-h_k = f(x^*) - f(x^k) \geq \langle \nabla f(x^k), x^* - x^k\rangle \geq \langle \nabla f(x^k), u^k - x^k\rangle = -\Gamma(x^k, r_k).$$

Proceeding as in the case $k = 0$, we get that for $t \in (0, \min\{1, 1/\delta_k\})$, by (**??**),

$$f(x^k + t(u^k - x^k)) \leq f(x^k) - t\Gamma(x^k, r_k) + \omega_\nu(t\delta_k)t^2\mathsf{e}(x^k)^2.$$

Whence, by the induction assumption,

$$\begin{aligned} f(x^k + t(u^k - x^k)) - f(x^*) &\leq h_k - th_k + \omega_\nu(t\delta_k)t^2\mathsf{e}(x^k)^2 = (1-t)h_k + \omega_\nu(t\delta_k)t^2\mathsf{e}(x^k)^2 \\ &\leq (1-t)\,\mathsf{Gap}(x^0)c_k + \omega_\nu(t\delta_k)t^2\mathsf{e}(x^k)^2 \\ &= (1 - t/2)\,\mathsf{Gap}(x^0)c_k - t/2\,\mathsf{Gap}(x^0)c_k + \omega_\nu(t\delta_k)t^2\mathsf{e}(x^k)^2 \\ &= (1 - t/2)\,\mathsf{Gap}(x^0)c_k - \frac{\mathsf{Gap}(x^0)c_k}{2}\left(t - \omega_\nu(t\delta_k)t^2\frac{2\mathsf{e}(x^k)^2}{\mathsf{Gap}(x^0)c_k}\right). \end{aligned}$$

If we choose the stepsize α_k as in (**??**), we are ensured that the expression in the brackets on the right-hand-side is non-negative. Consequently, we obtain $h_{k+1} \leq (1 - \alpha_k/2)\,\mathsf{Gap}(x^0)c_k \leq \mathsf{Gap}(x^0)c_k \exp(-\alpha_k/2) = \mathsf{Gap}(x^0)c_{k+1}$, which finishes the induction proof. ∎

It remains to lower bound the step size sequence $\alpha_k = \alpha_\nu(x^k)$. By definition of the step-size rule (**??**), we always have $\alpha_k \geq \min\{1, t_\nu^k\}$, where t_ν^k is (**??**) evaluated at $\delta_k = \delta_\nu(x^k)$ and $\xi_k = \frac{2\mathsf{e}_k^2}{\mathsf{Gap}(x^k)c_k}$. Note that for all values $\nu \in [2, 4]$, t_ν^* is an increasing function of $\frac{\delta}{\xi}$. Thus, our next steps are to lower bound $1/\delta$ and $\frac{\delta}{\xi}$. By definition of the LLOO, we have $\|u^k - x^k\|_2 \leq \min\{\rho r_k, \mathrm{diam}(\mathcal{X})\}$. Hence,

$$\frac{1}{\delta_k} = \begin{cases} \frac{1}{M_f\|u^k - x^k\|_2} & \text{if } \nu = 2, \\ \frac{1}{\frac{\nu-2}{2}M_f\|u^k - x^k\|_2^{3-\nu}\|u^k - x^k\|_{x^k}^{\nu-2}} & \text{if } \nu > 2, \end{cases}$$

If $\nu = 2$, we have

$$\frac{1}{\delta_k} \geq \frac{1}{M_f \min\{\rho r_k, \mathrm{diam}(\mathcal{X})\}} \geq \frac{1}{M_f \rho r_k},$$

while if $\nu > 2$, we observe

$$\begin{aligned} \frac{1}{\delta_k} &\geq \frac{1}{\frac{\nu-2}{2}M_f\|u^k - x^k\|_2^{3-\nu}L_{\nabla f}^{\frac{\nu-2}{2}}\|u^k - x^k\|_2^{\nu-2}} = \frac{1}{\frac{\nu-2}{2}M_f L_{\nabla f}^{\frac{\nu-2}{2}}\|u^k - x^k\|_2} \\ &\geq \frac{1}{\frac{\nu-2}{2}M_f L_{\nabla f}^{\frac{\nu-2}{2}}\min\{\rho r_k, \mathrm{diam}(\mathcal{X})\}} \geq \frac{1}{\frac{\nu-2}{2}M_f L_{\nabla f}^{\frac{\nu-2}{2}}\rho r_k}. \end{aligned}$$

Furthermore, from the identity $\frac{2\,\mathsf{Gap}(x^0)c_k}{\sigma_f} = r_k^2$, we conclude $\mathsf{Gap}(x^0)c_k = \frac{\sigma_f r_k^2}{2}$. Hence,

$$\frac{\delta_k}{\xi_k} = \frac{\delta(x^k)\,\mathsf{Gap}(x^0)c_k}{2\mathsf{e}(x^k)^2} = \begin{cases} \frac{M_f\|u^k - x^k\|_2\frac{\sigma_f r_k^2}{2}}{2\|u^k - x^k\|_{x^k}^2} & \text{if } \nu = 2, \\ \frac{\frac{\nu-2}{2}M_f\|u^k - x^k\|_2^{3-\nu}\mathsf{e}(x^k)^{\nu-2}\frac{\sigma_f r_k^2}{2}}{2\mathsf{e}(x^k)^2} & \text{if } \nu > 2. \end{cases}$$

If $\nu = 2$, we see that

$$\frac{\delta_k}{\xi_k} \geq \frac{M_f\|u^k - x^k\|_2\sigma_f r_k^2}{4L_{\nabla f}\|u^k - x^k\|_2^2} = \frac{M_f\sigma_f r_k^2}{4L_{\nabla f}\|u^k - x^k\|_2} \geq \frac{M_f\sigma_f r_k^2}{4L_{\nabla f}\min\{\rho r_k, \mathrm{diam}(\mathcal{X})\}} \geq \frac{M_f\sigma_f r_k}{4\rho L_{\nabla f}},$$

while if $\nu > 2$, we have in turn

$$\frac{\delta_k}{\xi_k} = \frac{(\nu-2)M_f\|u^k - x^k\|_2^{3-\nu}\sigma_f r_k^2}{8\mathsf{e}(x^k)^{4-\nu}} \geq \frac{(\nu-2)M_f\|u^k - x^k\|_2^{3-\nu}\sigma_f r_k^2}{8L_{\nabla f}^{\frac{4-\nu}{2}}\|u^k - x^k\|_2^{4-\nu}} = \frac{(\nu-2)M_f\sigma_f r_k^2}{8L_{\nabla f}^{\frac{4-\nu}{2}}\|u^k - x^k\|_2}$$

$$\geq \frac{(\nu-2)M_f\sigma_f r_k^2}{8L_{\nabla f}^{\frac{4-\nu}{2}}}\min\{\rho r_k, \mathrm{diam}(\mathcal{X})\} \geq \frac{(\nu-2)M_f\sigma_f r_k}{8\rho L_{\nabla f}^{\frac{4-\nu}{2}}} = \frac{(\nu-2)M_f L_{\nabla f}^{\frac{\nu-2}{2}}\sigma_f r_k}{8\rho L_{\nabla f}}.$$

Denoting $\gamma_\nu = \frac{\nu-2}{2}M_f L_{\nabla f}^{\frac{\nu-2}{2}}$ for $\nu > 2$ and $\gamma_\nu = M_f$ for $\nu = 2$, and substituting these lower bounds to the expression for t_ν^*, we obtain

$$t_\nu^k \geq \begin{cases} \frac{1}{\gamma_\nu \rho r_k}\ln\left(1 + \frac{\gamma_\nu\sigma_f r_k}{4\rho L_{\nabla f}}\right) & \text{if } \nu = 2, \\ \frac{1}{\gamma_\nu \rho r_k}\left[1 - \left(1 + \frac{\gamma_\nu\sigma_f r_k}{4\rho L_{\nabla f}}\frac{4-\nu}{\nu-2}\right)^{-\frac{\nu-2}{4-\nu}}\right] & \text{if } \nu \in (2,3)\cup(3,4), \\ \frac{1}{\gamma_\nu \rho r_k}\frac{1}{1+\frac{4\rho L_{\nabla f}}{\gamma_\nu\sigma_f r_k}} & \text{if } \nu = 3, \\ \frac{1}{\gamma_\nu \rho r_k}\left[1 - \exp\left(-\frac{\gamma_\nu\sigma_f r_k}{4\rho L_{\nabla f}}\right)\right] & \text{if } \nu = 4. \end{cases}$$

For all $\nu \in [2,4]$, the right-hand-side of the latter inequality has a limit $\frac{\sigma_f}{4\rho^2 L_{\nabla f}}$ as $r_k \to 0$. Moreover, it is a decreasing function of r_k, whence,

$$t_\nu^k \geq \bar{\alpha} = \begin{cases} \frac{1}{\gamma_\nu \rho r_0}\ln\left(1 + \frac{\gamma_\nu\sigma_f r_0}{4\rho L_{\nabla f}}\right) & \text{if } \nu = 2, \\ \frac{1}{\gamma_\nu \rho r_0}\left[1 - \left(1 + \frac{\gamma_\nu\sigma_f r_0}{4\rho L_{\nabla f}}\frac{4-\nu}{\nu-2}\right)^{-\frac{\nu-2}{4-\nu}}\right] & \text{if } \nu \in (2,3)\cup(3,4), \\ \frac{1}{\gamma_\nu \rho r_0}\frac{1}{1+\frac{4\rho L_{\nabla f}}{\gamma_\nu\sigma_f r_0}} & \text{if } \nu = 3, \\ \frac{1}{\gamma_\nu \rho r_0}\left[1 - \exp\left(-\frac{\gamma_\nu\sigma_f r_0}{4\rho L_{\nabla f}}\right)\right] & \text{if } \nu = 4. \end{cases} \tag{5.12}$$

Corollary 5.3. *Algorithm* `LLOO-FW` *guarantees linear convergence in terms of the approximation error:*

$$h_k \leq \mathsf{Gap}(x^0)\exp(-k\bar{\alpha}/2) \qquad \forall k \geq 0,$$

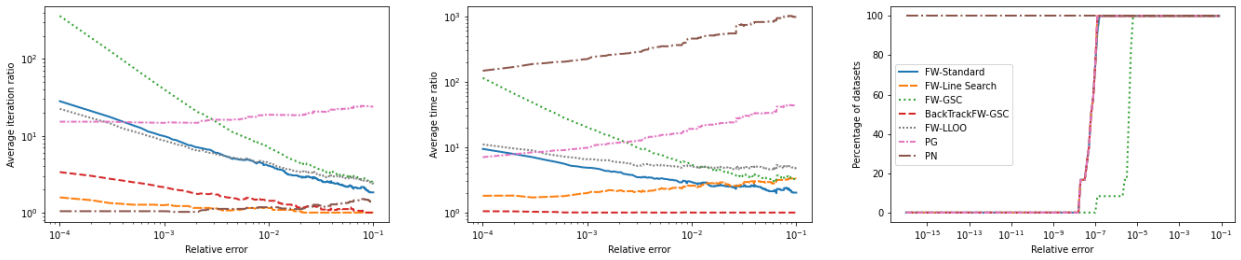
where $\bar{\alpha}$ *is defined in* (**??**).

Proof. By the lower bound (**??**) and the estimate (**??**), we see that $\alpha_k \geq \bar{\alpha}$ for all $k \geq 0$. Hence $\exp\left(-\frac{1}{2}\sum_{i=0}^{k-1}\alpha_i\right) \leq \exp(-k\bar{\alpha}/2)$, and the claim follows. ∎

6 Numerical Results

We provide four examples to compare our methods with existing methods in the literature. As competitors we choose vanilla FW (`FW-Standard`) using the standard step-size of $\frac{2}{k+2}$ (Standard) for which no general convergence proof for generalized self-concordance exists, as well as FW with exact line-search (`FW-Line Search`). As further benchmarks, we implement the self-concordant Proximal-Newton (`PN`) and the Proximal-Gradient (`PG`) of [**?**], as available in the SCOPT package[3].

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



(a) Average iteration ratio $\tilde{\rho}(\varepsilon)$. (b) Average time ratio $\hat{\rho}(\varepsilon)$. (c) Average success ratio $\rho(\varepsilon)$.

Figure 1: Performance Profile for the portfolio selection problem (**??**) obtained after averaging over 12 synthetically generated data sets.

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.[4]

We ran all first order methods for a maximum 50,000 iterations, and the PN method, which is more computationally expensive, for a maximum of 1,000 iterations. FW-Line Search is run with a tolerance of 10^{-10}. Within PN we use monotone FISTA [**?**], with at most 100 iteration and a tolerance of 10^{-5} to find the Newton direction, and the Lipschitz constant used in PG is determined by the Barzilai-Borwein method [**?**] with a limits of 100 iterations.

Our comparison is made by the construction of versions of performance profiles, following [**?**]. In order to present the result, we first estimate f^* by the best function value achieved by any of the algorithms, and compute the relative error attained by each of the methods at iteration k. More precisely, given the set of methods $\mathcal{S}$ and test problems $\mathcal{P}$, denote by F_{ij} the function value attained by method $i \in \mathcal{S}$ on problem $j \in \mathcal{P}$. We define the estimate of the optimal value of problem j by by $f_j^* = \min\{F_{sj} | s \in \mathcal{S}\}$. Denoting $(x_{ij}^k)_k$ the sequence produced by method i on problem j, we define the *relative error* as $r_{ij}^k = \frac{f(x_{ij}^k) - f_j^*}{f_j^*}$. Now, for all methods $i \in \mathcal{S}$ and any relative error ϵ, we compute the proportion of data sets that achieves a relative error of at most ϵ, that is $\rho_i(\epsilon) := \frac{1}{|\mathcal{P}|}|\{j \in \mathcal{P} | \exists k,\ r_{ij}^k \leq \varepsilon\}|$. We are also interested in comparing iteration complexity and CPU time. For that purpose, we define $N_{ij}(\varepsilon) = \min\{k \geq 0 | r_{ij}^k \leq \varepsilon\}$ as the first iteration in which method $i \in \mathcal{S}$ achieves a relative error ε on problem $j \in \mathcal{P}$. Analogously, $T_{ij}(\varepsilon)$ measures the minimal CPU time in which method $i \in \mathcal{S}$ achieves a relative error ε on problem $j \in \mathcal{P}$. We thus define $\tilde{\rho}_i(\varepsilon) := \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \frac{N_{ij}(\varepsilon)}{\min\{N_{sj}(\varepsilon) | s \in \mathcal{S}\}}$ for comparing the iteration complexity of all the methods, and the average time ratio $\hat{\rho}_i(\varepsilon) = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \frac{T_{ij}(\varepsilon)}{\min\{T_{sj}(\varepsilon) | s \in \mathcal{S}\}}$ for comparing the computational time of all the methods. In both cases, as the average ratio is closer to 1 the performance of the method is closer to the best performance. Besides average performance, we also report the performance of all tested methods on each data set.

6.1 Portfolio optimization with logarithmic utility

We study high-dimensional portfolio optimization problems with logarithmic utility. In this problem there are n assets with returns $r_t \in \mathbb{R}_+^n$ in period t of the investment horizon. The utility function

[4]The codes are publicly available on Github `https://github.com/kamil-safin/SCFW`.

of the investor is given as

$$f(x) = -\sum_{t=1}^{p} \log(r_t^\top x).$$

Our task is to design a portfolio x solving the problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t.: } x_i \geq 0, \sum_{t=1}^{p} x_t = 1. \tag{6.1}$$

Since f is the sum of n standard self-concordant functions, we see that $f \in \mathcal{F}_{2,3}(\mathbb{R}^n_{++})$ (Proposition **??**). We remark that this self-concordant minimization problem is relevant in the universal prediction problem in information theory [**?**].

For this example, computing a LLOO with $\rho = \sqrt{n}$ is simple, as described in [**?**]. Therefore, we also ran the `FW-LLOO` algorithm, where at each iteration σ_f is evaluated by the lowest eigenvalue of the Hessian observed until that iteration.

For conducting numerical experiments, we generated synthetic data, as in Section 6.4 in [**?**]. We generate matrix R with given price ratios as: $R_{i,j} = 1 + N(0, 0.1)$ for any $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$, which allows the closing price to vary by 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 data sets in total. We display the performance of all our implemented methods using the aggregate statistics $\rho, \tilde{\rho}, \hat{\rho}$ in Figures **??**. Table **??** reports our findings for each individual data set. `BacktrackFW-GSC` outperforms all other methods considered in terms of time to reach a certain relative error, including `PN` and `PG`. For large relative error the empirical advantage of `BacktrackFW-GSC` over `FW-Line Search` seems to be small. However, when the relative error's tolerance is reduced to 10^{-5}, the difference is more pronounced, and for some data sets `BactrackFW-GSC` is three times faster than the closest competitor `FW-Line Search`. Moreover, despite its linear convergence, the `FW-LLOO` performance is worse then that of `BacktrackFW-GSC`, indicating the strong convexity parameter σ_f here is very small resulting in a large convergence coefficient.

6.2 Distance weighted discrimination

In the context of binary classification, an interesting modification of the classical support-vector machine is the distance weighted discrimination (DWD) problem, introduced in [**?**]. In that problem, the classification loss attains the form

$$f(x) = \frac{1}{n} \sum_{i=1}^{p} (a_i^\top w + \mu y_i + \xi_i)^{-q} + c^\top \xi$$

over the compact set

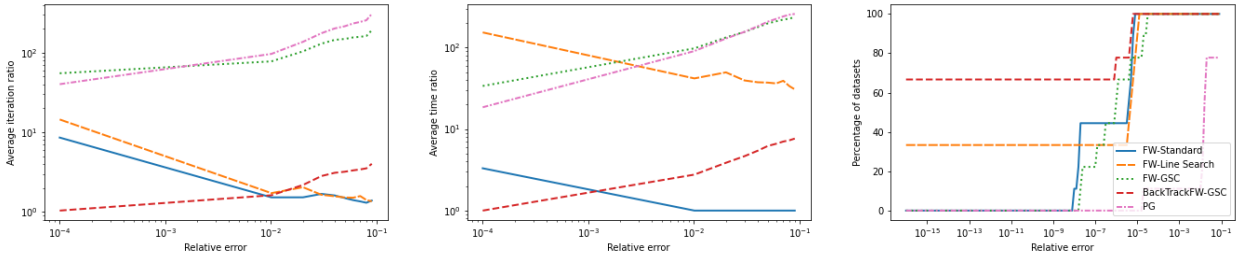
$$\mathcal{X} = \{x = (w, \mu, \xi) | \ \|w\|^2 \leq 1, \mu \in [-u, u], \|\xi\|^2 \leq R, \xi \in \mathbb{R}^p_+\},$$

where $R > 0$ is a hyperparameter that has to be learned via cross-validation.

Here $q \geq 1$ is a parameter to calibrate the statistical loss, and $(a_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}, (i = 1, 2, \ldots, p)$ are observed data. The decision variable is decoded as $x = (w, \mu, \xi) \in \mathbb{R}^n \cong \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p$, corresponding to a normal vector $w \in \mathbb{R}^d$, an intercept $\mu \in \mathbb{R}$ and a slack variable $\xi \in \mathbb{R}^p$. Since $\varphi(t) = t^{-q}, q \geq 1$

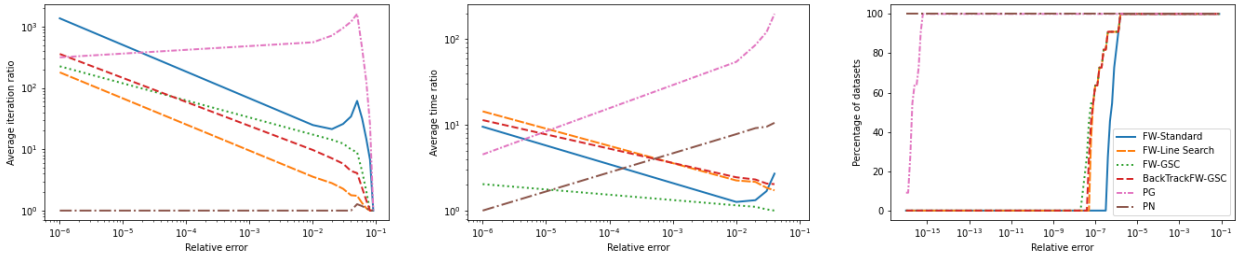
Problem			FW-Standard			FW-Line Search			BackTrackFW-GSC			FW-GSC			FW-LLOO			PN			PG		
name	p	n	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error
Relative error = 1e-03																							
syn_1000_800_10_50	1000	800	50	0.12	8.23e-04	5	0.03	2.81e-04	9	**0.02**	9.69e-04	201	0.64	9.98e-04	39	0.14	9.01e-04	5	3.84	1.36e-04	63	0.16	2.68e-04
syn_1000_800_10_50_1	1000	800	52	0.12	7.18e-04	8	0.05	2.10e-04	12	**0.03**	8.32e-04	263	0.78	9.98e-04	54	0.17	9.71e-04	5	4.33	1.99e-05	61	0.23	3.86e-04
syn_1000_800_10_50_2	1000	800	53	0.13	9.38e-04	8	0.05	8.96e-04	14	**0.05**	8.18e-04	268	0.71	9.99e-04	58	0.19	9.04e-04	5	1.20	1.41e-05	56	0.15	4.27e-05
syn_1000_800_10_50_3	1000	800	49	0.11	9.89e-04	7	0.05	2.87e-04	8	**0.02**	7.99e-04	183	0.49	9.93e-04	37	0.14	9.17e-04	5	4.49	2.49e-04	59	0.17	6.11e-04
syn_1000_1200_10_50	1000	1200	50	0.17	9.33e-04	6	0.05	4.80e-04	6	**0.02**	5.29e-04	6	**0.02**	2.66e-04	23	0.11	8.94e-04	6	17.36	1.28e-06	86	0.41	9.35e-04
syn_1000_1200_10_50_1	1000	1200	46	0.15	8.89e-04	5	0.04	1.26e-04	9	**0.03**	9.45e-04	198	0.79	9.99e-04	41	0.21	8.87e-04	5	3.84	7.10e-05	79	0.35	9.44e-04
syn_1000_1200_10_50_2	1000	1200	44	0.15	8.88e-04	4	0.03	6.31e-05	6	**0.02**	4.12e-04	127	0.52	9.93e-04	26	0.12	7.93e-04	6	4.96	1.36e-06	86	0.48	6.43e-04
syn_1000_1200_10_50_3	1000	1200	52	0.18	9.68e-04	8	0.09	3.13e-04	12	**0.04**	8.06e-04	245	0.97	9.94e-04	51	0.28	9.12e-04	5	5.42	2.28e-05	70	0.33	7.60e-04
syn_1000_1500_10_50	1000	1500	49	0.23	9.71e-04	6	0.08	1.94e-04	14	**0.07**	7.01e-04	220	1.13	9.98e-04	46	0.35	9.27e-04	5	7.30	4.84e-05	82	0.39	8.52e-04
syn_1000_1500_10_50_1	1000	1500	49	0.22	9.71e-04	6	0.08	1.94e-04	14	**0.06**	7.01e-04	220	1.23	9.98e-04	46	0.30	9.27e-04	5	7.97	4.84e-05	82	0.44	8.52e-04
syn_1000_1500_10_50_2	1000	1500	50	0.24	8.64e-04	7	0.10	4.94e-04	16	**0.07**	8.38e-04	250	1.24	9.97e-04	53	0.36	8.75e-04	5	6.04	2.72e-05	74	0.34	7.75e-04
syn_1000_1500_10_50_3	1000	1500	47	0.22	9.48e-04	5	0.06	2.01e-04	8	**0.03**	8.56e-04	201	1.04	9.97e-04	43	0.28	8.61e-04	5	13.11	1.55e-04	80	0.49	5.92e-05
Relative error = 1e-05																							
syn_1000_800_10_50	1000	800	448	1.09	9.65e-06	9	0.07	8.91e-06	20	**0.05**	7.71e-06	18823	52.62	1.00e-05	300	1.10	9.10e-06	6	4.01	3.54e-08	72	0.18	9.92e-06
syn_1000_800_10_50_1	1000	800	470	1.10	9.28e-06	35	0.27	9.42e-06	26	**0.07**	8.62e-06	25142	65.95	1.00e-05	458	1.55	9.95e-06	6	4.50	0.00e+00	70	0.27	5.52e-06
syn_1000_800_10_50_2	1000	800	477	1.16	9.01e-06	13	**0.09**	8.71e-06	35	0.11	8.00e-06	25660	67.35	1.00e-05	482	1.62	9.74e-06	6	1.42	0.00e+00	58	0.16	6.45e-06
syn_1000_800_10_50_3	1000	800	447	1.05	9.74e-06	18	0.16	5.53e-06	16	**0.04**	7.06e-06	17079	45.48	1.00e-05	243	0.98	9.78e-06	6	4.71	2.93e-08	78	0.25	9.10e-06
syn_1000_1200_10_50	1000	1200	448	1.59	9.89e-06	19	0.24	8.92e-06	12	**0.04**	4.50e-06	28	0.10	9.15e-06	116	0.60	9.23e-06	6	17.36	1.28e-06	117	0.57	9.11e-06
syn_1000_1200_10_50_1	1000	1200	442	1.58	9.58e-06	9	0.10	8.44e-06	21	**0.09**	6.45e-06	18571	74.16	1.00e-05	333	2.19	9.82e-06	6	4.37	7.21e-08	92	0.42	8.68e-06
syn_1000_1200_10_50_2	1000	1200	430	1.53	9.96e-06	5	0.04	7.29e-06	10	**0.03**	7.27e-06	11631	46.13	1.00e-05	145	0.73	9.75e-06	6	4.96	1.36e-06	103	0.58	7.76e-06
syn_1000_1200_10_50_3	1000	1200	478	1.68	9.92e-06	22	0.33	9.51e-06	27	**0.10**	9.96e-06	23293	92.17	1.00e-05	441	2.43	9.78e-06	6	6.08	0.00e+00	80	0.39	8.63e-06
syn_1000_1500_10_50	1000	1500	498	2.40	9.88e-06	13	0.22	9.49e-06	35	**0.20**	9.61e-06	20899	112.65	1.00e-05	373	2.61	9.57e-06	6	7.69	1.22e-07	96	0.48	8.96e-06
syn_1000_1500_10_50_1	1000	1500	498	2.29	9.88e-06	13	0.22	9.49e-06	35	**0.17**	9.61e-06	20899	108.46	1.00e-05	373	2.53	9.57e-06	6	8.34	1.22e-07	96	0.54	8.96e-06
syn_1000_1500_10_50_2	1000	1500	465	2.24	9.86e-06	23	0.43	9.52e-06	43	**0.21**	8.70e-06	23692	122.04	1.00e-05	446	3.17	9.88e-06	6	6.84	0.00e+00	85	0.40	9.75e-06
syn_1000_1500_10_50_3	1000	1500	445	2.15	9.66e-06	7	0.10	6.28e-06	18	**0.08**	5.24e-06	18797	93.26	1.00e-05	314	2.15	9.99e-06	6	13.44	0.00e+00	87	0.54	8.74e-06

Table 2: Results for portfolio selection problem (**??**). Number of iterations and CPU time in seconds to achieve a certain relative error or best relative error achieved by methods, as well as the relative error achieved at that iteration. We highlight in bold the best performance among all competitors.



(a) Average iteration ratio $\tilde{\rho}(\varepsilon)$. (b) Average time ratio $\hat{\rho}(\varepsilon)$. (c) Average success ratio $\rho(\varepsilon)$.

Figure 2: Performance Profile for the DWD problem averaged over binary classification problems.



(a) Average iteration ratio $\tilde{\rho}(\varepsilon)$. (b) Average time ratio $\hat{\rho}(\varepsilon)$. (c) Average success ratio $\rho(\varepsilon)$.

Figure 3: Performance Profile for Covariance estimation problem (**??**).

is generalized self-concordant with parameters $M_\varphi = \frac{q+2}{\sqrt[q+2]{q(q+1)}}$ and $\nu = \frac{2(q+3)}{q+2} \in (2,3)$ (cf. Table **??**) we get a GSC minimization problem over the compact set $\mathcal{X}$, with parameters $\nu = \frac{2(q+3)}{q+2}$ and $M_f = \frac{q+2}{\sqrt[q+2]{q(q+1)}} n^{1/(q+2)} \max\left\{\|(a_i^\top, y_i, e_i^\top)^\top\|_2^{q/(q+2)} : 1 \le i \le n\right\}$. The special case $q = 1$ is the loss function of [**?**] who solved this problem via a second-order cone reformulation. We test our algorithms using $q = 2$ where a_i and y_i are based on data sets a1a-a9a from the LIBSVM library [**?**], where a_i are normalized. We set $c_i = 1$ for all $i = 1, \ldots, p$, $u = 5$, and $R = 10$.

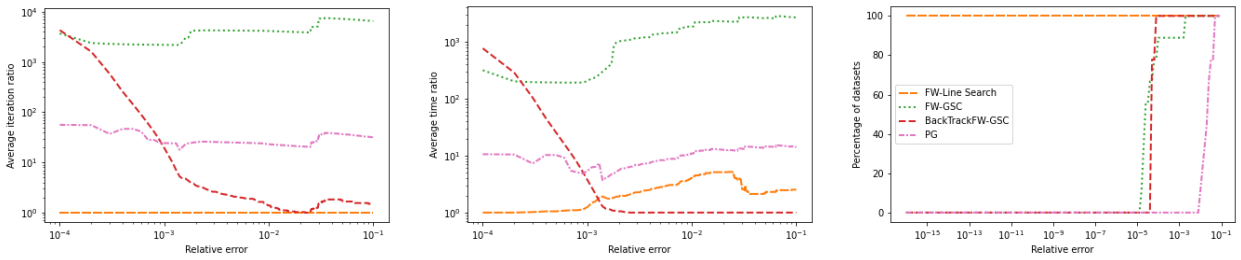
We note that we were not able to run `PN` on this example, and that the `PG` generally had the worst performance of all methods compared. Figure **??** collects results on the average performance of our methods and Table **??** shows the results obtained for each individual data set. Inspecting the realized performance values, a qualitatively similar picture to the Portfolio selection problem emerges: Across all instances investigated, `BacktrackFW-GSC` is the best performing algorithm. For large relative error the empirical advantage of our methods are not as pronounced. However, for more precise solutions with a smaller relative error the best method, `BacktrackFW-GSC`, works very well for the DWD problem. In some instances we see that the method runs 10 to 11 times faster than the best competitor `FW-standard`. It is important to note that since the DWD problem does not posses a Lipschitz continuous gradient, while `FW-standard` is the closest empirical competitor it comes with no theoretical guarantees on its performance.

6.3 Inverse covariance estimation

Undirected graphical models offer a way to describe and explain the relationships among a set of variables, a central element of multivariate data analysis. The principle of parsimony dictates that we should select the simplest graphical model that adequately explains the data. The typical

Problem			FW-Standard			FW-Line Search			BackTrackFW-GSC			FW-GSC			PG		
name	p	n	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error
Relative error = 1e-03																	
a1a	128	1605	109	0.03	9.83e-04	364	2.34	8.97e-04	39	**0.02**	7.08e-04	1140	0.50	9.90e-04	763	0.25	8.34e-04
a2a	128	2265	126	0.04	9.87e-04	401	4.17	8.12e-04	40	**0.03**	5.42e-04	1343	0.70	9.98e-04	1181	0.48	8.41e-04
a3a	128	3185	157	0.06	9.93e-04	35	0.33	8.99e-04	48	**0.04**	6.63e-04	1577	0.97	9.84e-04	1243	0.60	9.37e-04
a4a	128	4781	127	0.07	9.86e-04	79	1.21	4.93e-04	44	**0.06**	6.77e-04	1919	1.57	9.81e-04	1630	1.05	7.11e-04
a5a	128	6414	111	**0.07**	9.86e-04	219	3.82	6.50e-04	49	0.08	9.16e-04	2209	2.21	9.94e-04	50001	40.15	4.81e-03
a6a	128	11220	127	0.13	9.88e-04	180	4.83	5.56e-04	47	**0.12**	7.29e-04	2897	4.96	9.91e-04	3303	3.74	7.70e-04
a7a	128	16100	127	0.17	9.87e-04	317	12.43	8.74e-04	46	**0.17**	5.39e-04	3457	7.82	9.95e-04	3666	5.72	6.41e-04
a8a	128	22696	127	**0.22**	9.86e-04	390	19.38	3.78e-04	47	**0.22**	7.16e-04	4089	12.27	9.98e-04	6967	14.01	9.75e-04
a9a	128	32561	127	0.39	9.86e-04	115	8.26	9.95e-04	50	**0.37**	8.97e-04	4876	22.94	9.99e-04	8770	24.23	8.03e-04
Relative error = 1e-05																	
a1a	128	1605	1086	0.30	9.99e-06	565	3.66	1.14e-05	71	**0.04**	9.89e-06	1210	0.52	8.65e-06	770	0.25	9.74e-06
a2a	128	2265	1940	0.64	1.00e-05	770	7.45	9.33e-06	86	**0.06**	1.00e-05	1433	0.74	9.15e-06	50001	21.22	1.73e-05
a3a	128	3185	2041	0.81	1.00e-05	44	0.41	9.33e-06	77	**0.07**	9.98e-06	1676	1.04	9.79e-06	50001	24.97	1.51e-05
a4a	128	4781	1266	0.67	9.99e-06	104	1.81	8.91e-06	63	**0.09**	9.07e-06	2033	1.66	9.51e-06	50001	32.84	4.84e-04
a5a	128	6414	2113	1.41	1.00e-05	442	7.64	7.72e-06	65	**0.11**	8.84e-06	2340	2.34	9.78e-06	50001	40.15	4.81e-03
a6a	128	11220	1267	1.30	9.99e-06	824	22.16	9.49e-06	58	**0.15**	8.21e-06	3068	5.23	9.94e-06	*44191	49.36	1.33e-05
a7a	128	16100	1939	2.71	1.00e-05	522	21.48	9.70e-06	61	**0.24**	8.01e-06	3664	8.28	9.58e-06	50001	76.77	4.97e-04
a8a	128	22696	1266	2.27	9.99e-06	3670	183.00	9.62e-06	58	**0.28**	9.73e-06	4333	13.00	9.49e-06	50001	102.05	5.15e-04
a9a	128	32561	1562	4.44	9.99e-06	159	11.38	9.34e-06	60	**0.43**	9.30e-06	5167	24.38	9.67e-06	*18589	52.99	7.49e-04

Table 3: Results for distance weighted discrimination (DWD) problem. Number of iterations and CPU time in seconds to achieve a certain relative error or best relative error achieved by method after 50,000 iterations, as well as the relative error achieved at that iteration. * The algorithm did not change the solution from this point onward although the optimal solution was not reached.



(a) Average iteration ratio $\tilde{\rho}(\varepsilon)$. (b) Average time ratio $\hat{\rho}(\varepsilon)$. (c) Average success ratio $\rho(\varepsilon)$.

Figure 4: Performance Profile for Signal Reconstruction using KL divergence problem (**??**).

approach to tackle this problem is the following: Given a data set, we solve a maximum likelihood problem with an added low-rank penalty to make the resulting graph as sparse as possible. We consider learning a Gaussian graphical random field of p nodes/variables from a data set $\{\phi_1, \ldots, \phi_N\}$. Each random vector ϕ_j is an iid realization from a p-dimensional Gaussian distribution with mean μ and covariance matrix Σ. Let $\Theta = \Sigma^{-1}$ be the precision matrix. To satisfy conditional dependencies between the random variables, Θ must have zero in Θ_{ij} if i and j are not connected in the underlying graphical model. To learn the graphical model via an ℓ_1-regularization framework in its constrained formulation, we minimize the loss function

$$f(x) = -\log\det(\mathrm{mat}(x)) + \mathrm{tr}(\hat{\Sigma}\,\mathrm{mat}(x)) \tag{6.2}$$

over the ℓ_1-ball $\mathcal{X} = \{x \in \mathbb{R}^n | \, \|x\|_1 \leq 1\}$. The decision variables are vectors $x \in \mathbb{R}^n$ for $n = p^2$, so that $\mathrm{mat}(x)$ represents the $p \times p$ matrix constructed from the p^2-dimensional vector x. It can be seen that f is standard self-concordant with domain $\mathbb{S}^n_{++}$. Hence, $M_f = 2$ and $\nu = 3$. One can see that the gradient $\nabla f(x) = \hat{\Sigma} - \mathrm{mat}(x)^{-1}$ and Hessian $\nabla^2 f(x) = \mathrm{mat}(x)^{-1} \otimes \mathrm{mat}(x)^{-1}$. Since $\mathrm{mat}(x)$ is positive definite, we can compute the inverse via a Cholesky decomposition, which in the worst case needs $O(p^3)$ arithmetic steps. To compute the search direction, we have to solve the LP

$$\min_{s \in \mathcal{X}} \langle \hat{\Sigma} - \mathrm{mat}(x)^{-1}, s \rangle$$

which leads to identify a leading singular value. We test our method on synthetically generated data sets. We generated the data by first creating a the matrix $\hat{\Sigma}$ randomly, by generating a random orthonormal basis or $\mathbb{R}^p$, $B = \{v_1, \ldots, v_p\}$, and then set

$$\hat{\Sigma} = \sum_{i=1}^{p} \sigma_i v_i v_i^\top,$$

where σ_i are independently and uniformly distributed between 0.5 and 1. We generated 10 such data sets, for p ranging between 50 and 300. Figure **??** collects results on the average performance of our methods and Table **??** shows the results obtained for each individual data set. We observe that `FW-GSC` generally has the best performance for higher relative error, as well as for lower relative error when $p \leq 150$. The closest competitors are `FW-standard` for higher relative errors, and `PN` for the lower relative errors, which for larger value of p has slightly superior running times.

6.4 Signal retrieval using KL divergence

Consider a signal retrieval problem, where the noisy measurements y are obtained from signal θ passing through a linear transformation $\mathbf{W}$. One approach to retrieve the original signal θ consists

Problem		FW-Standard			FW-Line Search			BackTrackFW-GSC			FW-GSC			PN			PG		
name	p	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error
Relative error = 1e-03																			
cov_50	50	328	**0.05**	9.97e-04	92	0.18	9.93e-04	196	0.16	9.94e-04	220	**0.05**	9.99e-04	12	0.31	9.03e-04	532	0.27	9.36e-04
cov_80	80	487	0.17	9.99e-04	113	0.42	9.93e-04	241	0.37	9.95e-04	316	**0.11**	9.97e-04	16	0.70	9.24e-04	2622	2.13	9.95e-04
cov_120	120	714	0.39	9.98e-04	140	0.91	9.96e-04	398	1.03	9.94e-04	395	**0.29**	9.96e-04	17	1.12	5.85e-04	3909	5.81	9.42e-04
cov_150	150	880	0.73	9.97e-04	173	1.58	9.99e-04	356	1.43	9.97e-04	502	**0.57**	9.96e-04	20	2.24	6.69e-04	7446	16.37	9.92e-04
cov_170	170	986	1.04	1.00e-03	165	1.92	9.96e-04	503	3.36	9.96e-04	537	**0.78**	9.95e-04	20	2.90	8.95e-04	6688	19.72	9.83e-04
cov_200	200	1140	4.52	1.00e-03	216	8.32	9.97e-04	475	3.53	1.00e-03	636	**1.74**	1.00e-03	21	5.45	7.23e-04	11043	57.18	9.83e-04
cov_220	220	1237	**2.06**	1.00e-03	290	6.86	1.00e-03	590	5.51	9.98e-04	766	2.40	9.97e-04	17	4.95	9.62e-04	4856	26.98	9.88e-04
cov_250	250	1375	3.69	1.00e-03	298	10.24	9.98e-04	827	16.53	9.99e-04	850	**3.57**	9.98e-04	22	11.04	8.03e-04	14102	132.68	9.92e-04
cov_270	270	1466	5.09	9.99e-04	241	14.77	9.96e-04	577	9.98	9.98e-04	767	**4.49**	9.96e-04	21	12.40	9.19e-04	12694	122.53	9.85e-04
cov_300	300	1603	11.75	1.00e-03	315	17.69	9.95e-04	674	20.64	9.98e-04	965	**6.84**	9.99e-04	23	19.83	8.10e-04	13148	147.81	9.99e-04
Relative error = 1e-05																			
cov_50	50	3487	0.59	9.91e-06	590	1.16	9.97e-06	1340	1.10	9.99e-06	764	**0.19**	1.00e-05	15	0.42	2.04e-07	542	0.28	2.61e-06
cov_80	80	5022	1.31	9.98e-06	812	2.98	9.99e-06	1321	2.02	9.99e-06	1088	**0.43**	9.97e-06	20	0.92	1.88e-07	2646	2.15	6.89e-06
cov_120	120	7101	5.49	9.99e-06	1277	8.70	9.99e-06	1871	4.96	1.00e-05	1674	**1.14**	9.99e-06	20	1.38	2.44e-06	3952	5.87	4.95e-06
cov_150	150	8697	6.96	1.00e-05	1636	15.03	9.99e-06	3165	13.01	9.99e-06	2118	**2.42**	9.99e-06	24	2.87	1.62e-07	7509	16.51	9.00e-06
cov_170	170	9703	10.70	1.00e-05	1262	14.86	9.99e-06	1929	11.55	1.00e-05	1810	**2.57**	9.98e-06	24	3.90	1.88e-07	6765	19.93	8.71e-06
cov_200	200	11224	42.42	1.00e-05	2022	80.55	9.99e-06	4857	40.75	1.00e-05	2742	7.18	1.00e-05	25	**6.76**	8.04e-07	11141	58.14	8.53e-06
cov_220	220	12277	19.92	1.00e-05	2687	61.63	9.99e-06	4610	42.74	1.00e-05	3362	10.40	9.99e-06	21	**6.61**	6.30e-06	4967	27.54	8.86e-06
cov_250	250	13750	33.75	1.00e-05	2625	106.45	1.00e-05	4164	81.15	1.00e-05	3529	14.39	9.99e-06	26	**13.44**	8.70e-06	14230	133.56	9.60e-06
cov_270	270	14759	49.47	1.00e-05	2374	126.03	9.99e-06	5917	110.87	1.00e-05	3040	16.54	9.99e-06	26	**16.22**	5.12e-07	12861	124.23	9.32e-06
cov_300	300	16217	103.49	1.00e-05	2563	147.75	9.99e-06	5208	141.58	1.00e-05	3621	**24.83**	1.00e-05	27	26.20	6.01e-06	13332	149.67	9.99e-06

Table 4: Results for covariance estimation example. Number of iterations and CPU time in seconds to achieve a certain relative error or best relative error achieved by methods, as well as the relative error achieved at that iteration.

of minimizing some distance between the observed measurements y and the signal transformation $\mathbf{W}\theta$. Additional information, such as sparsity of the signal can be captured by constraining the ℓ_1 norm of θ to be no larger than R. Thus, we are tasked to minimize the following optimization problem:

$$\min_{\theta \in \mathbb{R}^n_+, \|\theta\|_1 \leq R} \{D_\phi(\mathbf{W}\theta, y)\}.$$

When $D_\phi(\mathbf{W}\theta, y)$ is the Euclidean distance we retrieve the well known constrained linear regression. However, other distance measures can be more appropriate when $\mathbf{W}$, θ and y are all assumed to be nonnegative. One such case is when $D_\phi(\mathbf{W}\theta, y)$ is the Kullback–Leibler (KL) divergence which measure the residuals between two nonnegative points (see e.g., [**?**], and the references therein) given by

$$D_\phi(\mathbf{W}\theta, y) \triangleq \sum_{i=1}^{N} \left\{ \langle w_i, \theta \rangle \log\left(\frac{\langle w_i, \theta \rangle}{y_i} \right) - \langle w_i, \theta \rangle \right\} + \sum_{i=1}^{N} y_i.$$

Ignoring the constant $\sum_{i=1}^{N} y_i$, we see that the optimization problem involves the function

$$g(\theta) \triangleq \sum_{i=1}^{N} \left\{ \langle w_i, \theta \rangle \log\left(\frac{\langle w_i, \theta \rangle}{y_i} \right) - \langle w_i, \theta \rangle \right\}.$$

Introduce the map $\varphi_c(t) \triangleq t \log(t/c) - t$ for $t, c > 0$, so that

$$g(\theta) = \sum_{i=1}^{N} \varphi_{b_i}(\langle w_i, \theta \rangle).$$

It can be easily checked that φ_{b_i} is $(1, 4)$-generalized self-concordant, with closed domain $[0, \infty)$. Therefore, we arrive at a GSC optimization problem of the form

$$\min_{\theta \in \mathbb{R}^d_+, \|\theta\|_1 \leq R} g(x).$$

We test our algorithms with $\mathbf{W}$ and y taken from data sets a1a-a9a from the LIBSVM library [**?**]. In all experiments we set $R = 30$ and all values of coordinates of y are set to 1.

Figure **??** collects results on the average performance of our methods and Table **??** shows the results obtained for each individual data set. It is interesting to note that both `FW-standard` and `PN` failed to run in for this example. This indicates that though `FW-standard` may perform well in some applications, there is no guarantee that the step-size chosen by this method would result in iterates which remain in the function's domain, and therefore it is both theoretically and in some cases practically ill-defined for GSC functions. For these data sets we see that `BackTrackFW-GSC` has the best performance for higher values of relative error, whereas `FW-Line Search` has significantly superior performance to all other methods for lower values of relative error. Moreover, similar to the case of DWD (in which `PN` failed to run as well) the performance of `PG` significantly deteriorates for lower values of relative errors, which we believe is due to the same numerical issues that cause `PN` to fail.

Problem			FW-Line Search			BackTrackFW-GSC			FW-GSC			PG		
name	p	n	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error	iter	time[s]	error
Relative error = 1e-02														
a1a	128	1605	4	0.007	2.27e-03	5	**0.003**	9.70e-03	2339	0.45	9.99e-03	35	0.02	9.39e-03
a2a	128	2265	4	0.009	2.00e-03	6	**0.002**	5.56e-03	3300	0.78	9.98e-03	54	0.03	9.71e-03
a3a	128	3185	4	0.009	2.19e-03	5	**0.004**	9.89e-03	4700	1.43	9.99e-03	60	0.04	9.87e-03
a4a	128	4781	4	0.014	1.86e-03	7	**0.009**	9.22e-03	7052	3.77	9.98e-03	70	0.06	9.35e-03
a5a	128	6414	4	0.026	1.59e-03	5	**0.006**	9.37e-03	9439	8.55	1.00e-02	75	0.08	9.16e-03
a6a	128	11220	4	0.043	1.55e-03	6	**0.018**	8.92e-03	16534	26.56	1.00e-02	110	0.18	9.79e-03
a7a	128	16100	4	0.063	1.55e-03	5	**0.015**	7.62e-03	23727	55.82	1.00e-02	113	0.24	9.60e-03
a8a	128	22696	4	0.089	1.56e-03	6	**0.031**	8.97e-03	33582	106.81	1.00e-02	148	0.53	9.72e-03
a9a	128	32561	4	0.153	1.62e-03	5	**0.032**	7.55e-03	48203	293.88	1.00e-02	168	0.65	9.87e-03
Relative error = 1e-04														
a1a	128	1605	6	**0.01**	8.42e-05	26353	12.63	1.00e-04	10949	2.04	1.00e-04	50001	33.82	2.45e-04
a2a	128	2265	6	**0.02**	7.03e-05	1965	1.22	1.00e-04	11375	2.54	1.00e-04	50001	34.91	3.15e-04
a3a	128	3185	6	**0.02**	7.51e-05	26986	21.94	1.00e-04	13594	4.35	1.00e-04	337	0.25	3.80e-05
a4a	128	4781	6	**0.03**	6.13e-05	28223	38.65	1.00e-04	15283	8.24	1.00e-04	50001	49.87	3.45e-04
a5a	128	6414	6	**0.05**	4.51e-05	26449	48.10	1.00e-04	18014	18.21	1.00e-04	50001	59.21	6.23e-04
a6a	128	11220	6	**0.09**	4.56e-05	30646	91.76	1.00e-04	26588	44.07	1.00e-04	50001	86.55	5.27e-04
a7a	128	16100	6	**0.13**	4.60e-05	31325	140.75	1.00e-04	34810	81.70	1.00e-04	50001	107.00	6.30e-04
a8a	128	22696	6	**0.18**	4.55e-05	30026	202.43	1.00e-04	46845	152.62	1.00e-04	50001	178.30	7.13e-04
a9a	128	32561	6	**0.31**	4.65e-05	30747	279.08	1.00e-04	50001	306.22	1.74e-03	50001	209.73	6.54e-04

Table 5: Results for KL example. Number of iterations and CPU time in seconds to achieve a certain relative error or best relative error achieved by methods, as well as the relative error achieved at that iteration.

7 Conclusion

Motivated by the recent interest in computational statistics and machine learning in functions displaying generalized self-concordant properties, this paper develops a set of projection-free algorithms for minimizing generalized self-concordant functions as defined in [**?**]. This function class covers several well-known examples, including logistic, power, reciprocal and, of course, standard self-concordant functions. In particular, members of this function class are potentially ill-conditioned: they may neither have a Lipschitz continuous gradient nor be strongly convex. Hence, no provably convergent Frank-Wolfe method has been available so far for minimizing generalized self-concordant functions. This paper fills this important gap by developing two provably convergent FW algorithms with sublinear convergence rates. Tightening the definition of the linear minimization oracle, we derive a projection-free method with linear convergence rate for the entire class of generalized self-concordant functions. With the help of extensive numerical experiments, we demonstrate the practical efficiency of our approach.

We conclude by mentioning some interesting extensions of the approach presented in this work. First, our theory can be used to develop distributed versions of the algorithms presented in this paper in order to develop a generalized version of the DISCO algorithm [**?**]. Second, it will be interesting to incorporate gradient sliding techniques [**?**], and stochastic versions of our algorithms. Recently, a Newton Frank-Wolfe method has been introduced in [**?**]. It seems natural to us that their algorithm can be extended to GSC functions. All these are important extensions we are planning to pursue in the near future.

Acknowledgments.

The authors sincerely thank Professor Shoham Sabach for his contribution in the early stages of this project, including his part in developing the basic ideas developed in this paper. We would also like to thank Professor Quoc Tranh-Dinh for sharing MATLAB codes on SCOPT with us. This research is supported by the COST Action CA16228 "European Network for Game Theory".

A Additional Facts about GSC functions

In order to make this paper self-contained we are collecting in this appendix finer estimates provided by self-concordance. For a complete treatise the reader should consult the seminal paper [**?**].

An important feature of GSC functions is their invariance under affine transformations. This is made precise in the following Lemma.

Lemma A.1 ([**?**], Prop. 2)**.** *Let $f \in \mathcal{F}_{M_f,\nu}(\operatorname{dom} f)$ and $\mathcal{A}(x) = Ax + b : \mathbb{R}^n \to \mathbb{R}^p$ a linear operator. Then*

(a) If $\nu \in [2,3]$, then $\tilde{f}(x) \triangleq f(\mathcal{A}(x))$ is $(M_{\tilde{f}}, \nu)$-GSC with $M_{\tilde{f}} = M_f \|A\|^{3-\nu}$.

(b) If $\nu > 3$ and $\lambda_{\min}(A^\top A) > 0$, then $\tilde{f}(x) = f(\mathcal{A}(x))$ is $(M_{\tilde{f}}, \nu)$-GSC with $M_{\tilde{f}} = M_f \lambda_{\min}(A^\top A)^{\frac{3-\nu}{2}}$, where $\lambda_{\min}(A^\top A)$ is the smallest eigenvalue of $A^\top A$.

When we apply 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 (**??**). 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 (**??**). 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 $x = A\tilde{x} \in \mathcal{X}, s = A\hat{s} \in \mathcal{X}$.

Beside affine invariance, we will use some stability properties of GSC functions.

Proposition A.2 ([**?**], Prop. 1)**.** *Let $f_i \in \mathcal{F}_{M_{f_i},\nu}(\operatorname{dom} f_i)$ where $M_{f_i} \geq 0$ and $\nu \geq 2$ for $i = 1, \ldots, N$. Then, given scalars $w_i > 0, 1 \leq i \leq N$, the function $f \triangleq \sum_{i=1}^N w_i f_i$ is well defined on $\operatorname{dom} f \triangleq \bigcap_{i=1}^N \operatorname{dom} f_i$ and belongs to $\mathcal{F}_{M_f,\nu}(\operatorname{dom} f)$, where $M_f \triangleq \max_{1 \leq i \leq N} w_i^{1-\frac{\nu}{2}} M_{f_i}$.*

As corollary of this Proposition and invariance under linear transformations, we obtain the next characterization theorem, which is of particular importance in machine learning applications.

Given N functions $\varphi_i \in \mathcal{F}_{M_{\varphi_i},\nu}(\operatorname{dom} \varphi_i)$. For $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}, q \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}$ a positive definite and symmetric matrix, consider the finite-sum model

$$f(x) \triangleq \sum_{i=1}^N \varphi_i(\langle a_i, x \rangle + b_i) + \langle q, x \rangle + \frac{1}{2}\langle Qx, x \rangle \tag{A.1}$$

Proposition A.3 ([**?**], Prop. 5)**.** *If $\varphi_i \in \mathcal{F}_{M_{\varphi_i},\nu}(\operatorname{dom} \varphi_i)$ for $\nu \in (0, 3]$, then $f : \mathbb{R}^n \to (-\infty, \infty]$ defined in* (**??**) *belongs to $\mathcal{F}_{M_f,3}(\operatorname{dom} f)$, where $M_f \triangleq \lambda_{\min}(Q)^{(\nu-3)/2} \max_{1 \leq i \leq N} M_{\varphi_i} \|a_i\|_2^{3-\nu}$.*

B Proof of Theorem ??

The proof of Theorem **??** is an application of the technical Lemma below.

Lemma B.1. *Consider the function*

$$\psi_\nu(t) \triangleq t - \xi \omega_\nu(t\delta) t^2, \tag{B.1}$$

where $\xi, \delta \geq 0$ are parameters and $\nu \geq 2$. For all $\nu \geq 2$, the function $t \mapsto \psi_\nu(t)$ is concave and differentiable. The unique maximum of this function is achieved at

$$t_\nu^* \triangleq \begin{cases} \frac{1}{\delta} \ln\left(1 + \frac{\delta}{\xi}\right) & \text{if } \nu = 2, \\ \frac{1}{\delta}\left[1 - \left(1 + \frac{\delta}{\xi}\frac{4-\nu}{\nu-2}\right)^{-\frac{\nu-2}{4-\nu}}\right] & \text{if } \nu \in (2,3) \cup (3,4), \\ \frac{1}{\delta+\xi} & \text{if } \nu = 3, \\ \frac{1}{\delta}\left[1 - \exp\left(-\frac{\delta}{\xi}\right)\right] & \text{if } \nu = 4. \end{cases} \tag{B.2}$$

Proof. We will organize the proof of Lemma **??** according to the generalized self-concordance parameter $\nu \in [2, 4]$.

The case $\nu = 2$: For this parameter we have

$$\omega_2(t) = \frac{1}{t^2}[e^t - t - 1],$$

and

$$\psi_2(t) = t - \frac{\xi}{\delta^2}[e^{t\delta} - t\delta - 1].$$

This is a strictly concave function with unique maximum at

$$t_2^* = \frac{1}{\delta} \ln\left(1 + \frac{\delta}{\xi}\right). \tag{B.3}$$

The case $\nu \in (2,3) \cup (3,4)$**:** Some simple algebra shows that in this case

$$\psi_\nu(t) = t\left(1 + \frac{\xi}{\delta}\frac{\nu-2}{4-\nu}\right) - \frac{\xi}{\delta^2}\frac{(\nu-2)^2}{2(3-\nu)(4-\nu)}\left[(1-t\delta)^{\frac{2(3-\nu)}{2-\nu}} - 1\right].$$

Setting $\psi'_\nu(t) = 0$, yields the value

$$t^*_\nu = \frac{1}{\delta}\left[1 - \left(1 + \frac{\delta}{\xi}\frac{4-\nu}{\nu-2}\right)^{-\frac{\nu-2}{4-\nu}}\right].$$

It is easy to check that $\psi''_\nu(t) = -\xi(1-t\delta)^{\frac{2}{2-\nu}} < 0$ so that t^* is the global maximum of $\psi_\nu(t)$.

The case $\nu = 3$**:** For this case, it is easy to see that

$$\psi_3(t) = t + \frac{\xi}{\delta^2}[t\delta + \ln(1-t\delta)] \qquad t \in (0, 1/\delta).$$

Therefore, for $t \in (0, 1/\delta)$, we see that

$$\psi'_3(t) = 1 + \frac{\xi}{\delta^2}\left(\delta - \frac{\delta}{1-t\delta}\right), \text{ and } \psi''_3(t) = -\frac{\xi}{\delta}(1-t\delta)^{-2} < 0.$$

The unique maximum is attained at

$$t^*_3 = \frac{1}{\delta + \xi}.$$

The case $\nu = 4$**:** For this case we have

$$\psi_4(t) = t - \frac{\xi}{\delta^2}[t\delta + (1-t\delta)\ln(1-t\delta)] \qquad t \in (0, 1/\delta).$$

Therefore, fro $t \in (0, 1/\delta)$,

$$\psi'_4(t) = 1 - \frac{\xi}{\delta^2}\ln(1-t\delta), \text{ and } \psi''_4(t) = -\frac{\xi}{1-t\delta} < 0.$$

From here, it is easy to see that the unique maximum is attained at

$$t^*_4 = \frac{1}{\delta}\left[1 - \exp(-\delta/\xi)\right].$$

■

B.1 Proof of Theorem ??

Identifying the parameters involved in (**??**) as $\delta = \delta_\nu(x)$, and $\xi = \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$ gives us

$$\eta_{x,\nu}(t) = \mathsf{Gap}(x)\psi_\nu(t).$$

Hence, the following explicit expressions for the step-size parameters is immediate.

$\nu = 2$: Since $\delta_2(x) = M_f\beta(x)$ we get the relation

$$\mathtt{t}_2(x) = \frac{1}{M_f\beta(x)}\ln\left(1 + \frac{M_f\beta(x)}{\mathsf{e}(x)^2}\,\mathsf{Gap}(x)\right).$$

$\nu \in (2,3) \cup (3,4)$: Set $\delta = \delta_\nu(x) = \frac{\nu-2}{2} M_f \beta(x)^{3-\nu} \mathsf{e}(x)^{\nu-2}$ and $\xi = \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, we get

$$\mathtt{t}_\nu(x) = \frac{2}{\nu-2}\frac{1}{M_f}\beta(x)^{\nu-3}\mathsf{e}(x)^{2-\nu}\left[1-\left(1+\frac{4-\nu}{2}M_f\beta(x)^{3-\nu}\mathsf{e}(x)^{\nu-4}\,\mathsf{Gap}(x)\right)^{\frac{2-\nu}{4-\nu}}\right].$$

$\nu = 3$: Since $\delta_3(x) = \frac{M_f}{2}\mathsf{e}(x)$, we get

$$\mathtt{t}_3(x) = \frac{\mathsf{Gap}(x)}{\frac{M}{2}\mathsf{e}(x)(\frac{2}{M}\mathsf{e}(x) + \mathsf{Gap}(x))}$$

$\nu = 4$: Since $\delta_4(x) = M_f \beta(x)^{-1}\mathsf{e}(x)^2$, we get

$$\mathtt{t}_4(x) = \frac{\beta(x)}{M_f \mathsf{e}(x)^2}\left[1-\exp\left(-\frac{M_f\,\mathsf{Gap}(x)}{\beta(x)}\right)\right].$$

This completes the proof of Theorem **??**

C Auxiliary Results needed in the proof of Theorem ??

C.1 Proof of Lemma ??

Set $x \equiv x^k$. Since $\mathtt{t}_\nu(x) > 1$, the decrease on the objective function is

$$\eta_{x,\nu}(1) = \mathsf{Gap}(x)\left(1-\frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}\omega_\nu(\delta_\nu(x))\right).$$

If $\nu > 2$ we know that $\delta_\nu(x) \le \mathtt{t}_\nu(x)\delta_\nu(x) < 1$, and the expression above is well-defined. If $\nu = 2$, the domain of the function ω_2 is full, and again the expression above is well-defined. Set $\zeta_\nu(x) := \omega_\nu(t\delta_\nu(x))t^2$ and $\xi(x) := \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, so that

$$\frac{\eta_{x,\nu}(t)}{\mathsf{Gap}(x)} = 1 - \zeta_\nu(t)\xi(x),$$

where $t \in (0,\infty)$ if $\nu = 2$ and $t \in (0, 1/\delta_\nu(x))$ for $\nu \in (2,4]$. By definition, $\mathtt{t}_\nu(x)$ is the unconstrained maximizer of the right-hand-side above. Therefore, $1 - \xi(x)\zeta_\nu'(\mathtt{t}_\nu(x)) = 0$. Since $t \mapsto \zeta_\nu(t)$ is convex, its derivative is a non-decreasing function. Thus, since $1 < \mathtt{t}_\nu(x)$, we conclude $\xi(x) = \frac{1}{\zeta_\nu'(\mathtt{t}_\nu(x))} \le \frac{1}{\zeta_\nu'(1)}$. Moreover, $\zeta_\nu(1) \ge 0$, so that

$$\begin{aligned}\frac{\eta_{x,\nu}(1)}{\mathsf{Gap}(x)} &= 1 - \xi(x)\zeta_\nu(1) = 1 - \frac{\zeta_\nu(1)}{\zeta_\nu'(\mathtt{t}_\nu(x))} \ge 1 - \frac{\zeta_\nu(1)}{\zeta_\nu'(1)} \\ &= 1 - \frac{\omega_\nu(\delta_\nu(x))}{2\omega_\nu(\delta_\nu(x)) + \delta_\nu(x)\omega_\nu(\delta_\nu(x))} \\ &\ge \frac{1}{2}.\end{aligned}$$

where we used that $\omega_\nu'(t) \ge 0$ for $t > 0$.

C.2 Proof of Lemma ??

We first prove a general lower estimate on the per-iteration progress.

Lemma C.1. *Suppose that $\mathtt{t}_\nu(x^k) \le 1$. Then, the per-iteration progress in the objective function value is lower bounded by*

$$\Delta_k \ge \begin{cases} \frac{2\ln(2)-1}{\mathsf{e}(x^k)} \min\left\{\frac{\mathsf{Gap}(x^k)^2}{\mathsf{e}(x^k)}, \frac{\mathsf{e}(x)\,\mathsf{Gap}(x^k)}{M_f\beta(x^k)}\right\} & \text{if } \nu = 2, \\ \tilde{\gamma}_\nu \min\left\{\frac{\mathsf{Gap}(x^k)}{\frac{\nu-2}{2}\beta(x^k)^{3-\nu}\mathsf{e}(x^k)^{\nu-2}}, \frac{-1}{\mathsf{b}}\frac{\mathsf{Gap}(x^k)^2}{\mathsf{e}(x^k)^2}\right\} & \text{if } \nu \in (2,3)\cup(3,4), \\ \frac{2(1-\ln(2))}{M_f\mathsf{e}(x^k)} \min\left\{\mathsf{Gap}(x^k), \frac{M_f\,\mathsf{Gap}(x^k)^2}{\mathsf{e}(x^k)}\right\} & \text{if } \nu = 3, \\ \frac{\exp(-1)\beta(x)}{M_f\mathsf{e}(x)^2} \min\left\{\mathsf{Gap}(x), \frac{M_f\,\mathsf{Gap}(x)^2}{\beta(x)}\right\} & \text{if } \nu = 4 \end{cases} \tag{C.1}$$

where $\tilde{\gamma}_\nu \triangleq 1 + \frac{4-\nu}{2(3-\nu)}\left(1 - 2^{2(3-\nu)/(4-\nu)}\right)$ and $\mathsf{b} \triangleq \frac{2-\nu}{4-\nu}$.

In fact, this result is a simple consequence of the technical lemma below.

Lemma C.2. *Consider function $t \mapsto \psi_\nu(t)$ defined in (**??**) with unique maximum t_ν^* as described in (**??**). It holds that*

$$\psi_\nu(t_\nu^*) = \begin{cases} \frac{1}{\delta}\left((1+\frac{\xi}{\delta})\ln\left(1+\frac{\delta}{\xi}\right)-1\right) & \text{if } \nu = 2, \\ \frac{1}{\delta}\left(1 - \frac{\mathsf{ab}\xi}{\delta} + \frac{\mathsf{ab}\xi}{\delta}\left(1-\frac{1}{\mathsf{b}}\frac{\delta}{\xi}\right)^{\mathsf{b}+1}\right) & \text{if } \nu \in (2,3)\cup(3,4), \\ \frac{1}{\delta}\left(1 - \frac{\xi}{\delta}\ln\left(1+\frac{\delta}{\xi}\right)\right) & \text{if } \nu = 3, \\ \frac{1}{\delta}\left[1 - \frac{\xi}{\delta} + \frac{\xi}{\delta}\exp\left(-\frac{\delta}{\xi}\right)\right] & \text{if } \nu = 4. \end{cases}, \tag{C.2}$$

where $\mathsf{a} \triangleq \frac{4-\nu}{2(3-\nu)}$ and $\mathsf{b} \triangleq \frac{2-\nu}{4-\nu} < 0$. Moreover, the following lower bound holds

$$\psi_\nu(t^*) \ge \begin{cases} \frac{2\ln 2-1}{\delta}\min\{1, \frac{\delta}{\xi}\} & \text{if } \nu = 2, \\ \frac{\tilde{\gamma}_\nu}{\delta}\min\left\{1, -\frac{\delta}{\xi\mathsf{b}}\right\} & \text{if } \nu \in (2,3)\cup(3,4), \\ \frac{1-\ln 2}{\delta}\min\{1, \frac{\delta}{\xi}\} & \text{if } \nu = 3, \\ \frac{\exp(-1)}{\delta}\min\{1, \frac{\delta}{\xi}\} & \text{if } \nu = 4. \end{cases}, \tag{C.3}$$

where

$$\tilde{\gamma}_\nu \triangleq 1 + \frac{4-\nu}{2(3-\nu)}\left(1 - 2^{2(3-\nu)/(4-\nu)}\right). \tag{C.4}$$

Proof. We organize the proof according to the value of $\nu \in [2,4]$.

The case $\nu = 2$: Since $\psi_2(t) = t - \frac{\xi}{\delta^2}[e^{t\delta} - t\delta - 1]$, once we plug in t_2^* from eq. (**??**) we arrive, after some computations, at

$$\psi_2(t_2^*) = \frac{1}{\delta}\left((1+\frac{\xi}{\delta})\ln(1+\frac{\delta}{\xi}) - 1\right)$$

We next establish the lower bound formulated in (**??**). Denote $\phi(t) \triangleq (1+t)\ln\left(1+\frac{1}{t}\right) - 1$. Then $\psi(t_2^*) = \phi(\frac{\xi}{\delta})/\delta$. At the same time,

$$\frac{d\phi(t)}{dt} = \ln\left(1+\frac{1}{t}\right) + (1+t)\cdot\frac{t}{1+t}\cdot\left(-\frac{1}{t^2}\right) = \ln\left(1+\frac{1}{t}\right) - \frac{1}{t} < 0.$$

Thus, $\phi(t)$ is decreasing and $\phi(t) \geq \phi(1) = 2\ln 2 - 1$ when $t \in (0,1]$. Let us now consider the function $\phi(t)/(1/t)$.

$$\frac{d}{dt}\left(\frac{\phi(t)}{1/t}\right) = \frac{d}{dt}(t\phi(t)) = (2t+1)\ln\left(1+\frac{1}{t}\right) + t(1+t)\cdot\frac{t}{1+t}\cdot\left(-\frac{1}{t^2}\right) - 1$$
$$= (2t+1)\ln\left(1+\frac{1}{t}\right) - 2 \geq 0.$$

Hence, $\phi(t)/(1/t) \geq \phi(1) = 2\ln 2 - 1$ when $t \in (1,+\infty)$. Combining these two cases, we see that

$$\psi_2(t_2^*) = \frac{1}{\delta}\phi(\xi/\delta) \geq (2\ln(2)-1)\min\{1/\delta, 1/\xi\}. \tag{C.5}$$

The case $\nu \in (2,3)$**:** A computation shows that

$$\psi_\nu(t_\nu^*) = \frac{1}{\delta}\left[1 - \frac{4-\nu}{2(3-\nu)}\left(1+\frac{\delta}{\xi}\frac{4-\nu}{\nu-2}\right)^{\frac{2-\nu}{4-\nu}}\right] + \frac{\xi}{\delta^2}\frac{(\nu-2)}{2(3-\nu)}\left[1-\left(1+\frac{\delta}{\xi}\frac{4-\nu}{\nu-2}\right)^{\frac{2-\nu}{4-\nu}}\right].$$

Set $\mathsf{a} \triangleq \frac{4-\nu}{2(3-\nu)} > 0$ and $\mathsf{b} \triangleq \frac{2-\nu}{4-\nu} < 0$. Then, setting $u = 1 - \frac{1}{\mathsf{b}}\frac{\delta}{\xi}$, we see that

$$\psi_\nu(t_\nu^*) = \frac{1}{\delta}\left(1 - \frac{\xi\mathsf{ab}}{\delta} - \mathsf{a}u^{\mathsf{b}} + \mathsf{ab}\frac{\xi}{\delta}u^{\mathsf{b}}\right)$$
$$= \frac{1}{\delta}\left[1 - \frac{\mathsf{ab}\xi}{\delta} + \frac{\mathsf{ab}\xi}{\delta}\left(1-\frac{1}{\mathsf{b}}\frac{\delta}{\xi}\right)^{\mathsf{b}+1}\right]$$

To verify the lower bound, we rewrite $\psi_\nu(t_\nu^*)$ as follows:

$$\psi_\nu(t_\nu^*) = \frac{1}{\delta}\left(1 - \mathsf{a}u^{\mathsf{b}} + \frac{\mathsf{a}}{u-1}(1-u^b)\right)$$
$$= \frac{1}{\delta}\left(1 + \frac{\mathsf{a}}{u-1} - \frac{\mathsf{a}u^{\mathsf{b}+1}}{u-1}\right)$$
$$= \frac{1}{\delta}\gamma(u),$$

where $\gamma(u) \triangleq 1 + \frac{\mathsf{a}}{u-1} - \frac{\mathsf{a}u^{\mathsf{b}+1}}{u-1}$. Our next goal is to show that, for $u \in [2,+\infty)$, $\gamma(u)$ is below bounded by some positive constant and, for $u \in (1,2]$, $\gamma(u)$ is below bounded by some positive constant multiplied by $u-1$.

1. $u \in [2,+\infty)$. We will show that $\gamma'(u) \geq 0$, whence $\gamma(u) \geq \gamma(2)$. Thus, we need to show that

$$0 \leq \gamma'(u) = -\frac{\mathsf{a}}{(u-1)^2}\underbrace{\left(1 - (\mathsf{b}+1)u^{\mathsf{b}} + \mathsf{b}u^{\mathsf{b}+1}\right)}_{=h(u)}.$$

Since $\mathsf{a} > 1$, to show that $\gamma'(u) \geq 0$ it is enough to show that $h(u) \leq 0$. Since $\mathsf{b} \in (-1,0)$ and $t \geq 2$,

$$h'(u) = \mathsf{b}(\mathsf{b}+1)u^{\mathsf{b}} - \mathsf{b}(\mathsf{b}+1)u^{\mathsf{b}-1} = \mathsf{b}(\mathsf{b}+1)u^{\mathsf{b}-1}(u-1) \leq 0.$$

Whence, $h(u) \leq h(2)$ for all $u \in [2,+\infty)$. It remains to show that $h(2) \leq 0$. Let us consider $h(2) = \varphi(\mathsf{b}) := 1-(\mathsf{b}+1)2^{\mathsf{b}} + \mathsf{b}2^{\mathsf{b}+1} = 1 + \mathsf{b}2^{\mathsf{b}} - 2^{\mathsf{b}}$ as a function of $\mathsf{b} \in (-1,0)$. Clearly, $\varphi(-1) = \varphi(0) = 0$,

and it is easy to check via the intermediate value theorem that $\varphi(b) < 0$ for all $b \in (-1, 0)$. We conclude that for $u \geq 2$ we get $\psi_\nu(t_2^*) \geq \frac{1}{\delta}\gamma(2)$.

2. $t \in (1, 2]$. We will show that $\frac{d}{du}\left(\gamma(u)/(u-1)\right) \leq 0$, whence $\gamma(u) \geq (u-1)\gamma(2)$. Thus, we need to show that

$$\begin{aligned} 0 &\geq \frac{d}{dt}\left(\frac{1}{u-1} + \frac{\mathsf{a}}{(u-1)^2} - \frac{\mathsf{a}u^{\mathsf{b}+1}}{(u-1)^2}\right) \\ &= \frac{1}{(u-1)^3}\left(-u + 1 - 2\mathsf{a} + \mathsf{a}(\mathsf{b}+1)u^{\mathsf{b}} - \mathsf{a}(\mathsf{b}-1)u^{\mathsf{b}+1}\right) \equiv \frac{1}{(u-1)^3}h(u). \end{aligned}$$

Therefore, our next step is to show that $h(u) \leq 0$. We have

$$\begin{aligned} h'(u) &= -1 + \mathsf{a}(\mathsf{b}+1)\mathsf{b}u^{\mathsf{b}-1} - \mathsf{a}(\mathsf{b}-1)(\mathsf{b}+1)u^{\mathsf{b}}, \\ h''(u) &= \mathsf{ab}(\mathsf{b}+1)(\mathsf{b}-1)u^{\mathsf{b}-2} - \mathsf{a}(\mathsf{b}-1)\mathsf{b}(\mathsf{b}+1)u^{\mathsf{b}-1} \\ &= \mathsf{ab}(\mathsf{b}+1)(\mathsf{b}-1)u^{\mathsf{b}-2}(1-u). \end{aligned}$$

By definition, $\mathsf{a}(\mathsf{b}+1) = 1$. Hence, since $u > 1$ and $\mathsf{b} \in (-1, 0)$, we observe that $h''(u) \leq 0$. Thus, $h'(u) \leq h'(1) = 0$, and consequently, $h(u) \leq h(1) = 0$, for all $u \in (1, 2]$. This proves the claim $\gamma(u)/(u-1) \geq \gamma(2)$ for $u \in (1, 2]$.

Combining both cases, we obtain that $\gamma(u) \geq \min\{\gamma(2), (u-1)\gamma(2)\}$, where $\gamma(2) = 1 - \mathsf{a} + \mathsf{a}2^{1/\mathsf{a}}$, using the fact that $\mathsf{b} + 1 = 1/\mathsf{a}$. Unraveling this expression by using the definition of the constant a, we see that $\gamma(2)$ depends only on the self-concordance parameter $\nu \in (2, 3)$. In light of this, let us introduce the constant

$$\tilde{\gamma}_\nu \triangleq 1 + \frac{4-\nu}{2(3-\nu)}\left(1 - 2^{2(3-\nu)/(4-\nu)}\right). \tag{C.6}$$

Observe that $\tilde{\gamma}_2 = 0$ and, by a simple application of l'Hôpital's rule, $\lim_{\nu\uparrow 3}\tilde{\gamma}_\nu = 1 - \log(2) \in (0, 1)$. Hence $\gamma(2) \equiv \tilde{\gamma}_\nu \in (0, 1)$ for all $\nu \in (2, 3)$. We conclude,

$$\psi_\nu(t_\nu^*) \geq \frac{\tilde{\gamma}_\nu}{\delta}\min\left\{1, \frac{-1}{\mathsf{b}}\frac{\delta}{\xi}\right\} \tag{C.7}$$

The case $\nu = 3$**:** A direct substitution for $\psi_3(t)$ gives us

$$\psi_3(t_3^*) = \frac{1}{\delta} + \frac{\xi}{\delta^2}\ln\left(\frac{\xi}{\delta+\xi}\right). \tag{C.8}$$

Denote $u = \xi/\delta$. Then $t_3^* = \frac{1}{\delta+\xi}$, so that

$$\psi_3(t_3^*) = \frac{1}{\delta}\left[1 + u\ln\left(\frac{u}{u+1}\right)\right].$$

Consider the function $\phi : (0, \infty) \to (0, \infty)$, given by $\phi(t) := 1 + t\ln\left(\frac{t}{1+t}\right)$. Then, $\psi_3(t_3^*) = \frac{1}{\delta}\phi(\xi/\delta)$. We use this identity to obtain the lower bound announced in eq. **??**.

When $t \in (0, 1)$, since

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

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 $t \geq 1$,

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

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 $t \geq 1$. Summarizing these two cases we see

$$\psi_3(t_3^*) \geq \frac{1}{\delta}\min\{1, \delta/\xi\}(1 - \ln(2)) = (1 - \ln(2))\min\{1/\delta, 1/\xi\}. \tag{C.9}$$

$\nu \in (3,4)$**:** Similarly to the case $\nu \in (2,3)$, denote $u = 1 - \frac{1}{\mathsf{b}}\frac{\delta}{\xi}$, where $\mathsf{a} = \frac{4-\nu}{2(3-\nu)} \in (-\infty, 0)$, $\mathsf{b} = \frac{2-\nu}{4-\nu} \in (-\infty, -1)$. Then the expression for the $\psi_\nu(t_\nu^*)$ is the same as in the case $\nu \in (2,3)$:

$$\psi_\nu(t_\nu^*) = \frac{1}{\delta}\left(1 + \frac{\mathsf{a}}{u-1} - \frac{\mathsf{a}u^{\mathsf{b}+1}}{u-1}\right) = \frac{1}{\delta}\gamma(u).$$

Our next goal is to show that, for $t \in [2, +\infty)$, $\gamma(t) := \left(1 + \frac{\mathsf{a}}{t-1} - \frac{\mathsf{a}t^{\mathsf{b}+1}}{t-1}\right)$ is below bounded by some positive constant and, for $t \in (1,2]$, $\gamma(t)$ is below bounded by some positive constant multiplied by $t-1$.

1. $t \in [2, +\infty)$. We will show that $\gamma'(t) \geq 0$, whence $\gamma(t) \geq \gamma(2)$. Thus, we need to show that for $t \geq 2$,

$$0 \leq \gamma'(t) = -\frac{\mathsf{a}}{(t-1)^2}\underbrace{\left(1 - (\mathsf{b}+1)t^{\mathsf{b}} + \mathsf{b}t^{\mathsf{b}+1}\right)}_{=:h(t)}.$$

Since, for $\nu \in (3,4)$, $\mathsf{a} \leq 0$, to show that $\gamma'(t) \geq 0$ it is enough to show that $h(t) \geq 0$. Since $\mathsf{b} < -1$,

$$h'(t) = \mathsf{b}(\mathsf{b}+1)t^{\mathsf{b}} - \mathsf{b}(\mathsf{b}+1)t^{\mathsf{b}-1} = \mathsf{b}(\mathsf{b}+1)t^{\mathsf{b}-1}(t-1) \geq 0,$$

whence, $h(t) \geq h(2)$, $t \in [2, +\infty)$. It remains to show that $h(2) \geq 0$. Let us consider $h(2) = 1 - (\mathsf{b}+1)2^{\mathsf{b}} + \mathsf{b}2^{\mathsf{b}+1} = 1 + \mathsf{b}2^{\mathsf{b}} - 2^{\mathsf{b}}$ as a function of b. For all possible values $\mathsf{b} \in (-\infty, -1)$ one can check numerically that $\psi(2) \in (0,1)$. Hence, $h(t) \geq 0$ for all $t \geq 2$.

2. $t \in (1,2]$. We will show that $\frac{d}{dt}\left(\gamma(t)/(t-1)\right) \leq 0$, whence $\gamma(t) \geq (t-1)\gamma(2)$. Thus, we need to show that

$$\begin{aligned}
0 &\geq \frac{d}{dt}\left(\frac{1}{t-1} + \frac{\mathsf{a}}{(t-1)^2} - \frac{\mathsf{a}t^{\mathsf{b}+1}}{(t-1)^2}\right) \\
&= -\frac{1}{(t-1)^2} - \frac{2\mathsf{a}}{(t-1)^3} - \frac{\mathsf{a}(\mathsf{b}+1)t^{\mathsf{b}}}{(t-1)^2} + \frac{2\mathsf{a}t^{\mathsf{b}+1}}{(t-1)^3} \\
&= \frac{1}{(t-1)^3}\left(-t + 1 - 2\mathsf{a} - \mathsf{a}(\mathsf{b}+1)t^{\mathsf{b}+1} + \mathsf{a}(\mathsf{b}+1)t^{\mathsf{b}} + 2\mathsf{a}t^{\mathsf{b}+1}\right) \\
&= \frac{1}{(t-1)^3}\underbrace{\left(-t + 1 - 2\mathsf{a} + \mathsf{a}(\mathsf{b}+1)t^{\mathsf{b}} - \mathsf{a}(\mathsf{b}-1)t^{\mathsf{b}+1}\right)}_{=:h(t)}.
\end{aligned}$$

Our next step is to show that $h(t) \leq 0$. We have

$$h'(t) = -1 + \mathsf{a}(\mathsf{b}+1)\mathsf{b}t^{\mathsf{a}-1} - \mathsf{a}(\mathsf{b}-1)(\mathsf{b}+1)t^{\mathsf{b}}$$

$$h''(t) = \mathsf{ab}(\mathsf{b}+1)(\mathsf{b}-1)t^{\mathsf{b}-2} - \mathsf{a}(\mathsf{b}-1)\mathsf{b}(\mathsf{b}+1)t^{\mathsf{b}-1}$$
$$= \mathsf{ab}(\mathsf{b}+1)(\mathsf{b}-1)t^{\mathsf{b}-2}(1-t).$$

Using the definition of a, b, and the fact that $\nu \in (3,4)$, we obtain that $\mathsf{a}(\mathsf{b}+1) = 1$. Hence, since $t > 1$, we obtain that $h''(t) \leq 0$. Thus, $h'(t) \leq h'(1) = 0$, $h(t) \leq h(1) = 0$, and $\gamma(t)/(t-1) \geq \gamma(2)$.

Combining both cases, we obtain that $\gamma(t) \geq \min\{\gamma(2), (t-1)\gamma(2)\}$. Note that $\gamma(2) = \tilde{\gamma}_\nu$ as defined in eq. (**??**). Consequently,

$$\psi_\nu(t_\nu^*) \geq \frac{\tilde{\gamma}_\nu}{\delta} \min\{1, -\frac{1}{\mathsf{b}}\frac{\delta}{\xi}\} = \tilde{\gamma}_\nu \min\{\frac{1}{\delta}, -\frac{1}{\mathsf{b}}\frac{1}{\xi}\}.$$

The case $\nu = 4$: A simple computation shows that

$$\psi_\nu(t_\nu^*) = \frac{1}{\delta}\left[1 - \frac{\xi}{\delta} + \frac{\xi}{\delta}\exp\left(-\frac{\delta}{\xi}\right)\right].$$

To analyze this expression, denote by $u := \frac{\xi}{\delta}$. Then

$$\psi_\nu(t_\nu^*) = \frac{1}{\delta}\left(1 - \frac{1}{u} + \frac{1}{u}\exp(-1/u)\right) \geq 0.$$

Let us define a function $\gamma(t)$ such that $\psi_\nu(t_\nu^*) = \frac{1}{\delta}\gamma(1/u)$. Our next goal is to show that, for $t \in (0,1]$, $\gamma(t)$ is below bounded by some positive constant and, for $t \geq 1$, $\gamma(t)$ is below bounded by some positive constant divided by t.

1. $t \in (0,1]$. We will show that $\gamma'(t) \leq 0$, whence $\gamma(t) \geq \gamma(1)$. Indeed, for $t \in (0,1]$,

$$\gamma'(t) = -1 + \exp(-1/t)(1+1/t) < -1 + 2\exp(1/t) \leq -1 + 2\exp(-1) < 0.$$

Thus, we have

$$\gamma(t) \geq \gamma(1) = \exp(-1).$$

2. $t \in [1, +\infty)$. We will show that $\frac{d}{dt}\left(\frac{\gamma(t)}{1/t}\right) \geq 0$, whence $\gamma(t) \geq \frac{\gamma(1)}{t}$.

$$\frac{d}{dt}\left(t\left(1 - t + t\exp\left(-\frac{1}{t}\right)\right)\right) = \exp\left(-\frac{1}{t}\right)(2t+1) + 1 - 2t. \tag{C.10}$$

Using the Taylor expansion for $\ln(1+x)$ and $\ln(1-x)$ for $x \in (0, 0.5]$, we have

$$\ln(1+x) - \ln(1-x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \sum_{k=4}^{\infty}\frac{(-1)^k x^k}{k} - \left(-x - \frac{x^2}{2} + \frac{x^3}{3} - \sum_{k=4}^{\infty}\frac{x^k}{k}\right)$$
$$= 2x + \frac{2x^3}{3} + \sum_{k=2}^{\infty}\frac{2x^{2k+1}}{2k+1} \geq 2x.$$

Setting $x = \frac{1}{2t}$ for $t \geq 1$, we obtain

$$\log(1 + 1/(2t)) - \log(1 - 1/(2t)) \geq 1/t$$
$$\Leftrightarrow \log(2t(1 + 1/(2t))) - \log(2t(1 - 1/(2t))) \geq 1/t$$
$$\Leftrightarrow \log(2t+1) - \log(2t-1) \geq 1/t$$

$$\Leftrightarrow -1/t + \log(2t+1) \geq \log(2t-1)$$
$$\Leftrightarrow \exp\left(-\frac{1}{t}\right)(2t+1) + 1 - 2t \geq 0.$$

which, combined with (**??**) proves that $\frac{d}{dt}\left(\frac{\gamma(t)}{1/t}\right) \geq 0$ for $t \geq 1$. This, we have that, for $t \geq 1$, $\gamma(t) \geq \frac{\gamma(1)}{t} = \frac{\exp(-1)}{t}$. Combining the two cases, we obtain the lower bound

$$\gamma(t) \geq \exp(-1)\min\{1, 1/t\} \qquad \forall t > 0.$$

Since $1/u = \frac{\delta}{\xi}$, this lower bound implies that

$$\psi_4(t_4^*) = \frac{1}{\delta}\gamma(1/u) \geq \frac{\exp(-1)}{\delta}\min\left\{1, \frac{\delta}{\xi}\right\}. \tag{C.11}$$

■

Proof of Lemma **??**. By identifying the parameters appropriately, we can give the proof of Lemma **??** as a straightforward exercise derived from Lemma **??**. We provide the explicit derivation for each GSC parameter ν below.

$\nu = 2$: Substitute in (**??**) the parameter values $\xi = \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$ and $\delta = \delta_2(x) = M_f\beta(x)$, the lower bound turns into

$$\psi_2(\mathtt{t}_2(x)) \geq \frac{2\ln(2)-1}{\mathsf{e}(x)}\min\left\{\frac{\mathsf{Gap}(x)}{\mathsf{e}(x)}, \frac{\mathsf{e}(x)}{M_f\beta(x)}\right\}. \tag{C.12}$$

$\nu \in (2,3)$: Substitute in (**??**) the parameter values $\delta \equiv \delta_\nu(x) = M_f\frac{\nu-2}{2}\beta(x)^{3-\nu}\mathsf{e}(x)^{\nu-2}$, $\xi \equiv \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, so that

$$\psi_\nu(\mathtt{t}_\nu(x)) \geq \tilde{\gamma}_\nu\min\left\{\frac{1}{\frac{\nu-2}{2}\beta(x)^{3-\nu}\mathsf{e}(x)^{\nu-2}}, \frac{-1}{\mathsf{b}}\frac{\mathsf{Gap}(x)}{\mathsf{e}(x)^2}\right\}. \tag{C.13}$$

$\nu = 3$: Substitute in (**??**) the parameter values $\delta \equiv \delta_3(x) = \frac{M_f}{2}\mathsf{e}(x)$, $\xi \equiv \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, to get

$$\psi_3(\mathtt{t}_3(x)) \geq \frac{2(1-\ln(2))}{M_f\mathsf{e}(x)}\min\left\{1, \frac{M_f\,\mathsf{Gap}(x)}{\mathsf{e}(x)}\right\}. \tag{C.14}$$

$\nu \in (3,4)$: Same as $\nu \in (2,3)$.

$\nu = 4$: Substitute in (**??**) the parameter values $\delta \equiv \delta_4(x) = M_f\beta(x)^{-1}\mathsf{e}(x)^2$, $\xi \equiv \frac{\mathsf{e}(x)^2}{\mathsf{Gap}(x)}$, so that

$$\psi_4(\mathtt{t}_4(x)) \geq \frac{\exp(-1)\beta(x)}{M_f\mathsf{e}(x)^2}\min\left\{1, \frac{M_f\,\mathsf{Gap}(x)}{\beta(x)}\right\}. \tag{C.15}$$

■

Proof of Lemma **??**. Use the estimates $\beta(x) \leq \mathrm{diam}(\mathcal{X})$ and $\mathsf{e}(x) \leq \sqrt{L_{\nabla f}}\beta(x) \leq \sqrt{L_{\nabla f}}\,\mathrm{diam}(\mathcal{X})$ in the expressions provided in Lemma **??**. ■

D Proofs for Section ??

Lemma D.1. *For all $t \in [0,1]$ 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,

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

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 $t \in [0,1]$. ∎

Lemma D.2. *We have $\mathcal{L}_k \leq \max\{\mathcal{L}_{-1}, \gamma_u L_{\nabla f}\}$.*

Proof. By construction of the backtracking procedure we know that if the sufficient decrease condition is evaluated successfully at the first run, then $\mathcal{L}_{k-1} \geq \mathcal{L}_k \geq \gamma_d\mathcal{L}_{k-1}$. If not, then it is clear that $\mathcal{L}_k \leq \gamma_d L_{\nabla f}$. Hence, for all $k \geq 0$, $\mathcal{L}_k \leq \max\{\gamma_d L_{\nabla f}, \mathcal{L}_{k-1}\}$. By backwards induction, it follows then $\mathcal{L}_k \leq \max\{\mathcal{L}_{-1}, \gamma_u L_{\nabla f}\}$. ∎

References

[] Francis Bach. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010. doi: 10.1214/09-EJS521. URL https://projecteuclid.org:443/euclid.ejs/1271941980.

[] Michel Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009.

[] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2017/09/01 2009. doi: 10.1137/080716542. URL https://doi.org/10.1137/080716542.

[] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1):1–27, 2017. doi: 10.1007/s10107-016-1069-4. URL https://doi.org/10.1007/s10107-016-1069-4.

[] Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

[] Amir Beck and Marc Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1):1 – 6, 2014.

[] Lev Bogolubsky, Pavel Dvurechensky, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, Aleksey Tikhonov, and Maksim Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4914–4922. Curran Associates, Inc., 2016. arXiv:1603.00717.

[] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), May 2011. ISSN 2157-6904. doi: 10.1145/1961189.1961199. URL https://doi.org/10.1145/1961189.1961199.

[] Imre Csiszar. Why Least Squares and Maximum Entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, 19(4):2032–2066, 1991. doi: 10.1214/aos/1176348385. URL https://projecteuclid.org:443/euclid.aos/1176348385.

[] Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002. doi: 10.1007/s101070100263. URL https://doi.org/10.1007/s101070100263.

[] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, 2018. arXiv:1802.04367.

[] Pavel Dvurechensky, Shimrit Shtern, Mathias Staudigl, Petr Ostroukhov, and Kamil Safin. Self-concordant analysis of Frank-Wolfe algorithms. *Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.*, 2020.

[] Marina Epelman and Robert M. Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Mathematical Programming*, 88(3):451–485, 2000. doi: 10.1007/s101070000136. URL https://doi.org/10.1007/s101070000136.

[] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 2019/09/05 1956. doi: 10.1002/nav.3800030109. URL https://doi.org/10.1002/nav.3800030109.

[] Robert M. Freund, Paul. Grigas, and Rahul. Mazumder. An extended Frank–Wolfe method with "in-face"directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2020/02/05 2017. doi: 10.1137/15M104726X. URL https://doi.org/10.1137/15M104726X.

[] Dan Garber and Elad Hazan. 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, 2020/01/01 2016. doi: 10.1137/140985366. URL https://doi.org/10.1137/140985366.

[] Jacques GuéLat and Patrice Marcotte. Some comments on wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986. doi: 10.1007/BF01589445. URL https://doi.org/10.1007/BF01589445.

[] David H. Gutman and Javier F. Peña. The condition number of a function relative to a set. *Mathematical Programming*, 2020. ISSN 1436-4646. URL https://doi.org/10.1007/s10107-020-01510-4.

[] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1):75–112, 2015. doi: 10.1007/s10107-014-0778-9. URL https://doi.org/10.1007/s10107-014-0778-9.

[] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435, 2013.

[] Rahul G Krishnan, Simon Lacoste-Julien, and David Sontag. Barrier Frank-Wolfe for marginal inference. In *Advances in Neural Information Processing Systems*, pages 532–540, 2015.

[] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. pages 496–504, 2015.

[] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

[] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016. doi: 10.1137/140992382. URL https://doi.org/10.1137/140992382.

[] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966. doi: https://doi.org/10.1016/0041-5553(66)90114-5. URL http://www.sciencedirect.com/science/article/pii/0041555366901145.

[] Yen-Huan Li and Volkan Cevher. Convergence of the exponentiated gradient method with armijo line search. *Journal of Optimization Theory and Applications*, 181(2):588–607, May 2019. ISSN 1573-2878. doi: 10.1007/s10957-018-1428-9. URL https://doi.org/10.1007/s10957-018-1428-9.

[] Deyi Liu, Volkan Cevher, and Quoc Tran-Dinh. A Newton Frank-Wolfe method for constrained self-concordant minimization. *preprint arXiv:2002.07003*, 2020.

[] J. S. Marron, Michael J. Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007. ISSN 01621459. URL http://www.jstor.org/stable/27639976.

[] Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/marteau-ferey19a.html.

[] Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[] Yu. Nesterov and A. Nemirovski. *Interior Point Polynomial methods in Convex programming*. SIAM Publications, 1994.

[] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. doi: 10.1007/s10107-012-0629-5. URL https://doi.org/10.1007/s10107-012-0629-5.

[] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer International Publishing, 2018.

[] Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1):311–330, 2018. doi: 10.1007/s10107-017-1188-6. URL https://doi.org/10.1007/s10107-017-1188-6.

[] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2000.

[] Gergely Odor, Yen-Huan Li, Alp Yurtsever, Ya-Ping Hsieh, Quoc Tran-Dinh, Marwa El Halabi, and Volkan Cevher. Frank-Wolfe works for non-Lipschitz continuous gradient objectives: Scalable poisson phase retrieval. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6230–6234, 2016.

[] Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *arXiv preprint arXiv:1810.06838*, 2018.

[] Art B. Owen. Self-concordance for empirical likelihood. *Canadian Journal of Statistics*, 41(3): 387–397, 2020/02/05 2013. doi: 10.1002/cjs.11183. URL https://doi.org/10.1002/cjs.11183.

[] Fabian Pedregosa, Geoffrey Negiar, Armin Askari, and Martin Jaggi. Linearly convergent Frank-Wolfe with backtracking line-search. In *International Conference on Artificial Intelligence and Statistics*, pages 1–10. PMLR, 2020.

[] Javier Peña and Daniel Rodríguez. Polytope conditioning and linear convergence of the Frank–Wolfe algorithm. *Mathematics of Operations Research*, 44(1):1–18, 2020/01/07 2018. doi: 10.1287/moor.2017.0910. URL https://doi.org/10.1287/moor.2017.0910.

[] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 2018. doi: 10.1007/s10107-018-1282-4. URL https://doi.org/10.1007/s10107-018-1282-4.

[] Quoc Tran-Dinh, Anastasios Kyrillidis, and Volkan Cevher. Composite self-concordant minimization. *The Journal of Machine Learning Research*, 16(1):371–416, 2015.

[] Levent Tunçel and Arkadi Nemirovski. Self-concordant barriers for convex approximations of structured convex sets. *Foundations of Computational Mathematics*, 10(5):485–525, 2010. ISSN 1615-3383. URL https://doi.org/10.1007/s10208-010-9069-x.

[] Yuchen Zhang and Xiao Lin. DiSCO: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 362–370. PMLR, 06 2015. URL http://proceedings.mlr.press/v37/zhangb15.html.