Fast Quasar-Convex Optimization with Constraints

David Martínez-Rubio
IMDEA Software Institute
Madrid, Spain
david.martinezrubio@imdea.org

Abstract

Quasar-convex functions form a broad nonconvex class with applications to linear dynamical systems, generalized linear models, and Riemannian optimization, among others. Current nearly optimal algorithms work only in affine spaces due to the loss of one degree of freedom when working with general convex constraints. Obtaining an accelerated algorithm that makes nearly optimal $\widetilde{O}(1/(\gamma\sqrt{\varepsilon}))$ first-order queries to a γ-quasar convex smooth function *with constraints* was independently asked as an open problem in Martínez-Rubio (2022); Lezane, Langer, and Koolen (2024). In this work, we solve this question by designing an inexact accelerated proximal point algorithm that we implement using a first-order method achieving the aforementioned rate and, as a consequence, we improve the complexity of the accelerated geodesically Riemannian optimization solution in Martínez-Rubio (2022). We also analyze projected gradient descent and Frank-Wolfe algorithms in this constrained quasar-convex setting. To the best of our knowledge, our work provides the first analyses of first-order methods for quasar-convex smooth functions with general convex constraints.

Most of the non-local notations in this work have a link to their definitions, using this code, such as γ, which links to where this notation is defined as the quasar-convexity parameter of the functions we optimize in this work.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Nonconvex optimization has become central to modern machine learning, yet our theoretical understanding of why simple first-order methods succeed in these settings remains limited. Algorithms such as stochastic gradient descent routinely achieve both efficient optimization and strong statistical performance, even though the underlying problems are nonconvex and, in principle, could exhibit a highly irregular landscape with spurious local minima or saddle points. This disconnect motivates a line of research seeking structural properties of nonconvex problems that explain the empirical successes of local-search algorithms.

Classical theory is built around convex optimization, where every local minimum is global, and where powerful tools of duality are available. However, the binary distinction between convex and nonconvex problems is too coarse: many important nonconvex problems exhibit benign structure that allows for efficient optimization. A rich body of work has introduced relaxations of convexity that retain the global optimality of local minima, including star-convexity, essential strong convexity, restricted secant inequalities, one-point strong convexity, variational coherence, quasiconvexity, pseudoconvexity, invexity, the Polyak-Łojasiewicz (PL) condition, tilted convexity, and the strict-saddle property without spurious local minima (Karimi, Nutini, and Schmidt, 2016; Ge, Huang, et al., 2015; Hinder, Sidford, and Sohoni, 2019; Martínez-Rubio, 2022).

Besides, it has been shown that all local minimizers are global in a number of machine learning tasks such as problems in phase retrieval (Sun, Qu, and Wright, 2018), tensor decomposition (Ge, Huang, et al., 2015), dictionary learning (Sun, Qu, and Wright, 2017), matrix sensing (Bhojanapalli, Neyshabur, and Srebro, 2016; Park et al., 2017), and matrix completion (Ge, Lee, and Ma, 2016). Moreover, under overparameterization assumptions, gradient descent provably finds global minimizers in neural networks (Allen-Zhu, Li,

and Song, 2019; Du, Lee, et al., 2019; Nguyen and Mondelli, 2020; Zou et al., 2018; Du, Zhai, et al., 2019). These results highlight a growing recognition that the landscape of many machine learning objectives is more favorable than worst-case nonconvexity suggests and motivates the study of optimization under benign nonconvexity.

Among the relaxations of convexity, *quasar convexity* has recently emerged as a particularly compelling property. It is a generalization of star-convexity, where the slope of the classical affine lower bound given by a gradient and its function value is multiplied by a number greater than 1, and one only requires for this to bound the function at a specific solution, cf. Section 2. This class of functions allows for acceleration under smooth objectives, in the *unconstrained case* and the essentially equivalent case where there is an affine constraint. When the problem has general constraints, current solutions lose one degree of freedom, and current accelerated solutions and analyses do not work.

Important examples of quasar-convex functions include linear dynamical systems (Hardt, Ma, and Recht, 2018), several generalized linear models (GLMs) (Foster, Sekhari, and Sridharan, 2018; Wang and Wibisono, 2023), and geodesically convex problems in constant-curvature Riemannian spaces after appropriate reductions (Martínez-Rubio, 2022). Further, the landscape of neural networks has been observed to empirically satisfy quasar-convexity along the trajectory of gradient descent methods with respect to the reached solution (Kleinberg, Li, and Yuan, 2018; Zhou et al., 2019; Tran, Zhang, and Cutkosky, 2024).

Despite the progress in the study of optimization with quasar convexity, constrained optimization in this regime remains poorly understood. Whether acceleration is possible by first-order methods for smooth quasar-convex problems with general convex constraints has been posed as an open question in independent works (Martínez-Rubio, 2022; Lezane, Langer, and Koolen, 2024), where a fast method in the presence of a compact convex constrained would improve the accelerated Riemannian optimization method in (Martínez-Rubio, 2022). In this paper, we resolve this question in the affirmative by providing an accelerated *constrained* first-order method that matches the lower bound of (Hinder, Sidford, and Sohoni, 2019), up to logarithmic factors. Our results extend the scope of quasar-convex optimization to restricted domains, thereby broadening its applicability to machine learning problems with natural structural restrictions. To the best of our knowledge, we provide the first analyses of first-order methods for smooth quasar-convex problems with general constraints.

Main Contributions. Our contributions can be summarized as follows:

1. We design an accelerated *constrained* first-order method for L-smooth γ-quasar-convex functions with respect to a minimizer x^* in a compact convex feasible set of diameter D, finding an ε-minimizer in the nearly optimal $\widetilde{O}\left(\gamma^{-1}\sqrt{LD^2/\varepsilon}\right)$ queries to a first-order oracle, where x_0 is an initial point, solving the open question in (Martínez-Rubio, 2022; Lezane, Langer, and Koolen, 2024). The solution involved designing an implementable inexact accelerated proximal method along with the design of a line search under adversarial noise due to the inexactness in the proximal solution.

2. Our algorithm implies improved complexity over the Riemannian optimization solution in (Martínez-Rubio, 2022), compared to prior work on tilted convexity, we achieve faster acceleration and under weaker assumptions.

3. We show that projected gradient descent as well as for the Frank-Wolfe algorithm converge at the unaccelerated rate $O(LD^2/(\gamma^2\varepsilon))$ for L-smooth γ-quasar convex functions with constraints.

2 PRELIMINARIES AND SETTING

Notation. We denote by $\arg\min^{\delta}_{x\in\mathcal{X}} f(x)$ the set of δ-minimizers of f over a feasible set $\mathcal{X}$. A function is said to be differentiable on a closed (possibly non-open) set $\mathcal{X}$ if is differentiable in an open neighborhood of $\mathcal{X}$. The set indicator function is $\mathbf{1}_X(x) = 0$ if $x \in \mathcal{X}$ and $\mathbf{1}_X(x) = +\infty$ otherwise. We use $\widetilde{O}(\cdot)$ as the big-O notation omitting logarithmic factors. For a set of points, we denote its convex hull by $\mathrm{conv}\{S\}$. We denote the Euclidean projection operator by $\mathrm{Proj}_{\mathcal{X}}(x) \stackrel{\text{def}}{=} \arg\min_{y\in\mathcal{X}} \|y - x\|_2$.

A first-order oracle for f is an operator that, given a query point $x \in \mathbb{R}^d$, returns the pair $(f(x), \nabla f(x))$.

γ-quasar convexity. For $\gamma \in (0, 1]$, a differentiable function is said to be γ-quasar convex in $\mathcal{X}$ with center $x^* \in \mathcal{X}$ if $x^* \in \arg\min_{x\in\mathcal{X}} f(x)$ and for every $x \in \mathcal{X}$:

$$f(x^*) \;\geq\; f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x\rangle.$$

If $\gamma = 1$, the function is said to be star convex. Note that by the definition above, any point x satisfying $\nabla f(x) = 0$ is also a minimizer.

L-smoothness. A differentiable function f has L-Lipschitz gradients in $\mathcal{X}$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2, \qquad \forall x, y \in \mathcal{X}.$$

Equivalently, f satisfies

$$|f(x) - f(y) - \langle \nabla f(y), x-y\rangle| \;\leq\; \frac{L}{2}\|x-y\|_2^2,$$

for all $x, y \in \mathcal{X}$. This property is also often called L-smoothness, more commonly in convex scenarios, where the absolute value above is redundant. Lastly, we say that a differentiable function f is μ-strongly convex in $\mathcal{X}$, for $\mu > 0$, if for all $x, y \in \mathcal{X}$, we have

$$f(x) - f(y) - \langle \nabla f(y), x-y\rangle \geq \frac{\mu}{2}\|x-y\|_2^2.$$

Problem setting. We study the problem

$$\min_{x\in\mathcal{X}} f(x), \tag{1}$$

for methods with access to a first-order oracle of f, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact convex set of diameter D, and $f : \mathbb{R}^d \to \mathbb{R}$ is both L-smooth and γ-quasar convex in $\mathcal{X}$ with center denoted by $x^* \in \mathcal{X}$. Our aim is to obtain an ε-minimizer $\hat{y}$ of the problem above, that is, $f(\hat{y}) - \min_{x\in\mathcal{X}} f(x) \leq \varepsilon$.

A relaxation of convexity stronger than γ-quasar convexity is (γ, γ_p)-tilted convexity. For $\gamma, \gamma_p \in (0, 1]$ and all x, y in a closed convex set $\mathcal{X}$, a tilted-convex function satisfies

$$f(x) + \frac{1}{\gamma}\langle \nabla f(x), y-x\rangle \leq f(y) \text{ if } \langle \nabla f(x), y-x\rangle \leq 0,$$
$$f(x) + \gamma_p\langle \nabla f(x), y-x\rangle \leq f(y) \text{ if } \langle \nabla f(x), y-x\rangle \geq 0.$$

This is clearly stronger than γ-quasar convexity since the first line above includes the quasar-convex property if $x = x^*$. We note that (Martínez-Rubio, 2022, Theorem 5) provided a first-order method for optimizing a (γ, γ_p)-tilted convex function with L-Lipschitz gradients in a compact convex set $\mathcal{X}$ of diameter D in time $\widetilde{O}(\sqrt{LD^2/(\gamma^2\gamma_p\varepsilon)})$, where the complexity contains a logarithmic dependence on the Lipschitz constant of f in $\mathcal{X}$. Our main result, cf. Theorem 5, shows that under the weaker assumption of γ-quasar convexity, we obtain faster convergence $\widetilde{O}(\sqrt{LD^2/(\gamma^2\varepsilon)})$, where the dependence of the complexity on γ_p and the logarithm of the Lipschitz constant of f disappear. Note that a smooth constrained optimization problem could have a moderate smoothness constant and at the same time have an arbitrary large Lipschitz constant.

Our solution uses the structure of an approximate proximal point method. The following related notions are important. For a function f and a regularization parameter $\lambda > 0$, the Moreau envelope is defined as

$$M_{\lambda f}(x) \stackrel{\text{def}}{=} \inf_y \left\{ f(y) + \frac{1}{2\lambda}\|y-x\|_2^2 \right\}.$$

We call the subproblem in the definition of $M_{\lambda f}(x)$ the proximal subproblem, and denote $\text{prox}_{\lambda f}(x) \stackrel{\text{def}}{=} \arg\min \left\{ f(y) + \frac{1}{2\lambda}\|y-x\|_2^2 \right\}$. If f and λ are clear from context, we will simply use $M(x)$, $\text{prox}(x)$. If the proximal subproblem has a unique solution, then the Moreau envelope is differentiable and it is

$$\nabla M(x) = \frac{1}{\lambda}(x - \text{prox}(x)),$$

see (Bertsekas, Nedic, and Ozdaglar, 2003).

3 RELATED WORK

Hardt, Ma, and Recht (2018) introduced the quasar convex class, where it was named weak quasi-convexity. They provided an analysis of stochastic gradient descent for a weakly smooth quasar convex objective. Guminov, Gasnikov, and Kuruzov (2023) proposed an accelerated algorithm for smooth quasar convex functions with a search over a 2-dimensional affine space. Their solution was inspired by the first nearly accelerated first-order method for convex smooth functions (Nemirovski, 1982b; Nemirovski, 1982a) and also by Narkiss and Zibulevsky (2005), that builds on the former. Nesterov et al. (2021) followed up on the previous work, providing a universal algorithm that simplified the plane search to a line search and allowed for affine constraints. Because the feasible set is still an affine space, no degree of freedom is lost with respect to the unconstrained case. Later, Hinder, Sidford, and Sohoni (2020) also proposed an accelerated algorithm with line search, quantifying the time that it would take for a binary search to run, which required bounding the algorithm's iterates, and they proved nearly matching lower bounds, and coined the term quasar-convexity. They also studied strong quasar-convex problems.

Wang and Wibisono (2023) extended the continuized approach for acceleration of Even et al. (2021) in order to obtain a randomized method that with high probability obtains an ε minimizer in $O(\sqrt{\frac{LR^2}{\varepsilon}})$ iterations, that is, without log factors. Lezane, Langer, and Koolen (2024) generalized accelerated solutions to non-Euclidean and weak smoothness function classes. Jin (2020) studied gradient norm minimization under quasar-convexity. Gower, Sebbouh, and Loizou (2021); Vaswani, Dubois-Taine, and Babanezhad (2022); Fu, Xu, and Wilson (2023) studied the optimization of stochastic quasar convex functions with adaptivity guarantees. Saad, Lee, and

Orabona (2025) provided lower bounds for stochastic quasar-convex functions. When the problem is Lipschitz *non-smooth*, standard regret-minimization algorithms generalize to work for the best iterate with constrained settings for quasar convex problems, see Joulani, György, and Szepesvári (2020).

Regarding the applicability of quasar-convexity, Hardt, Ma, and Recht (2018) show that several linear dynamical systems optimization tasks are quasar-convex. Foster, Sekhari, and Sridharan (2018) show that GLMs, that use losses of the form $w \mapsto (\sigma(\langle w, x\rangle - y))$ for a datapoint (x, y), are quasar convex if the link function σ satisfies boundedness conditions on its first two derivatives, although they stated a weaker property. Wang and Wibisono (2023) provide further examples of families of GLMs that are quasar-convex. Foster, Sekhari, and Sridharan (2018) also showed that a robust linear regression problem is quasar convex. Martínez-Rubio (2022) reduced geodesically-convex optimization in constant curvature Riemannian Manifolds to constrained tilted-convex problems in the Euclidean space, and therefore to quasar-convex problems.

Our main algorithm is built as an inexact accelerated proximal point method. Obtaining accelerated algorithms by designing an inexact accelerated proximal point method and implementing the approximate proximal subproblem with first-order methods is a technique that has yielded general and fruitful frameworks, cf. (Monteiro and Svaiter, 2013; Lin, Mairal, and Harchaoui, 2017; Frostig et al., 2015; Ivanova et al., 2021; Carmon et al., 2022), among many others. A few works exploit the structure of some non-convex functions, such as weak convexity, in order to obtain a convex or strongly convex subproblem for the proximal point method, the first one of which could be (Kaplan and Tichatschke, 1998, Proposition 1). In a similar vein, (Davis and Drusvyatskiy, 2019) used the proximal point method with a regularizer value guaranteeing good properties for the proximal subproblem, and provide guarantees of stationarity for stochastic algorithms.

Millan, Ferreira, and Ugon (2025) developed an analysis showing rates for the Frank-Wolfe method under smoothness and star convexity, a special case of our problem setting.

4 THE ACCELERATED METHOD

Accelerated optimization for first-order methods in convex smooth optimization is an extensively studied topic. Initially considered an algebraic trick devoid of intuition, Nesterov's accelerated gradient descent (Nesterov, 1983) has been explained and generalized via many efforts and interpretations yielding powerful frameworks (Nesterov, 2005; Monteiro and Svaiter, 2013; Zhu and Orecchia, 2017; Levy, Yurtsever, and Cevher, 2018; Joulani, Raj, et al., 2020; Wang, Abernethy, and Levy, 2024). One point of view is that of linear coupling (Zhu and Orecchia, 2017) where the authors reinterpret an optimal algorithm like Nesterov (2005, Section 3) as a combination of gradient descent and an online learning algorithm like mirror descent, where the function value progress of gradient descent compensates the per-iterate regret of mirror descent. This analysis is done separately for unconstrained and constrained cases.

Mirror descent or FTRL online-learning algorithms, cf. (Orabona, 2019), on quasar-convex functions suffer a greater regret with respect to the convex counterpart, by a factor depending on $1/\gamma$. In the *unconstrained* case this remains manageable: both the progress from gradient descent and the regret that one wants to compensate, still become proportional to the squared norm of the gradient at the currently computed point. A suitable coupling between the gradient descent and FTRL points can balance the proportionality constants and restore acceleration, provided one does a low-dimensional search (Hinder, Sidford, and Sohoni, 2020; Guminov, Gasnikov, and Kuruzov, 2023; Nesterov et al., 2021). In the *constrained* case, however, this symmetry breaks: the usual descent proxy does not seem to necessarily compensate the increased regret, and a direct linear-coupling proof does not seem to go through. We lose a degree of freedom on how to choose the coupling.

Our remedy is an inexact implementation of the proximal point step, that is, computing an inexact version of $\mathrm{prox}_\lambda(x)$. This approach offers the possibility of greater descent in exchange for greater computational expense in the step implementation. And further, one step of an approximate proximal point can be seen as an approximate gradient descent step on the associated Moreau envelope, with the property that this gradient descent step always lands in the feasible set $\mathcal{X}$ (Bauschke and Combettes, 2011), analogously to the unconstrained case. This leads us to investigate the properties of accelerated proximal point methods and the associated Moreau envelope for quasar convex smooth problems.

The proximal subproblem of a γ-quasar convex function f is not necessarily quasar convex in general, not even if we allow for a different resulting quasar-convex constant. Indeed, since a center x^* of a quasar convex function is a minimizer, it is enough to consider some quasar convex function plus a small quadratic centered at some point x_0 that shifts the minimizer of the sum towards x_0 with respect to x^*. The func-

tion f does not even necessarily satisfy star convexity around x_0, which may make the sum not quasar convex. Alternatively, see Example 1 for a concrete counterexample, which we display in Figure 1. Hence, the quasar-convex class is not closed under addition, even if we allow the parameter γ to change after the sum. We claim that this non-additivity is natural and holds more generally for relaxations of convexity. Indeed, Nesterov (2018, Section 2.1.1) showed that if a class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfies (A) $\nabla f(x^*) = 0$ implies x^* is a global solution, (B) the class is closed under conic combinations and (C) affine functions belong to the class, then $\mathcal{F}$ is the class of convex functions. As a consequence, there is no class larger than the one of convex functions that satisfies (A) and (B) at the same time. Therefore, relaxations satisfying (A) like quasar convexity, star convexity, tilted convexity, etc. cannot be closed under conic combinations. Similarly, a classical reduction adds the regularizer $x \mapsto \frac{\varepsilon}{R^2}\|x_0 - x\|_2^2$, where R is a bound for the distance from x_0 to a minimizer in order to exploit some strong convexity. The above observation suggests that this reduction is unlikely to hold for many relaxations of convexity satisfying (A). Likewise, computing the proximal point operator for general $\lambda > 0$ may lead to subproblems that either lack a minimizer, fail to have a unique solution, or are non-convex.

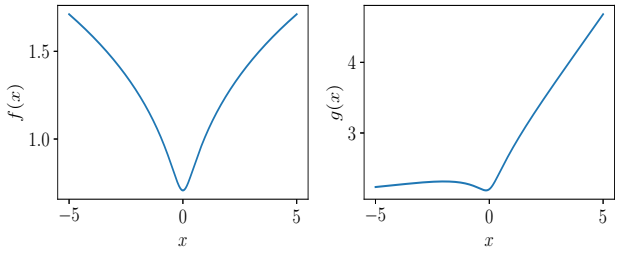

Figure 1: Example 1; a (1/2)-quasar convex function $f(x) \stackrel{\text{def}}{=} (x^2 + 1/8)^{1/6}$ in $[-5, 5]$ (left) and its proximal subproblem $g(x) \stackrel{\text{def}}{=} f(x) + \frac{1}{2\lambda}(x - x_0)^2$, for $\lambda = 50$ and $x_0 \approx -12.23$ (right). The latter has a local maximizer at $x = -2$ and a local minimizer near $x = 0$ and so it cannot be quasar convex in $[-5, 5]$.

The considerations above are negative results in the general case regarding combining functions for classes weaker than convexity. However, there are some special cases that we can exploit: by definition, if we add two functions that are γ-quasar *with respect to the same center* x^*, then the sum becomes γ-quasar-convex. In particular, when adding $\mathbf{1}_{\mathcal{X}}$ to f, provided $x^* \in \mathcal{X}$ and f is γ-quasar-convex on $\mathcal{X}$ with respect to x^*, the composite function $f + \mathbf{1}_{\mathcal{X}}$ preserves quasar-convexity around the same center. Similarly, for (γ, γ_p)-tilted convex problems any restriction to a compact convex set $\mathcal{X}$ results in a γ-quasar convex problem in $\mathcal{X}$. This is the case the problem appearing after the reduction from Riemannian optimization in (Martínez-Rubio, 2022). However, in order to exploit an inexact proximal point step, we further prove several properties of the Moreau envelope. If our original problem is L-smooth, and if we choose $\lambda < 1/L$, we obtain a convex proximal subproblem where the strong convexity from the regularizer makes the subproblem be strongly convex with $O(1)$ condition number, as the following lemma shows.

Lemma 1. [↓] *Let $f : \mathbb{R}^d \to \mathbb{R}$ have L-Lipschitz gradients on $\mathcal{X}$, and $\lambda = 1/(2L)$. For any $x \in \mathcal{X}$, the proximal problem $\min_{y\in\mathcal{X}} f(y) + \frac{1}{2\lambda}\|y - x\|_2^2$ is L-strongly convex, $3L$-smooth, and the associated Moreau envelope $M_{\lambda f + \mathbf{1}_{\mathcal{X}}}$ is $2L$-smooth.*

The property above on the Moreau envelope and the subproblems allows to solve proximal subproblems fast and to approximate Moreau envelope's gradients and function values: it would take $\widetilde{O}(1)$ iterations of projected gradient descent (PGD), cf. (2), to achieve a δ-minimizer of the subproblem (Bubeck, 2015, Section 3.4.2). Now, a key property is that at least when the proximal subproblem has a unique solution, as guaranteed by Lemma 1, the Moreau envelope inherits the γ-quasar-convexity from f.

Proposition 2. [↓] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth γ-quasar convex in a closed convex set $\mathcal{X}$ with respect to a center $x^* \in \mathcal{X}$. For any $\lambda < \frac{1}{L}$, the Moreau envelope $M_{\lambda f + \mathbf{1}_{\mathcal{X}}}$ is γ-quasar convex with respect to x^*.*

Algorithm 1 Constrained Quasar-Cvx Acceleration

Input: γ-quasar convex L-smooth f in a compact convex set $\mathcal{X}$, accuracy ε, initial point x_0, tolerance δ.

$y_0 \leftarrow z_0 \leftarrow x_0$;
$\lambda \leftarrow 1/(2L)$
$T \leftarrow \frac{4}{\gamma}\sqrt{\frac{LD^2}{\varepsilon}}$
for $t \leftarrow 1$ **to** T **do**
  $a_t \leftarrow \gamma^2 t/(8L)$; $A_t = \sum_{i=1}^t a_i = \gamma^2 t(t+1)/(16L)$
  $x_t \leftarrow$ BinaryLineSearch$(f, y_{t-1}, z_{t-1}, L, \delta, \frac{A_{t-1}\gamma}{a_t})$
  $y_t \in \arg\min^{\delta}_{y\in\mathcal{X}}\{f(y) + \frac{1}{2\lambda}\|x_t - y\|_2^2\}$ ◇ Using, for instance, PGD (2).
  $\widetilde{\nabla} M(x_t) \leftarrow \frac{1}{\lambda}(x_t - y_t)$
  $z_t \leftarrow \arg\min_{z\in\mathcal{X}}\{\sum_{i=1}^t \frac{a_i}{\gamma}\langle \widetilde{\nabla} M(x_i), z\rangle + \frac{\|z - x_0\|_2^2}{2}\}$
end for
return $\hat{y}_T \in \arg\min^{\delta}_{y\in\mathcal{X}}\{f(y) + \frac{1}{2\lambda}\|y_T - y\|_2^2\}$.

Since we are only able to approximate the Moreau envelope, we have to deal with errors in our algorithm design and analysis. For the inexact Moreau envelope function value and gradient, omitting the notation's dependence on λ and f, we use:

$$\widetilde{M}(x) \stackrel{\text{def}}{=} f(y_x)+\tfrac{1}{2\lambda}\|y_x-x\|_2^2 \text{ and } \widetilde{\nabla} M(x) \stackrel{\text{def}}{=} \tfrac{1}{\lambda}(x-y_x),$$

where y_x is any δ-minimizer of $y \mapsto f(y)+\frac{1}{2\lambda}\|y-x\|_2^2$ over $\mathcal{X}$. We note that if we computed an exact minimizer, these notions would correspond to the exact function values and gradients for the envelope M.

We set the total number of iterations to $T \stackrel{\text{def}}{=} \frac{4}{\gamma}\sqrt{\frac{LD^2}{\varepsilon}}$ and fix the inner accuracy at $\delta \stackrel{\text{def}}{=} LD^2/(10T^6)$ and thus the time taken for PGD (2) to compute δ-minimizers of proximal subproblems, like y_t in Line 7 of Algorithm 1, is nearly a constant

$$O(\log\big(\frac{L\|x-\text{prox}_\lambda(x)\|_2^2}{\delta}\big)) = O(\log\big(\frac{LD^2}{\varepsilon}\big)) = \widetilde{O}(1).$$

There is a simple stopping criterion that guarantees we achieved δ-optimality, see Lemma 11 in Section B.1.

Our accelerated Algorithm 1 uses three sequences: (i) y_t are approximate proximal points with parameter $\lambda = 1/(2L)$; (ii) z_t are iterates from FTRL that control regret on the approximate envelope gradients; and (iii) $x_t \in \text{conv}\{y_{t-1}, z_{t-1}\}$ are coupling points chosen by a noisy binary search (Algorithm 2), where the noise comes from our inexactness to compute the proximal point and so it is bounded but otherwise it can be arbitrary.

The following lemma is an approximate version to the descent lemma of gradient descent, where the progress made after the inexact gradient descent step $y_t = x_t - \lambda\widetilde{\nabla} M(x_t)$ is proportional to the squared norm of the inexact gradient, up to an additive error.

Lemma 3. [↓] *For $\lambda = 1/(2L)$, and x_t, y_t from Algorithm 1, let $\widetilde{M}(x_t) \stackrel{\text{def}}{=} f(y_t)+\frac{1}{2\lambda}\|y_t-x_t\|_2^2$ and $\widetilde{M}(y_t) = f(\hat{y}_t)+\frac{1}{2\lambda}\|\hat{y}_t-y_t\|_2^2$, for $\hat{y}_t \in \arg\min_{y\in\mathcal{X}}^{\delta}\{f(y)+\frac{1}{2\lambda}\|y-y_t\|_2^2\}$. Then*

$$\widetilde{M}(y_t) - \widetilde{M}(x_t) \le -\frac{1}{8L}\|\widetilde{\nabla} M(x_t)\|_2^2 + \delta.$$

Our accelerated solution also requires to find a point $x_t \in \text{conv}\{y_{t-1}, z_{t-1}\}$ such that an inequality approximately like convexity between y_{t-1} and x_t holds for M. If we had convexity of M and $x_t = \alpha y_{t-1} + (1-\alpha)z_{t-1}$, for some $\alpha \in [0,1]$, we would have $\langle \nabla M(x_t), x_t - z_{t-1}\rangle = \frac{\alpha}{1-\alpha}\langle \nabla M(x_t), y_{t-1} - x_t\rangle \le \frac{\alpha}{1-\alpha}(M(y_{t-1}) - M(x_t))$ and varying $\alpha \in [0,1]$ we could get any positive weight on the right hand side. Instead, we guarantee a similar property on $\widetilde{M}$ by performing a binary search over the segment $\text{conv}\{y_{t-1}, z_{t-1}\}$, but we only have access to noisy evaluations of the function and gradient, which could potentially make the binary search branch to the wrong path. However, we keep an invariant that, by using smoothness of the objective, guarantees the maintained interval in the search contains an interval satisfying the termination condition that is large enough to guarantee early termination. The oracles $h, \hat{h}$ required by Algorithm 2 are implemented via $\widetilde{M}(\cdot)$ and $\widetilde{\nabla} M(\cdot)$, respectively.

Algorithm 2 Binary Line Search(f, y, z, L, δ, c)

Input: L-smooth function f with domain $\mathcal{X}$, points y, z, smoothness constant L, tolerance δ, constant $c \ge 0$. Define $\lambda \stackrel{\text{def}}{=} 1/(2L)$, $g(\alpha) \stackrel{\text{def}}{=} M_{\lambda f+\mathbf{1}_\mathcal{X}}(\alpha y + (1-\alpha)z)$. Assume access to $h(\alpha), \hat{h}(\alpha) \in \mathbb{R}$ satisfying $g(\alpha) \le h(\alpha) \le g(\alpha)+\delta_1$ and $|g'(\alpha)-\hat{h}(\alpha)| \le \delta_2$. Target error: $\tilde{\varepsilon}$, cf. (13).

```
if ĥ(1) ≤ ε̃ then return 1
if h(1) − h(0) ≥ −δ_1 then return 0
a_0 ← 0;  b_0 ← 1;  α ← (a_0 + b_0)/2;  k ← 0
while α ĥ(α) > c(h(1) − h(0)) + ε̃ do
    if h(α) ≥ h(1) − δ_1 then
        a_{k+1} ← α;  b_{k+1} ← b_k
    else
        a_{k+1} ← a_k;  b_{k+1} ← α
    end if
    k ← k + 1
    α ← (a_k + b_k)/2;
end while
return αy + (1 − α)z
```

Lemma 4. [↓] *With the notation of Lemma 3 and Line 6 of Algorithm 1, and for any target constant $c \ge 0$, then Algorithm 2 returns a point x_t satisfying*

$$\langle \widetilde{\nabla} M(x_t), x_t - z_{t-1}\rangle \le c(\widetilde{M}(y_{t-1}) - \widetilde{M}(x_t)) + \sqrt{8LD^2\delta} + (9+5c)\delta,$$

and it computes x_t after δ-approximating a $\text{prox}_{\lambda f+\mathbf{1}_\mathcal{X}}$ at most $O\left(\log\left(\frac{LD^2}{\delta}\right)\right)$ times.

Since each δ-approximation of a $\text{prox}_{\lambda f+\mathbf{1}_\mathcal{X}}$ takes at most $O(\log(LD^2/\delta))$ iterations for PGD, Line 6 of Algorithm 1 can be implemented in $\widetilde{O}(1)$ first-order oracle queries. Using these tools, we obtain our near-optimal constrained accelerated method. The proof can be found in Section B.

Theorem 5. [↓] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and γ-quasar convex with center x^* in a compact convex set*

$\mathcal{X}$ of diameter D. Algorithm 1 obtains an ε-minimizer in

$$\widetilde{O}\left(\frac{1}{\gamma}\sqrt{\frac{LD^2}{\varepsilon}}\right)$$

queries to a first-order oracle of f.

5 UNACCELERATED METHODS

In this section, we present an analysis for the convergence of the projected gradient descent method as well as for the Frank-Wolfe algorithm, for L-smooth γ-quasar convex objectives with constraints.

Projected gradient descent minimizes a quadratic model of the function, which is an upper bound on the function, if $\eta \leq 1/L$. The update rule, starting from a point x, is the following:

$$x^+ \leftarrow \arg\min_{y\in\mathcal{X}} \left\{ f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\eta}\|y-x\|_2^2 \right\} = \mathrm{Proj}_{\mathcal{X}}(x - \eta\nabla f(x)). \tag{2}$$

We note that the problem defining x^+ is strongly convex and thus the solution is unique. Define the gradient mapping with respect to f as

$$g_{\mathcal{X}}(x) \stackrel{\text{def}}{=} \frac{1}{\eta}(x - x^+).$$

We do not make the dependence on η explicit in the notation since it will be clear from context.

Our analysis works via making use of the gradient mapping and its properties, which we state in the next two lemmas. These are known facts about the gradient mapping (Nesterov, 2004) but we provide proofs in Section C.1 for completeness.

Lemma 6. [↓] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function in a closed convex set $\mathcal{X}$. For any $x, y \in \mathcal{X}$, we have*

$$\langle \nabla f(x), x^+ - y\rangle \leq \langle g_{\mathcal{X}}(x), x^+ - y\rangle.$$

Lemma 7. [↓] *Let $\mathcal{X}$ be a closed convex set and let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth in $\mathcal{X}$. Starting at $x_t \in \mathcal{X}$, we compute x_{t+1} from x_t via (2) with $\eta = 1/L$. Then*

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|g_{\mathcal{X}}(x_t)\|_2^2.$$

Now we present our result for quasar convex objectives. A technical innovation in both projected gradient descent and the Frank-Wolfe analyses is that we devise a lower bound estimate of $f(x^*)$ that can be obtained from the algorithms, that differs from classical Frank-Wolfe analyses or analyses of PGD using the gradient mapping. This bound allows to shift certain weights by some indices to allow for errors coming from quasar convexity to be canceled.

Surprisingly, the structure of this new lower bound is very similar for PGD's and Frank-Wolfe's analyses. See Section C.

Theorem 8 (Projected Gradient Descent Rate)**.** [↓] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and γ-quasar convex with center x^* in a compact convex set $\mathcal{X}$. Let $D \stackrel{\text{def}}{=} \mathrm{diam}(\{x \in \mathcal{X} \mid f(x) \leq f(x_0)\})$ Then, for all $t \geq 1$, the iterates of projected gradient descent (2) with step size $\eta = 1/L$ satisfy*

$$f(x_t) - f(x^*) \leq \frac{20LD^2}{(t+1)\gamma^2}.$$

Now we present our result on the Frank-Wolfe algorithm, whose pseudocode can be seen in Section C.2. This algorithm operates by solving linear subproblems rather than projections, which are quadratic problems. The algorithm is not modified with respect to the classical algorithm in convex optimization, but our analysis differs to deal with the quasar convex assumption.

Theorem 9 (Frank-Wolfe Rate)**.** [↓] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and γ-quasar convex with center x^* in a compact convex set $\mathcal{X}$ of diameter D. Then, for all $t \geq 1$, the iterates of the Frank-Wolfe Algorithm 3 satisfy*

$$f(x_t) - f(x^*) \leq \frac{6LD^2}{(t+1)\gamma^2}.$$

Remark 10. *An interesting property of both the results in Theorem 8 and Theorem 9 is that the algorithms achieve their convergence rates without requiring any knowledge of the parameter γ. The fact that the algorithms adapt automatically to the unknown γ highlights their robustness and makes the results easily applicable in practical machine learning settings.*

6 CONCLUSION

In this work, we addressed the open question (Martínez-Rubio, 2022; Lezane, Langer, and Koolen, 2024) of whether smooth quasar-convex functions can be efficiently optimized in the presence of general convex constraints.

Prior work established accelerated rates in the unconstrained case, but these solutions did not work for the constrained case due to the loss of one degree of freedom in the algorithm's design. Our main contribution consisted of closing this gap: we presented

an accelerated first-order method that achieves an ε-approximate solution in $\widetilde{O}\left(\frac{1}{\gamma}\sqrt{\frac{LD^2}{\varepsilon}}\right)$ first-order oracle queries, and we further extended the scope of quasar-convex optimization beyond unconstrained domains analyzing unaccelerated algorithms. As a result of our accelerated method, we improved over prior Riemannian optimization algorithms via a reduction to quasar-convex optimization (Martínez-Rubio, 2022).

Our results provide a new tools for analyzing and understanding algorithms in structured nonconvex optimization, which may shed further light on why simple gradient-based algorithms are so effective in large-scale machine learning.

Acknowledgements

Project supported by a 2025 Leonardo Grant for Scientific Research and Cultural Creation from the BBVA Foundation.

References

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). "A convergence theory for deep learning via over-parameterization". In: *International conference on machine learning*. PMLR, pp. 242–252 (cit. on p. 1).

Bauschke, Heinz H. and Patrick L. Combettes (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer. ISBN: 978-1-4419-9466-0 (cit. on p. 4).

Bertsekas, Dimitri, Angelia Nedic, and Asuman Ozdaglar (2003). "Convex analysis and optimization". In: vol. 1. Athena Scientific, pp. 245–247. ISBN: 9781886529458 (cit. on p. 3).

Bhojanapalli, Srinadh, Behnam Neyshabur, and Nati Srebro (2016). "Global Optimality of Local Search for Low Rank Matrix Recovery". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, pp. 3873–3881 (cit. on p. 1).

Bubeck, Sébastien (2015). "Convex Optimization: Algorithms and Complexity". In: *Found. Trends Mach. Learn.* 8.3-4, pp. 231–357 (cit. on pp. 5, 13).

Carmon, Yair, Arun Jambulapati, Yujia Jin, and Aaron Sidford (2022). "RECAPP: Crafting a More Efficient Catalyst for Convex Optimization". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 2658–2685 (cit. on p. 4).

Davis, Damek and Dmitriy Drusvyatskiy (2019). "Stochastic Model-Based Minimization of Weakly Convex Functions". In: *SIAM J. Optim.* 29.1, pp. 207–239 (cit. on p. 4).

Du, Simon S., Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai (2019). "Gradient Descent Finds Global Minima of Deep Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1675–1685 (cit. on p. 2).

Du, Simon S., Xiyu Zhai, Barnabás Póczos, and Aarti Singh (2019). "Gradient Descent Provably Optimizes Over-parameterized Neural Networks". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net (cit. on p. 2).

Even, Mathieu, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Pierre Gaillard, Hadrien Hendrikx, Laurent Massoulié, and Adrien Taylor (2021). "A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip". In: *Advances in Neural Information Processing Systems*, pp. 1–32 (cit. on p. 3).

Foster, Dylan J., Ayush Sekhari, and Karthik Sridharan (2018). "Uniform Convergence of Gradients for Non-Convex Learning and Optimization". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8759–8770 (cit. on pp. 2, 4).

Frostig, Roy, Rong Ge, Sham M. Kakade, and Aaron Sidford (2015). "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 2540–2548 (cit. on p. 4).

Fu, Qiang, Dongchu Xu, and Ashia Camage Wilson (2023). "Accelerated Stochastic Optimization Methods under Quasar-convexity". In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 10431–10460 (cit. on p. 3).

Ge, Rong, Furong Huang, Chi Jin, and Yang Yuan (2015). "Escaping From Saddle Points - Online Stochastic Gradient for Tensor Decomposition". In: *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*. Ed. by Peter Grünwald, Elad Hazan, and Satyen

Kale. Vol. 40. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 797–842 (cit. on p. 1).

Ge, Rong, Jason D. Lee, and Tengyu Ma (2016). "Matrix Completion has No Spurious Local Minimum". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, pp. 2973–2981 (cit. on p. 1).

Gower, Robert M., Othmane Sebbouh, and Nicolas Loizou (2021). "SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation". In: *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 1315–1323 (cit. on p. 3).

Guminov, Sergey, Alexander V. Gasnikov, and Ilya A. Kuruzov (2023). "Accelerated methods for weakly-quasi-convex optimization problems". In: *Comput. Manag. Sci.* 20.1, p. 36 (cit. on pp. 3, 4).

Hardt, Moritz, Tengyu Ma, and Benjamin Recht (2018). "Gradient Descent Learns Linear Dynamical Systems". In: *J. Mach. Learn. Res.* 19, 29:1–29:44 (cit. on pp. 2–4).

Hinder, Oliver, Aaron Sidford, and Nimit Sharad Sohoni (2019). "Near-Optimal Methods for Minimizing Star-Convex Functions and Beyond". In: *Optimization for Machine Learning Workshop 2019, December 14, 2019, Vancouver, Canada* (cit. on pp. 1, 2).

– (2020). "Near-Optimal Methods for Minimizing Star-Convex Functions and Beyond". In: *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 1894–1938 (cit. on pp. 3, 4).

Ivanova, Anastasiya, Dmitry Pasechnyuk, Dmitry Grishchenko, Egor Shulgin, Alexander V. Gasnikov, and Vladislav Matyukhin (2021). "Adaptive Catalyst for Smooth Convex Optimization". In: *Optimization and Applications - 12th International Conference, OPTIMA 2021, Petrovac, Montenegro, September 27 - October 1, 2021, Proceedings*. Ed. by Nicholas N. Olenev, Yuri G. Evtushenko, Milojica Jacimovic, Michael Yu. Khachay, and Vlasta Malkova. Vol. 13078. Lecture Notes in Computer Science. Springer, pp. 20–37 (cit. on p. 4).

Jin, Jikai (2020). "On The Convergence of First Order Methods for Quasar-Convex Optimization". In: *CoRR* abs/2010.04937 (cit. on p. 3).

Joulani, Pooria, András György, and Csaba Szepesvári (2020). "A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, variance reduction, and variational bounds". In: *Theor. Comput. Sci.* 808, pp. 108–138 (cit. on p. 4).

Joulani, Pooria, Anant Raj, András György, and Csaba Szepesvári (2020). "A simpler approach to accelerated optimization: iterative averaging meets optimism". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4984–4993 (cit. on p. 4).

Kaplan, Alexander and Rainer Tichatschke (1998). "Proximal Point Methods and Nonconvex Optimization". In: *J. Glob. Optim.* 13.4, pp. 389–406 (cit. on p. 4).

Karimi, Hamed, Julie Nutini, and Mark Schmidt (2016). "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*. Ed. by Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken. Vol. 9851. Lecture Notes in Computer Science. Springer, pp. 795–811 (cit. on p. 1).

Kleinberg, Robert, Yuanzhi Li, and Yang Yuan (2018). "An Alternative View: When Does SGD Escape Local Minima?" In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2703–2712 (cit. on p. 2).

Levy, Kfir Yehuda, Alp Yurtsever, and Volkan Cevher (2018). "Online Adaptive Methods, Universality and Acceleration". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, pp. 6501–6510 (cit. on p. 4).

Lezane, Clément, Sophie Langer, and Wouter M Koolen (2024). "Accelerated Mirror Descent for Non-Euclidean Star-convex Functions". In: *arXiv preprint arXiv:2405.18976* (cit. on pp. 1–3, 7).

Lin, Hongzhou, Julien Mairal, and Zaïd Harchaoui (2017). "Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice". In: *J. Mach. Learn. Res.* 18, 212:1–212:54 (cit. on p. 4).

Martínez-Rubio, David (2022). "Global Riemannian Acceleration in Hyperbolic and Spherical Spaces". In: *International Conference on Algorithmic Learning Theory, 29 March - 1 April 2022, Paris, France*.

Vol. 167. Proceedings of Machine Learning Research. PMLR, pp. 768–826 (cit. on pp. 1–5, 7, 8).

Millan, R Diaz, Orizon Pereira Ferreira, and Julien Ugon (2025). "Frank-Wolfe algorithm for star-convex functions". In: *arXiv preprint arXiv:2507.17272* (cit. on p. 4).

Monteiro, Renato D. C. and Benar Fux Svaiter (2013). "An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and Its Implications to Second-Order Methods". In: *SIAM J. Optim.* 23.2, pp. 1092–1125 (cit. on p. 4).

Narkiss, Guy and Michael Zibulevsky (2005). *Sequential subspace optimization method for large-scale unconstrained problems*. Technion-IIT, Department of Electrical Engineering (cit. on p. 3).

Nemirovski, Arkadi (1982a). [Online; accessed 25-September-2025] (cit. on p. 3).

– (1982b). "Orth-method for smooth convex optimization". In: *Izvestia AN SSSR, Transl.: Eng. Cybern. Soviet J. Comput. Syst. Sci* 2, pp. 937–947 (cit. on p. 3).

Nesterov, Yurii (1983). "A method for solving the convex programming problem with convergence rate O $(1/k^2)$". In: *Dokl. akad. nauk Sssr*. Vol. 269. 3, pp. 543–547 (cit. on p. 4).

– (2018). *Lectures on convex optimization*. Vol. 137. Springer (cit. on p. 5).

Nesterov, Yurii, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky (2021). "Primal–dual accelerated gradient methods with small-dimensional relaxation oracle". In: *Optimization Methods and Software* 36.4, pp. 773–810 (cit. on pp. 3, 4).

Nesterov, Yurii E. (2004). *Introductory Lectures on Convex Optimization - A Basic Course*. Vol. 87. Applied Optimization. Springer. ISBN: 978-1-4613-4691-3 (cit. on p. 7).

– (2005). "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1, pp. 127–152 (cit. on p. 4).

Nguyen, Quynh and Marco Mondelli (2020). "Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (cit. on p. 2).

Orabona, Francesco (2019). "A Modern Introduction to Online Learning". In: *CoRR* abs/1912.13213 (cit. on p. 4).

Park, Dohyung, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi (2017). "Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*. Ed. by Aarti Singh and Xiaojin (Jerry) Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 65–74 (cit. on p. 1).

Saad, El Mehdi, Weicheng Lee, and Francesco Orabona (2025). "New Lower Bounds for Stochastic Non-Convex Optimization through Divergence Composition". In: *arXiv preprint arXiv:2502.14060* (cit. on p. 3).

Sun, Ju, Qing Qu, and John Wright (2017). "Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture". In: *IEEE Trans. Inf. Theory* 63.2, pp. 853–884 (cit. on p. 1).

– (2018). "A Geometric Analysis of Phase Retrieval". In: *Found. Comput. Math.* 18.5, pp. 1131–1198 (cit. on p. 1).

Tran, Hoang, Qinzi Zhang, and Ashok Cutkosky (2024). "Empirical Tests of Optimization Assumptions in Deep Learning". In: *CoRR* abs/2407.01825 (cit. on p. 2).

Vaswani, Sharan, Benjamin Dubois-Taine, and Reza Babanezhad (2022). "Towards Noise-adaptive, Problem-adaptive (Accelerated) Stochastic Gradient Descent". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 22015–22059 (cit. on p. 3).

Wang, Jun-Kun, Jacob D. Abernethy, and Kfir Y. Levy (2024). "No-regret dynamics in the Fenchel game: a unified framework for algorithmic convex optimization". In: *Math. Program.* 205.1, pp. 203–268 (cit. on p. 4).

Wang, Jun-Kun and Andre Wibisono (2023). "Continuized Acceleration for Quasar Convex Functions in Non-Convex Optimization". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net (cit. on pp. 2–4).

Zhou, Yi, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh (2019). "SGD Converges to Global Minimum in Deep Learning via Star-convex Path". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net (cit. on p. 2).

Zhu, Zeyuan Allen and Lorenzo Orecchia (2017). "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent". In: *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*. Ed. by Christos H. Papadimitriou. Vol. 67. LIPIcs. Schloss

Dagstuhl - Leibniz-Zentrum für Informatik, 3:1–3:22 (cit. on p. 4).

Zou, Difan, Yuan Cao, Dongruo Zhou, and Quanquan Gu (2018). "Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks". In: *CoRR* abs/1811.08888 (cit. on p. 2).

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Fast Quasar-Convex Optimization with Constraints: Supplementary Materials

A PROPERTIES

Proof. (Lemma 1) Since f has L-Lipschitz gradients, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{L}{2}\|x - y\|_2^2 \text{ for all } x, y \in \mathcal{X}.$$

Thus, adding the $2L$ strong convexity inequality for $\phi_z(y) = L\|x - y\|_2^2$ for any $z \in \mathbb{R}^d$, we obtain:

$$(f + \phi_z)(y) \geq (f + \phi_z)(x) + \langle \nabla (f + \phi_z)(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2 \text{ for all } x, y \in \mathcal{X}.$$

So $f + \phi_z$ is L-strongly convex. A similar argument shows that it is $3L$-smooth, since ϕ_z has $2L$-Lipschitz gradients.

Finally, for $\lambda = 1/(2L)$, denote $M \overset{\text{def}}{=} M_{\lambda f + \mathbf{1}_{\mathcal{X}}}$ for ease of notation. Note that since the proximal subproblem is strongly convex, we have by the first-order optimality condition that

$$\nabla f(\text{prox}_\lambda(x)) + g_{\text{prox}_\lambda(x)} + 2L(\text{prox}_\lambda(x) - x) = 0,$$

where $\text{prox}_\lambda(x) \overset{\text{def}}{=} \arg\min_{y \in \mathcal{X}} \{f(y) + \frac{1}{2\lambda}\|y - x\|_2^2\}$ and $g_{\text{prox}_\lambda(x)} \in \partial \mathbf{1}_{\mathcal{X}}(\text{prox}_\lambda(x))$. Using this equation in ② below and by the envelope theorem in ①, we have

$$\nabla M(x) \overset{①}{=} \frac{1}{\lambda}(x - \text{prox}_\lambda(x)) \overset{②}{=} \nabla f(\text{prox}_\lambda(x)) + g_{\text{prox}_\lambda(x)}. \tag{3}$$

Obtaining the smoothness of the Moreau envelope is standard: the proof does not rely on the convexity of f. Indeed, by definition, $\frac{1}{\lambda}$-smoothness means that the isotropic quadratic with leading coefficient $\frac{1}{2\lambda}$ whose zeroth- and first-order information coincide with that one of M at x, is above M. The quadratic we are looking for takes the form $Q_y : \mathbb{R}^d \to \mathbb{R}, x \mapsto f(y) + \mathbf{1}_{\mathcal{X}}(y) + \frac{1}{2\lambda}\|x - y\|_2^2$ for $y = \text{prox}_\lambda(x)$, for which indeed $M(x) = Q_y(x)$ and $\nabla M(x) = \nabla Q_y(x)$. Moreover, for any y, it is $M \leq Q_y$ by the definition of M as a min. Alternatively, we can provide this argument in an algebraic form:

$$\begin{aligned}
M(y) &\overset{①}{=} f(\text{prox}_\lambda(y)) + \mathbf{1}_X(\text{prox}_\lambda(y)) + \frac{1}{2\lambda}\|\text{prox}_\lambda(y) - y\|_2^2 \\
&\overset{②}{\leq} f(\text{prox}_\lambda(x)) + \mathbf{1}_X(\text{prox}_\lambda(y)) + \frac{1}{2\lambda}\|\text{prox}_\lambda(x) - y\|_2^2 \\
&\overset{③}{=} M(x) - \frac{1}{2\lambda}\|\text{prox}_\lambda(x) - x\|_2^2 + \frac{1}{2\lambda}\|\text{prox}_\lambda(x) - y\|_2^2 \\
&\overset{④}{=} M(x) - \frac{1}{\lambda}\langle \text{prox}_\lambda(x) - x, y - x \rangle + \frac{1}{2\lambda}\|y - x\|_2^2 \\
&\overset{⑤}{=} M(x) + \langle \nabla M(x), y - x \rangle + \frac{1}{2\lambda}\|x - y\|_2^2,
\end{aligned} \tag{4}$$

where ① and ③ use the definition of the Moreau envelope and the prox, ② holds by the optimality of $\text{prox}_\lambda(y)$, ④ is the cosine theorem and ⑤ uses the expression of $\nabla M(x)$ the gradient that we computed in (3).

□

Proof. (Proposition 2) Recall that we have the following, for $\lambda = 1/(2L)$, cf. (3):

$$\nabla M(x) = 2L(x - \mathrm{prox}_\lambda(x)) = \nabla f(\mathrm{prox}_\lambda(x)) + g_{\mathrm{prox}_\lambda(x)}.$$

Note that we can use the reasoning leading to (3) since f is assumed to be L-smooth.

Using quasar convexity of f and $\mathbf{1}_{\mathcal{X}}$ (convexity implies γ-quasar convexity centered at a minimizer, $\gamma \in (0,1]$, and for the convex $\mathbf{1}_{\mathcal{X}}$, every point in $\mathcal{X}$ is a minimizer), we obtain:

$$\begin{aligned}
M(x^*) &= f(x^*) + \mathbf{1}_{\mathcal{X}}(x^*) \\
&\geq f(\mathrm{prox}_\lambda(x)) + \frac{1}{\gamma}\langle \nabla f(\mathrm{prox}_\lambda(x)), x^* - \mathrm{prox}_\lambda(x)\rangle + \mathbf{1}_{\mathcal{X}}(\mathrm{prox}_\lambda(x)) + \frac{1}{\gamma}\langle g_{\mathrm{prox}_\lambda(x)}, x^* - \mathrm{prox}_\lambda(x)\rangle \\
&= \left(M(x) - L\|x - \mathrm{prox}_\lambda(x)\|_2^2\right) + \frac{1}{\gamma}\langle \nabla M(x), x^* - x\rangle + \frac{1}{\gamma}\langle \nabla M(x), x - \mathrm{prox}_\lambda(x)\rangle \\
&= M(x) + \frac{1}{\gamma}\langle \nabla M(x), x^* - x\rangle + L\left(\frac{2}{\gamma} - 1\right)\|x - \mathrm{prox}_\lambda(x)\|_2^2 \\
&\geq M(x) + \frac{1}{\gamma}\langle \nabla M(x), x^* - x\rangle,
\end{aligned} \tag{5}$$

where we used $M(x^*) = f(x^*) = f(x^*) + \mathbf{1}_{\mathcal{X}}(x^*)$ which holds since $\mathbf{1}_{\mathcal{X}} \equiv 0$ in $\mathcal{X}$ and

$$f(x^*) \overset{①}{\leq} f(\mathrm{prox}_\lambda(x^*)) + \frac{1}{2\lambda}\|x^* - \mathrm{prox}_\lambda(x^*)\|_2^2 \overset{②}{=} M(x^*) = \min_{y\in\mathcal{X}}\{f(y) + \frac{1}{2\lambda}\|x^* - y\|_2^2\} \overset{③}{\leq} f(x^*),$$

where ① uses x^* is a minimizer of f, ② holds by definition of $\mathrm{prox}_\lambda(\cdot)$, and ③ holds by substituting y by x^*. Note that the inequality above also shows that $\mathrm{prox}_\lambda(x^*) = x^*$ since the inequalities become equalities. □

Lemma 11. *Let F be an L-strongly convex function with minimizer at y^*, that is also $3L$-smooth as in our Lemma 1. Then, if we have a point y such that $\|\nabla F(y)\|_2 \leq \sqrt{2L\delta}$, then $F(y) - F(y^*) \leq \delta$. PGD starting at distance at most D from y^* obtains such a point in at most $O(\log(LD^2/\delta))$ iterations.*

Proof. The following holds for a function F with minimizer at y^* that is $\widetilde{\mu}$-strongly convex and $\widetilde{L}$ smooth:

$$\frac{1}{2\widetilde{L}}\|\nabla F(y)\|_2^2 \leq F(y) - F(y^*) \leq \frac{1}{2\widetilde{\mu}}\|\nabla F(y)\|_2^2. \tag{6}$$

For the function we consider, it is $\widetilde{\mu} \leftarrow L$ and $\widetilde{L} \leftarrow 3L$. Thus, if we detect that y satisfies $\|\nabla F(y)\|_2 \leq \sqrt{2L\delta}$, then we have $F(y) - F(y^*) \leq \delta$. At the same time, we have this gradient norm guarantee whenever $F(y) - F(y^*) \leq \delta/3$, by the first inequality above.

Since PGD starting at y_0 takes at most $O(\frac{\widetilde{L}}{\widetilde{\mu}}\log(\frac{\widetilde{L}\|y_0 - y^*\|_2^2}{\hat{\delta}}))$ iterations to obtain a $\hat{\delta}$-minimizer, cf. (Bubeck, 2015, Section 3.4.2) for instance, then for our function and $\hat{\delta} = \delta/3$, we take $O(\log\frac{LD^2}{\delta})$ iterations to find a gradient with norm at most $\sqrt{2L\delta}$ guaranteeing a function gap of at most δ. □

Example 1. *We consider $f(x) = (x^2 + \frac{1}{8})^{1/6}$ with center $x^* = 0$ and show that it is $\frac{1}{2}$-quasar convex in $[-5, 5]$. Such a property means:*

$$(x^2 + \frac{1}{8})^{1/6} - \frac{1}{8^{1/6}} = f(x) - f(x^*) \leq 2xf'(x) = \frac{2}{3}(x^2 + \frac{1}{8})^{-5/6}x^2, \qquad \text{for } x \in [-5, 5].$$

Showing the inequality above is simple, by a change of variable $t = (x^2 + \frac{1}{8})^{1/6}$ (yielding $x^2 = t^6 - \frac{1}{8}$) gives that it is enough to show, for $t \in I \overset{\text{def}}{=} [\frac{1}{8^{1/6}}, (25 + \frac{1}{8})^{1/6}]$, that the minimum of the following function is at least 0:

$$g(t) \overset{\text{def}}{=} \frac{2}{3}t^{-5}(t^6 - \frac{1}{8}) - t + \frac{1}{8^{1/6}} \overset{?}{\geq} 0 \text{ for all } t \in I.$$

This function is concave since $g''(t) = -\frac{5}{2}t^{-7} \leq 0$ and $I \subset \mathbb{R}_{>0}$, and so it is enough to check the endpoints of the interval to look for the minimum, which indeed is nonnegative.

It is direct to check that $g(x) \overset{\text{def}}{=} f(x) + \frac{1}{2\lambda}(x - x_0)^2$ for $\lambda = 50$ and a $x_0 \approx -12.23$, there is a local maximizer of g at $x = -2$, by computing the value of x_0 that makes the derivative of g be 0 at $x = -2$. In that case $g''(-2) \leq 0$.

B ACCELERATED CONSTRAINED METHOD FOR SMOOTH QUASAR-CONVEX OPTIMIZATION

Here we present the proof of our main theorem, the proofs of Lemmas 3 and 4, that we use for the theorem, follow.

Proof. (Theorem 5) Throughout this proof, we use the value $\lambda = 1/(2L)$ and denote $M(\cdot) \stackrel{\text{def}}{=} M_{\lambda f + \mathbf{1}_{\mathcal{X}}}(\cdot)$ and $\text{prox}(\cdot)$ its corresponding proximal operator, without making explicit the dependence on λ, f and $\mathcal{X}$. Denote the diameter of the feasible set $\mathcal{X}$ by D.

Denote $F_k(y) \stackrel{\text{def}}{=} f(y) + \frac{1}{2\lambda}\|y - x_k\|_2^2$. For $\delta > 0$ to be chosen later, assume that we optimize F_k so that $\|y_k - y_k^*\|_2 \le \delta_1$, and so $F_k(y_k) - \delta_2 \le F_k(y_k^*) = M(x_k) \le F_k(y_k)$, where $y_k^* = \arg\min_y F_k(y) = \text{prox}_\lambda(x_k)$ for $\delta_2 = \delta$. Since the condition number of $F_k(y_k)$ is $O(1)$, we can do so with PGD in $O(\log(\frac{LD^2}{\delta}))$ iterations, cf. Lemma 11.

We denote our approximation of $\nabla M(x_k)$ by $\widetilde{\nabla} M(x_k) \stackrel{\text{def}}{=} \frac{1}{\lambda}(x_k - y_k)$, as in Line 8 of Algorithm 1. Also denote by $\widetilde{M}(x_k) \stackrel{\text{def}}{=} f(y_k) + \frac{1}{2\lambda}\|y_k - x_k\|_2^2$ the approximation of the Moreau envelope value. Using the approximate optimality condition of y_k and optimality of y_k^*, that is, $F_k(y_k) - F_k(y_k^*) \le \delta$, we obtain

$$\widetilde{M}(x_k) - \delta = F_k(y_k) - \delta \le F_k(y_k^*) = M(x_k) \le F_k(y_k) \tag{7}$$

$$\|\widetilde{\nabla} M(x_k) - \nabla M(x_k)\|_2 \le \frac{1}{\lambda}\|y_k - y_k^*\|_2 \overset{①}{\le} \sqrt{8L\delta}, \tag{8}$$

where in ① we that F_k is L-strongly convex, cf. Lemma 1, and so $\frac{L}{2\lambda^2}\|y_k - y_k^*\|_2^2 \le \frac{1}{\lambda^2}(F_k(y_k) - F_k(y_k^*)) \le 4L^2\delta$.

Let $a_t > 0$ for $t \ge 1$, whose value will be decided later, and $A_t \stackrel{\text{def}}{=} A_{t-1} + a_t = \sum_{i=1}^t a_i$. Denote $A_0 = 0$ and $A_0 L_0 = 0$. We first build the following lower bound L_t on $f(x^*) = M(x^*)$, for $t \ge 1$:

$$\begin{aligned}
A_t M(x^*) &\overset{①}{\ge} \sum_{i=1}^t a_i M(x_i) + \frac{a_i}{\gamma}\langle \nabla M(x_i), x^* - x_i\rangle \\
&\overset{②}{\ge} (-A_t\delta + \sum_{i=1}^t a_i \widetilde{M}(x_i)) + \sum_{i=1}^t \frac{a_i}{\gamma}\langle \widetilde{\nabla} M(x_i), x^* - x_i\rangle \pm \frac{1}{2}\|x^* - x_0\|_2^2 - \sqrt{8L\delta}\frac{A_t D}{\gamma} \\
&\overset{③}{\ge} (-A_t\delta + \sum_{i=1}^t a_i \widetilde{M}(x_i)) + \sum_{i=1}^t \frac{a_i}{\gamma}\langle \widetilde{\nabla} M(x_i), z_t - x_i\rangle + \frac{1}{2}\|z_t - x_0\|_2^2 \\
&\quad - \frac{1}{2}\|x^* - x_0\|_2^2 - \sqrt{8L\delta}\frac{A_t D}{\gamma} \\
&\stackrel{\text{def}}{=} A_t L_t.
\end{aligned}$$

where above, ① holds by the γ-quasar convexity of M. In ②, we used (7), and we also added and subtracted our approximate gradients, and used Cauchy-Schwartz along with (8) and $x^*, x_i \in \mathcal{X}$ to obtain

$$-\sum_{i=1}^t \frac{a_i}{\gamma}\langle \nabla M(x_i) - \widetilde{\nabla} M(x_i), x^* - x_i\rangle \ge -\sqrt{8L\delta}\frac{A_t D}{\gamma}.$$

We also added and subtracted half the initial squared distance to x^* to use it in ③, where we bound some x^* away by using the definition of z_t. Now, we define the gap $G_t \stackrel{\text{def}}{=} \widetilde{M}(y_t) - L_t$. For $t \ge 1$, we are going to bound

$$A_t G_t - A_{t-1} G_{t-1} - \mathbb{I}(t=1)\frac{1}{2}\|x_0 - x^*\|_2^2 \le E_t, \tag{9}$$

for some value E_t, where $\mathbb{I}(A)$ denotes the event indicator that is 1 if A holds true and 0 otherwise. In the sequel, we use the notation $\widetilde{M}(y_t) \stackrel{\text{def}}{=} f(\hat{y}_t) + \frac{1}{2\lambda}\|\hat{y}_t - y_t\|_2^2$ for $\hat{y}_t \in \arg\min_y^\delta \{f(y) + \frac{1}{2\lambda}\|y - y_t\|_2^2\}$. For the corresponding

$\hat{y}_T$, and concatenating (9), we obtain:

$$f(\hat{y}_T) - f(x^*) \leq \widetilde{M}(y_T) - M(x^*) \leq G_T \leq \frac{1}{A_T}\left(\frac{1}{2}\|x^* - x_0\|_2^2 + \sum_{i=1}^{T} E_i\right). \tag{10}$$

Thus, we want A_t to grow rapidly whereas E_t should be moderate. We note that the extra term at $t = 1$ in (9) is due to the lower bound $A_1 L_1$, which is not canceled by $A_0 L_0 \stackrel{\text{def}}{=} 0$, whereas for other values of t this same term appears in both lower bounds $A_t L_t$ and $A_{t-1} L_{t-1}$, so they cancel. We now show (9), for all $t \geq 1$:

$$\begin{aligned}
A_t G_t - A_{t-1} G_{t-1} - \mathbb{I}(t=1)\frac{1}{2}\|x_0 - x^*\|_2^2 &\stackrel{①}{\leq} A_t(\widetilde{M}(y_t) - \widetilde{M}(x_t)) + \cancel{a_t \widetilde{M}(x_t)} + A_{t-1}(\widetilde{M}(x_t) - \widetilde{M}(y_{t-1})) \\
&- \left(\cancel{\sum_{i=1}^{t} a_i \widetilde{M}(x_i)} + \sum_{i=1}^{t-1} \frac{a_i}{\gamma}\langle \widetilde{\nabla} M(x_i), z_t - x_i\rangle + \frac{\|z_t - x_0\|_2^2}{2}\right) + A_t\delta + \frac{A_t D}{\gamma}\sqrt{8L\delta} - \frac{a_t}{\gamma}\langle \widetilde{\nabla} M(x_t), z_t - x_t\rangle \\
&+ \left(\cancel{\sum_{i=1}^{t-1} a_i \widetilde{M}(x_i)} + \sum_{i=1}^{t-1} \frac{a_i}{\gamma}\langle \widetilde{\nabla} M(x_i), z_{t-1} - x_i\rangle + \frac{\|z_{t-1} - x_0\|_2^2}{2}\right) - A_{t-1}\delta - \frac{A_{t-1} D}{\gamma}\sqrt{8L\delta} \\
&\stackrel{②}{\leq} \left(-\frac{A_t}{8L}\|\widetilde{\nabla} M(x_t)\|_2^2 + A_t\delta\right) + \left(\frac{a_t}{\gamma}\langle \widetilde{\nabla} M(x_t), z_{t-1} - z_t\rangle + \frac{a_t}{\gamma}\left(\sqrt{8LD^2\delta} + (9 + \frac{5a_t}{A_{t-1}})\delta\right)\mathbb{I}(t=1)\right) \\
&\quad - \frac{1}{2}\|z_t - z_{t-1}\|_2^2 + a_t\delta + \frac{a_t D}{\gamma}\sqrt{8L\delta}) \\
&\stackrel{③}{\leq} \underbrace{\left(\frac{a_t^2}{2\gamma^2} - \frac{A_t}{8L}\right)}_{\leq 0}\|\widetilde{\nabla} M(x_t)\|_2^2 + \frac{a_t}{\gamma}(\sqrt{8LD^2\delta} + (9 + \frac{5a_t}{\gamma A_{t-1}})\delta)\mathbb{I}(t=1) + A_{t+1}\delta + \frac{a_t D}{\gamma}\sqrt{8L\delta} \\
&\leq \frac{a_t}{\gamma}(\sqrt{8LD^2\delta} + (9 + \frac{5a_t}{\gamma A_{t-1}})\delta)\mathbb{I}(t=1) + A_{t+1}\delta + \frac{a_t D}{\gamma}\sqrt{8L\delta} \stackrel{\text{def}}{=} E_t.
\end{aligned}$$

where ① holds by definition of G_t. In ②, for the first summand we apply Lemma 3, and for the second summand we apply the guarantee of the binary line search, cf. Lemma 4 with $c = A_{t-1}\gamma/a_t$ after multiplying it by a_t/γ, and we merged the resulting term with the last one in the second line after ①. Also, in ②, we canceled some errors depending on δ and we also bounded the remaining terms in big parentheses after canceling terms, by noting that they consist of the subtraction $-\ell_t(z_t) + \ell_t(z_{t-1})$ of a 1-strongly convex function ℓ_t, with minimizer z_{t-1} by its definition, and so we upper bound it by $-\frac{1}{2}\|z_t - z_{t-1}\|_2^2$ in ②. Now, in ③, we used Cauchy-Schwartz and Young's inequality $\langle a, b\rangle \leq \frac{1}{2}(a^2 + b^2)$ to cancel the terms depending on $z_{t-1} - z_t$ and note that the remaining terms proportional to $\|\widetilde{\nabla} M(x_t)\|_2^2$ are non-positive since the choice $a_t = \gamma^2 t/(8L)$ makes $A_t/(8L) = \gamma^2 t(t+1)/(128L^2) \geq \gamma^2 t^2/(128L^2) = a_t^2/(2\gamma^2)$.

Thus, it only remains to bound the right hand side of (10) with the value of E_t above:

$$\begin{aligned}
f(\hat{y}_T) - f(x^*) \leq \frac{1}{A_T}\left(\frac{1}{2}\|x^* - x_0\|_2^2 + \sum_{i=1}^{T} E_i\right) &\stackrel{①}{\leq} \frac{8L\|x^* - x_0\|_2^2}{\gamma^2 T^2} + \frac{8L}{\gamma^2 T^2}\sum_{i=1}^{T} E_i \\
&\stackrel{②}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon T}{16LD^2}\left[T^2(2\sqrt{8LD^2\delta} + (19 + 4T^2)\delta)\right] \\
&\stackrel{③}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\end{aligned}$$

where in ② we used $T = \frac{4}{\gamma}\sqrt{\frac{LD^2}{\varepsilon}}$, and we bounded $a_t \leq \gamma^2 T^2/(8L)$, $A_{t+1} \leq 4\gamma^2 T^2/(16L)$, $\gamma \leq 1$ and $a_t \leq 2A_{t-1}$ for $t \geq 2$. In ③, we used the value $\delta = LD^2/(10T^6)$.

Finally the total number of first-order oracle queries corresponds to T times the complexity of the binary search, which is $O(\frac{LD^2}{\varepsilon})$ by Lemma 4, so the total complexity is $\widetilde{O}(\frac{1}{\gamma}\sqrt{\frac{LD^2}{\varepsilon}})$ as desired. □

Proof. (Lemma 3) We have the following, by adding and subtracting $\frac{L}{2}\|y_t - x_t\|_2^2$:

$$\widetilde{M}(y_t) - \delta \le M(y_t) \le f(y_t) \pm \frac{L}{2}\|y_t - x_t\|_2^2 = \widetilde{M}(x_t) - \frac{L}{2}\|y_t - x_t\|_2^2 = \widetilde{M}(x_t) - \frac{1}{8L}\|\widetilde{\nabla} M(x_t)\|_2^2.$$

□

B.1 BINARY SEARCH

Proof. (Lemma 4) Denote $v_\alpha \stackrel{\text{def}}{=} \alpha y_{t-1} + (1-\alpha) z_{t-1}$. For f and $\lambda = 1/(2L)$, define the following function based on the Moreau envelope: $g(\alpha) = M(v_\alpha)$, for some $t \ge 1$. We have that $g'(\alpha) = \langle \nabla M(v_\alpha), y_{t-1} - z_{t-1}\rangle = \frac{1}{\alpha}\langle \nabla M(v_\alpha), v_\alpha - z_{t-1}\rangle$ is $\hat{L}$-Lipschitz, with $\hat{L} \stackrel{\text{def}}{=} 2L\|y_{t-1} - z_{t-1}\|_2^2$. Indeed:

$$|g'(\alpha_1) - g'(\alpha_2)| \le \|\nabla M(v_{\alpha_1}) - \nabla M(v_{\alpha_2})\|_2 \|y_{t-1} - z_{t-1}\|_2 \le 2L\|y_{t-1} - z_{t-1}\|_2^2,$$

where in the last inequality we used the $2L$-smoothness of M, cf. Lemma 1. We will use $h(\alpha) \stackrel{\text{def}}{=} \widetilde{M}(v_\alpha) \stackrel{\text{def}}{=} f(w_\alpha) + \frac{1}{2\lambda}\|w_\alpha - v_\alpha\|_2^2$, for $w_\alpha \stackrel{\text{def}}{\in} \arg\min_w^\delta\{f(w) + \frac{1}{2\lambda}\|w - v_\alpha\|_2^2\}$, and $\hat{h}(\alpha) \stackrel{\text{def}}{=} \langle \widetilde{\nabla} M(v_\alpha), y_{t-1} - z_{t-1}\rangle \stackrel{\text{def}}{=} \langle \frac{1}{\lambda}(v_\alpha - w_\alpha), y_{t-1} - z_{t-1}\rangle$ in order to implement the oracles in Algorithm 2. So the number δ-approximations of a $\text{prox}_{\lambda f + \mathbf{1}_\mathcal{X}}$ in the lemma statement equals the number of times that we require access to $h(\alpha), \hat{h}(\alpha)$. By their definitions, we have:

$$g(\alpha) \le h(\alpha) \le g(\alpha) + \delta_1 \text{ and } |\hat{h}(\alpha) - g'(\alpha)| \le \delta_2 \tag{11}$$

where $\delta_1 = \delta$, cf. (7), and $\delta_2 = \sqrt{8L\delta}\|y_{t-1} - z_{t-1}\|_2 \le \sqrt{8L\delta}D$ by using Cauchy-Schwartz on the expression resulting from substituting $g'(\alpha)$, $\hat{h}(\alpha)$ by their definitions, and applying (8), and $y_{t-1}, z_{t-1} \in \mathcal{X}$, $\text{diam}(\mathcal{X}) = D$.

We will show that we can do a binary search on g up to certain accuracy by only having access to the adversarially noisy function values and derivatives h and $\hat{h}$.

The stopping test is

$$\alpha\hat{h}(\alpha) \le c\big(h(1) - h(\alpha)\big) + \tilde{\varepsilon}. \tag{12}$$

for

$$\tilde{\varepsilon} \stackrel{\text{def}}{=} \delta_2 + (9 + 5c)\delta_1, \tag{13}$$

where the output x_t will be $v_\alpha = \alpha y_{t-1} + (1-\alpha)z_{t-1}$ for the final α that we compute. Note that $\alpha\hat{h}(\alpha) = \langle \widetilde{\nabla} M(x_t), \alpha(y_{t-1} - z_{t-1})\rangle = \langle \widetilde{\nabla} M(x_t), v_\alpha - z_{t-1}\rangle$.

Define as well the $\tilde{\varepsilon}_g \stackrel{\text{def}}{=} \tilde{\varepsilon} - c\delta_1 - \delta_2 = (9 + 4c)\delta_1$. If for some α we have

$$\alpha g'(\alpha) \;\le\; c\big(g(1) - g(\alpha)\big) \;+\; \tilde{\varepsilon}_g, \tag{14}$$

then (12) holds at α. Indeed, by (11):

$$\alpha(\hat{h}(\alpha) - \delta_2) \le \alpha g'(\alpha) \le c(g(1) - g(\alpha)) + \tilde{\varepsilon}_g \le c(g(1) - g(\alpha)) + \tilde{\varepsilon}_g + c\delta_1.$$

We first establish some invariants of our algorithm. If the check in Line 1 of Algorithm 2 succeeds, that is $\hat{h}(1) \le \tilde{\varepsilon}$, then (12) directly holds for $\alpha = 1$. The same is true if the second check succeeds, i.e., Line 2 in Algorithm 2. This is because then ① below holds, for $\alpha = 0$:

$$0 \cdot \hat{h}(0) = 0 \overset{①}{\le} c(h(1) - h(0)) + c\delta_1 \overset{②}{\le} c(h(1) - h(0)) + \tilde{\varepsilon}$$

and ② holds by (13). Note that this second check succeeds for the first iteration of Algorithm 2. So from now on, we can assume

$$\hat{h}(1) > \tilde{\varepsilon} \text{ and } h(0) > h(1) + \delta_1.$$

We have the following invariants, for all $k \ge 0$, throughout the execution of Algorithm 2:

(1) $h(1) \ge h(b_k)$.

(2) $h(a_k) \geq h(1) - \delta_1 \geq h(b_k) - \delta_1$.

(3) $\hat{h}(b_k) \geq \tilde{\varepsilon}$.

Indeed, (1) holds since it starts being $b_0 = 1$ and after any update in Line 8, it is $h(b_k) < h(1) - \delta_1 \leq h(1)$. The first inequality of (2) holds at $k = 0$ by the initial properties shown above and by any update made in Line 8, whereas the second inequality uses (1). For (3), we have that this property holds at $k = 0$ by the initial properties shown above. For any time $k \geq 0$ for which $b_{k+1} \neq b_k$, we have that $b_{k+1} \in (0, 1]$ equals an α that did not trigger the termination of the while loop, and thus ① below holds:

$$\hat{h}(b_{k+1}) \geq b_{k+1}\hat{h}(b_{k+1}) \overset{①}{>} c(h(1) - h(b_{k+1})) + \tilde{\varepsilon} \overset{②}{\geq} \tilde{\varepsilon}.$$

where ② holds by the second inequality of (1).

We will now show by induction that throughout the execution of the algorithm, if we did not stop, the interval $[a_k, b_k]$ contains an interval of length

$$\Delta \overset{\text{def}}{=} \frac{(\tilde{\varepsilon}_g - \delta_1)}{(1 + c/2)\hat{L}} = \frac{8\delta_1}{\hat{L}} > 0. \tag{15}$$

such that (14) holds. For this, it is enough to show that there is a point $\alpha^* \in [a_k, b_k]$ such that $g(\alpha^*) = 0$ and $g(\alpha) \leq g(b_k)$. This is true since by $g(\alpha) \leq g(b_k) \leq h(b_k) \leq h(1) \leq g(1) + \delta_1$, and $\hat{L}$-smoothness we have that for any $t \in [0, 1]$ with $|t - \alpha^*| \leq \Delta/2$:

$$g(1) - g(t) \geq g(\alpha^*) - \delta_1 - g(t) \geq -\frac{\hat{L}}{2}(\alpha - t)^2 - \delta_1 \geq -\frac{\hat{L}}{2}|\alpha - t| - \delta_1 \geq -\frac{\hat{L}\Delta}{4} - \delta_1,$$

$$tg'(t) \leq |g'(t)| \leq \hat{L}|t - \alpha^\star| \leq \hat{L}\frac{\Delta}{2},$$

and thus

$$tg'(t) \leq c\big(g(1) - g(t)\big) + \frac{\hat{L}\Delta}{2}(1 + \frac{c}{2}) + \delta_1 \leq c\big(g(1) - g(t)\big) + \tilde{\varepsilon}_g.$$

So we only need to prove the existence of such an α^* inductively. Note that as long as we do not stop, the check of (12) failed at the endpoints of $[a_k, b_k]$ so the good interval of length Δ would be completely in $[a_k, b_k]$.

We analyze two cases. By the invariant (2), these cases cover all possibilities.

Case 1, $h(a_k) > h(b_k) + \delta_1$ Recall that we have $\hat{h}(b_k) > \tilde{\varepsilon}$ by (3). Then

$$g(a_k) \geq h(a_k) - \delta_1 > h(b_k) \geq g(b_k), \qquad g'(b_k) \geq \hat{h}(b_k) - \delta_2 > \tilde{\varepsilon} - \delta_2 > 0.$$

Then, by continuity, there exists $\alpha^* \in (a_k, b_k)$ with $g'(\alpha^*) = 0$ and $g(\alpha^*) \leq g(b_k)$.

Case 2. $h(b_k) - \delta_1 \leq h(a_k) \leq h(b_k) + \delta_1$ Using (3) again, $g'(b_k) \geq \hat{h}(b_k) - \delta_2 > \tilde{\varepsilon} - \delta_2 \geq \tilde{\varepsilon}_g > \Delta\hat{L}$. Since we did not stop in the previous iteration and we halved the interval that inductively was of size greater than Δ, we now have $b_k - a_k > \Delta/2$. Define $\beta \overset{\text{def}}{=} b_k - \frac{\Delta}{2} \in (a_k, b_k]$ and note that by smoothness and (15):

$$g'(\beta) \geq g'(b_k) - \frac{\Delta\hat{L}}{2} > \frac{\Delta\hat{L}}{2} > 0.$$

Applying smoothness again and using $g'(b_k) \geq \Delta\hat{L} > 0$:

$$g(\beta) \leq g(b_k) - g'(b_k)\frac{\Delta}{2} + \frac{\hat{L}}{2}\frac{\Delta^2}{4} \leq g(b_k) - \frac{3\Delta\hat{L}}{8} \overset{①}{\leq} g(b_k) - 3\delta_1.$$

where ① holds by definition of Δ, cf. (15). Finally, using this last result, since $g \leq h \leq g + \delta_1$ and $h(a_k) \geq h(b_k) - \delta_1$, we have:

$$g(\beta) \leq g(b_k) - 3\delta_1 \leq h(b_k) - 2\delta_1 \leq h(a_k) - \delta_1 \leq g(a_k).$$

We are now in the same situation as before. Since $a_k < \beta$, $g'(\beta) > 0$ and $g(a_k) \geq g(\beta)$, by continuity there must exist a point $\alpha^* \in (a_k, b_k)$ such that $g'(\alpha^*) = 0$ and $g(\alpha^*) \leq g(\beta) \leq g(b_k)$.

To conclude, the width of the interval $[a_k, b_k]$ was halved at each iteration, and the algorithm must stop before the length of this interval is $\leq \Delta$, so we successfully stop after at most

$$\left\lceil \log_2\left(\frac{1}{\Delta}\right)\right\rceil = \left\lceil \log_2\left(\frac{8\hat{L}}{\delta_1}\right)\right\rceil \leq \left\lceil \log_2\left(\frac{8LD^2}{\delta_1}\right)\right\rceil$$

iterations of the while loop. At each iteration we query h, $\hat{h}$ a constant number of times and before the loop we also do a constant number of queries. The total number of queries is $O(\log(\frac{LD^2}{\delta}))$. □

C PROOFS FOR UNACCELERATED METHODS

C.1 Steepest Descent

Proof. (Lemma 6) Define the quadratic $Q_x : \mathcal{X} \to \mathbb{R}$ as

$$Q_x(z) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), z - x\rangle + \frac{1}{2\eta}\|z - x\|_2^2. \tag{16}$$

For any x, define $x^+ \stackrel{\text{def}}{=} \arg\min_{z\in\mathcal{X}} Q_x(z)$ as the resulting point from taking a projected gradient descent step from x, cf. (2). Note that Q_x is $(1/\eta)$-strongly convex. Thus, for all $y \in \mathcal{X}$, we have

$$\begin{aligned} f(x)+\langle \nabla f(x), x^+ - x\rangle + \frac{1}{2\eta}\|x^+ - x\|_2^2 + \frac{1}{2\eta}\|x^+ - y\|_2^2 &= Q_x(x^+) + \frac{1}{2\eta}\|x^+ - y\|_2^2 \\ \leq Q_x(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\eta}\|y - x\|_2^2. \end{aligned}$$

Or equivalently,

$$\begin{aligned} \langle \nabla f(x), x^+ - y\rangle &\leq \frac{1}{2\eta}\|y - x\|_2^2 - \frac{1}{2\eta}\|x^+ - x\|_2^2 - \frac{1}{2\eta}\|x^+ - y\|_2^2 \\ &= \frac{1}{\eta}\langle x - x^+, x^+ - y\rangle = \langle g_{\mathcal{X}}(x), x^+ - y\rangle. \end{aligned}$$

□

Proof. (Lemma 7) Using the smoothness inequality, we get the result, after applying Lemma 6 in ①:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\overset{①}{\leq} \langle g_X(x_t), x_{t+1} - x_t\rangle + \frac{1}{2L}\|g_{\mathcal{X}}(x_t)\|_2^2 = -\frac{1}{2L}\|g_{\mathcal{X}}(x_t)\|_2^2. \end{aligned}$$

□

Proof. (Theorem 8) First, we argue that all the iterates stay in the set $\mathcal{Y} \stackrel{\text{def}}{=} \{x \in \mathcal{X} \mid f(x) \leq f(x_0)\}$ of diameter D, which contains x^* by definition. By induction, assuming x_t is in the set, which holds by definition for the first iterate x_0, we show that x_{t+1} is also in the set. The update rule implies:

$$f(x_{t+1}) = \min_{x\in\mathcal{X}}\left\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{L}{2}\|x - x_t\|_2^2\right\} \overset{①}{\leq} f(x_t) \overset{②}{\leq} f(x_0),$$

where ① is obtained by setting $x = x_t$ and ② is by the induction hypothesis. Thus, the property is proven.

Now, for $t \geq 0$, let a_t be weights to be determined and let $A_t = \sum_{i=0}^t = a_t$. Consequently, define $A_{-1} \stackrel{\text{def}}{=} 0$ and denote $g_t \stackrel{\text{def}}{=} g_{\mathcal{X}}(x_t)$. For a minimizer $x^* \in \arg\min_{x\in\mathcal{X}} f(x)$, we define the following lower bound L_t on $f(x^*)$:

$$
\begin{aligned}
A_t f(x^*) &\overset{①}{\geq} \sum_{i=0}^{t} a_i f(x_{i+1}) + \sum_{i=0}^{t} \frac{a_i}{\gamma}\Big(\langle \nabla f(x_i), x^* - x_{i+1}\rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x^* - x_{i+1}\rangle\Big) \\
&\overset{②}{\geq} \sum_{i=0}^{t} a_i f(x_{i+1}) + \sum_{i=0}^{t} \frac{a_i}{\gamma}\Big(\langle g_i, x^* - x_{i+1}\rangle - L\|x_{i+1} - x_i\|_2 \cdot \|x^* - x_{i+1}\|_2\Big) \\
&\overset{③}{\geq} \sum_{i=0}^{t} a_i f(x_{i+1}) - \sum_{i=0}^{t} \frac{2a_i}{\gamma}\|g_i\|_2 D \\
&\overset{\text{def}}{=} A_t L_t,
\end{aligned}
$$

where in ① we applied a combination of the quasar-convexity inequalities given by points visited by the algorithm, and added and subtracted some terms so that in ② we use Lemma 6 on one term and Cauchy-Schwarz and gradient Lipschitzness of f on the other. Finally in ③ we applied Cauchy-Schwarz as well and used the bounds $\|x^* - x_{i+1}\|_2 \leq D$, which is due to the iterates staying in the set $\mathcal{Y}$.

Define the gap $G_t = f(x_{t+1}) - L_t$. Our aim is to show $A_t G_t - A_{t-1} G_{t-1} \leq E_t$, for some value E_t for all $t \geq 0$, so we can conclude $f(x_t) - f(x^*) \leq G_{t-1} \leq \frac{A_{t-2} G_{t-2} + E_{t-1}}{A_{t-1}} \leq \cdots \leq \frac{1}{A_{t-1}} \sum_{i=0}^{t-1} E_i$. Let $D_t \overset{\text{def}}{=} \max_{i \in \{0,\ldots,t\}} \|x^* - x_t\|_2 \leq D$. First, for $t \geq 1$, we have:

$$
\begin{aligned}
A_t G_t - A_{t-1} G_{t-1} &= A_{t-1}(f(x_{t+1}) - f(x_t)) + a_t f(x_{t+1}) \\
&\quad - \left(a_t f(x_{t+1}) + \sum_{i=0}^{t-1} a_i f(x_{i+1}) - \frac{2a_t}{\gamma}\|g_t\|_2 D + \sum_{i=0}^{t-1} \frac{2a_i}{\gamma}\|g_i\|_2 D\right) \\
&\quad + \left(\sum_{i=0}^{t-1} a_i f(x_{i+1}) + \sum_{i=0}^{t-1} \frac{2a_i}{\gamma}\|g_i\|_2 D\right) \\
&\overset{①}{\leq} -\frac{A_{t-1}}{2L}\|g_t\|_2^2 + \frac{2a_t}{\gamma}\|g_t\|_2 D \overset{②}{\leq} -\frac{A_{t-1}}{2L}\|g_t\|_2^2 + \frac{A_{t-1}}{2L}\|g_t\|_2^2 + \frac{2La_t^2}{\gamma^2 A_{t-1}} D^2 \overset{\text{def}}{=} E_t.
\end{aligned}
$$

where in ① we used Lemma 7 and in ② we used Young's inequality and canceled some terms.

For $t = 0$, the inequalities above hold up to before ②. Taking into account that $A_{-1} = 0$, we have $A_0 G_0 = A_0 G_0 - A_{-1} G_{-1} = E_0 \leq 2a_0 L D^2/\gamma$, where we have $\|g_t\|_2 \leq LD$ due to the fact that x_t, x_{t+1} are in the level set of x_0.

Finally, choosing $a_t = 2t + 2$ so that $A_t = (t+1)(t+2)$, we have, for all $t \geq 1$:

$$
\begin{aligned}
f(x_t) - f(x^*) \leq G_{t-1} &\leq \frac{1}{A_{t-1}} \sum_{i=0}^{t-1} E_i = \frac{1}{A_{t-1}}\left(\sum_{i=1}^{t-1} \frac{2LD^2 a_i^2}{\gamma^2 A_{i-1}} + \frac{4LD^2}{\gamma}\right) \\
&= \frac{1}{\gamma^2 t(t+1)}\left(\sum_{i=1}^{t-1} \frac{16LD^2(i+1)}{2i} + 4\gamma LD^2\right) \leq \frac{(16 + 4\gamma/t)LD^2}{\gamma^2(t+1)} = O\left(\frac{LD^2}{\gamma^2 t}\right).
\end{aligned}
$$

The numeric constant 20 in the theorem statement comes from bounding the above using $\gamma \in (0, 1]$ and $t \geq 1$. □

C.2 Frank-Wolfe

Proof. (Theorem 9) We note that actually, this proof works for any norm as long as L-smoothness and the diameter D are taken with respect to that norm. Let $a_t > 0$ to be determined later and define $A_t = A_{t-1} + a_t = \sum_{i=0}^{t} a_i$. Let $v_t \overset{\text{def}}{\in} \arg\min_{v \in \mathcal{X}} \langle \nabla f(x_t), v\rangle$ and let $x_{t+1} \overset{\text{def}}{=} \frac{A_{t-1}}{A_{t-1} + a_t/\gamma} x_t + \frac{a_t/\gamma}{A_{t-1} + a_t/\gamma} v_t$ be defined as a convex combination of x_t and v_t. Note $A_{-1} = 0$ and $A_0 = a_0$, so $x_1 = v_0$. We define the following lower bound on $f(x^*)$

Algorithm 3 Frank-Wolfe algorithm

Input: Function f that is L-smooth and γ-quasar convex in a compact convex set $\mathcal{X}$. Initial point $x_0 \in \mathcal{X}$.

$A_{-1} \leftarrow 0$

for $t \leftarrow 0$ **to** $T-1$ **do**

  $a_t \leftarrow 2t+2$; $A_t \leftarrow A_{t-1} + a_t = \sum_{i=0}^{t} a_i = (t+1)(t+2)$

  $v_t \in \arg\min_{v \in \mathcal{X}} \{\langle \nabla f(x_t), v \rangle\}$

  $x_{t+1} \leftarrow \frac{A_{t-1}}{A_t} x_t + \frac{a_t}{A_t} v_t = \frac{1}{A_t} \sum_{i=0}^{t} a_i v_i$ $\diamond = \frac{t}{t+2} x_t + \frac{2}{t+2} v_t$

end for

return x_T.

for a minimizer $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$:

$$
\begin{aligned}
A_t f(x^*) &\overset{①}{\geq} \sum_{i=0}^{t} a_i f(x_{i+1}) + \sum_{i=0}^{t} \frac{a_i}{\gamma} \langle \nabla f(x_{i+1}), x^* - x_{i+1} \rangle \\
&\overset{②}{\geq} \sum_{i=0}^{t} a_i f(x_{i+1}) + \sum_{i=0}^{t} \frac{a_i}{\gamma} \Big(\langle \nabla f(x_i), v_{i+1} - x_{i+1} \rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), v_{i+1} - x_{i+1} \rangle \Big) \\
&\overset{③}{\geq} \sum_{i=0}^{t} a_i f(x_{i+1}) + \sum_{i=0}^{t} \frac{a_i}{\gamma} \Big(\langle \nabla f(x_i), v_{i+1} - x_{i+1} \rangle - L \|x_{i+1} - x_i\| \cdot D \Big) \overset{\text{def}}{=} A_t L_t,
\end{aligned}
$$

where we applied quasar-convexity of f in ①, the optimality of v_{i+1} in ②, along with adding and subtracting some terms. And finally, we used Cauchy-Schwarz in ③ along with using gradient Lipschitzness of f, and also bounding $\|v_{i+1} - x_{i+1}\|$ by $\operatorname{diam}(\mathcal{X}) = D$.

Now we define the gap $G_t \overset{\text{def}}{=} f(x_{t+1}) - L_t$. If we show $A_t G_t - A_{t-1} G_{t-1} \leq E_t$, for some number E_t for all $t \geq 0$ (note $A_{-1} = 0$), then we will have $f(x_t) - f(x^*) \leq G_{t-1} \leq \frac{1}{A_{t-1}} \sum_{i=0}^{t-1} E_i$. Our aim is thus to have small E_t and large A_t. We obtain the following, for $t \geq 0$:

$$
\begin{aligned}
A_t G_t - A_{t-1} G_{t-1} &= A_{t-1}(f(x_{t+1}) - f(x_t)) + a_t f(x_{t+1}) \\
&\quad - \Big(\sum_{i=0}^{t} a_i f(x_{i+1}) + \frac{a_t}{\gamma} (\langle \nabla f(x_t), v_{t+1} - x_{t+1} \rangle - LD \|x_{t+1} - x_t\|) \\
&\quad + \sum_{i=0}^{t-1} \frac{a_i}{\gamma} (\langle \nabla f(x_i), v_{i+1} - x_{i+1} \rangle - LD \|x_{i+1} - x_i\|) \Big) \\
&\quad + \left(\sum_{i=0}^{t-1} a_i f(x_{i+1}) + \sum_{i=0}^{t-1} \frac{a_i}{\gamma} (\langle \nabla f(x_i), v_{i+1} - x_{i+1} \rangle - LD \|x_{i+1} - x_i\|) \right) \\
&\overset{①}{\leq} \langle \nabla f(x_t), A_{t-1}(x_{t+1} - x_t) - \frac{a_t}{\gamma}(v_{t+1} - x_{t+1}) \rangle + \frac{A_{t-1} L}{2} \|x_{t+1} - x_t\|^2 + \frac{a_t}{\gamma} LD \|x_{t+1} - x_t\| \\
&\overset{②}{\leq} \frac{a_t}{\gamma} \langle \nabla f(x_t), v_t - v_{t+1} \rangle + \frac{a_t^2 A_{t-1} LD^2}{2\gamma^2 (A_{t-1} + a_t/\gamma)^2} + \frac{a_t^2 LD^2}{\gamma^2 (A_{t-1} + a_t/\gamma)} \\
&\overset{③}{\leq} \frac{3LD^2 a_t^2}{2A_t \gamma^2} \overset{\text{def}}{=} E_t.
\end{aligned}
$$

In ①, we applied smoothness on the first term, and canceled and grouped some terms. In ② we used that the definition of x_{t+1} implies $A_{t-1}(x_{t+1} - x_t) = \frac{a_t}{\gamma}(v_t - x_{t+1})$ in order to obtain a term depending on $\nabla f(x_t)$ that is nonpositive by the optimality condition of the problem defining v_t. We also use that the definition of x_{t+1} also implies $x_{t+1} - x_t = \frac{a_t/\gamma}{A_{t-1} + a_t/\gamma}(v_t - x_t)$, so we use this to bound the other two distance terms and then bound $\|v_t - x_t\| \leq D$. In ③ we dropped the gradient term and we also used $\gamma \leq 1$ for the term $A_{t-1} + a_t/\gamma \geq A_t$ in two denominators. Note that all of these computations are valid for $t = 0$.

Finally, choosing $a_t = 2t + 2$, we have $A_t = \sum_{i=0}^{t} a_i = (t+1)(t+2)$, so for all $t \geq 1$:

$$f(x_t) - f(x^*) \leq G_{t-1} \leq \frac{1}{A_{t-1}} \sum_{i=0}^{t-1} E_i = \frac{1}{A_{t-1}} \sum_{i=0}^{t-1} \frac{3LD^2 a_i^2}{2A_i \gamma^2} = \frac{1}{t(t+1)} \sum_{i=0}^{t-1} \frac{6LD^2(i+1)}{(i+2)\gamma^2} < \frac{6LD^2}{(t+1)\gamma^2}.$$

□