BROXIMAL GRADIENT DESCENT: A PROJECTION-FREE SISTER OF PROJECTED GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Broximal Gradient Descent (BroxGD), a projection-free sister method to projected gradient descent for constrained optimization. Its forward–backward construction replaces the proximal backward operation by the broximal operation of Gruntkowska et al. (2025). Instead of projecting, each step minimizes a linear function over the intersection of the constraint set $\mathcal{X}$ and a ball $\mathcal{B}(x_k, t_k)$ centered at the current iterate x_k, of suitable radius $t_k > 0$:

$$x_{k+1} \in \operatorname{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle.$$

We develop a comprehensive convergence theory spanning a wide range of optimization regimes and radius rules. We expect BroxGD to find many applications and inspire numerous extensions, much like projected gradient descent. Our contribution is theoretical; potential applications and toy experiments illustrate the method and suggest directions for future work.

1 INTRODUCTION

Constrained optimization problems play a fundamental role in optimization theory and machine learning (ML) (Boyd & Vandenberghe, 2004). In this work, we revisit the classical formulation

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

where the set $\mathcal{X} \subseteq \mathbb{R}^d$ represents constraints, and $f : \mathbb{R}^d \to \mathbb{R}$ represents a loss/cost/objective function. Hence, we seek to minimize the loss f subject to the constraint $\mathcal{X}$. We denote the set of all minimizers by $\mathcal{X}_\star$. Such problems arise throughout ML (Pokutta et al., 2020; Sangalli et al., 2021), statistics (Hathaway, 1985; Hall & Presnell, 1999; Behr et al., 2013), and signal processing (Mattingley & Boyd, 2010; Zhong & Shen, 2026), where constraints are used to encode prior structure, enforce regularization, or impose physical feasibility. Constrained formulations play an increasingly important role in modern ML systems. As learning algorithms are deployed in safety-critical domains—including finance, healthcare, and autonomous systems—there is a growing demand for models that are safe, robust, interpretable, and fair (Donini et al., 2018; Yang et al., 2022). Constrained optimization offers a framework for encoding such requirements directly into the training objective (Cotter et al., 2019; Dai et al., 2023). Naturally, the extremely rich optimization literature offers a multitude of algorithms for handling such problems. A useful way of mapping out the landscape of these methods is to group them by the oracle used to access the constraint set $\mathcal{X}$. Most approaches fall into one of these three categories[1]:

(i) Projection oracle. Methods in this class assume access to a projection operator of the form

$$\operatorname{Proj}_{\mathcal{X}}^{\phi}(x) \in \arg\min_{z \in \mathcal{X}} D_\phi(z, x),$$

where $D_\phi(z, x) := \phi(z) - \phi(x) - \langle \nabla \phi(x), z - x \rangle$ is the *Bregman divergence* induced by a strictly convex, differentiable function ϕ. This includes Projected Gradient Descent (ProjGD) (Bertsekas, 1999) when $\phi(x) = \frac{1}{2}\|x\|^2$, mirror descent (Nemirovski & Yudin, 1983) (and its special cases such as exponentiated gradient methods (Kivinen & Warmuth, 1997)), and dual averaging (Nesterov, 2009).

[1]These three oracle models are representation-free: they do not assume an explicit parametrization of $\mathcal{X}$, but only access it through queries revealing geometric information about the set. When additional structure of $\mathcal{X}$ is available, it can be exploited to design more specialized algorithms–see Appendix B.4, Appendix B.5, Appendix B.6 and Boyd & Vandenberghe (2004).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(ii) Linear minimization oracle (LMO). These methods, which include the Frank–Wolfe method (Frank & Wolfe, 1956; Levitin & Polyak, 1966) and its variants, access the constraint set only through linear optimization queries of the form

$$\operatorname{argmin}_{z\in\mathcal{X}}\langle g, z\rangle.$$

Such oracles are attractive in settings where projections may be expensive or intractable.

(iii) Separation oracle. Such oracles provide, for a given query point x, either a certificate that $x \in \mathcal{X}$ or a hyperplane separating x and $\mathcal{X}$. Classical methods based on this oracle include the ellipsoid method (Grötschel & Schrijver, 1993), cutting-plane schemes (Gilmore & Gomory, 1961), and bundle-type epigraph algorithms (Lemaréchal et al., 1995). Membership oracles (Grötschel & Schrijver, 1993) are sometimes considered in this context as a weaker primitive, but are typically converted into separation access for algorithmic purposes.

1.1 From proximal to broximal steps

Our starting point is the Ball-Proximal ("Broximal") Method (BPM) of Gruntkowska et al. (2025). For a proper closed function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, a center $x \in \operatorname{dom} h$, and a radius $t > 0$, its broximal operator is the mapping

$$\operatorname{brox}_h^t(x) \coloneqq \operatorname{argmin}_{\|z-x\|\le t} h(z). \tag{2}$$

BPM repeatedly applies this operator to the objective itself, replacing the proximal quadratic penalty by a ball constraint. Remarkably, for a proper closed convex objective with a minimizer, exact BPM with any fixed positive radius enjoys linear convergence of objective values and reaches a minimizer in finitely many steps, without smoothness or strong convexity (Gruntkowska et al., 2025, Theorem 8.1 and Corollary 8.2).

These striking guarantees concern exact minimization over each ball; they motivate the design of a new forward–backward (FB) algorithm for solving (1), analogous to projected gradient descent, but *somehow* using the broximal operator instead of the projection (proximal) operator (Moreau, 1965).

In order to design such a method, we first move the constraint to the objective, which is a standard procedure in convex optimization. In particular, we write the constrained optimization problem (1) in the form

$$\min_{x\in\mathbb{R}^d} F(x), \qquad F \coloneqq f + \iota_{\mathcal{X}},$$

where $\iota_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$. We then linearize f, and form the model

$$m_k(z) \coloneqq \langle \nabla f(x_k), z - x_k\rangle + \iota_{\mathcal{X}}(z).$$

Adding a quadratic movement penalty to the model and minimizing leads to projected gradient descent:

$$x_{k+1}^{\mathsf{ProjGD}} = \operatorname{argmin}_{z\in\mathbb{R}^d}\left\{m_k(z) + \tfrac{\|z-x_k\|^2}{2\gamma_k}\right\}, \tag{3}$$

where $\{\gamma_k\}_{k\ge 0}$ is an appropriately chosen sequence of positive stepsizes.

1.2 Broximal gradient descent

Our broximal viewpoint instead suggests replacing this penalty by a *hard movement constraint*[2]. The resulting method minimizes the same model, but over a ball centered at the current iterate:

$$\boxed{x_{k+1} \in \operatorname*{argmin}_{\|z-x_k\|\le t_k} m_k(z),} \tag{4}$$

where $\{t_k\}_{k\ge 0}$ is an appropriately chosen sequence of positive radii. This is *Broximal Gradient Descent* (BroxGD)—the method we propose and analyze in this paper. Letting $\mathcal{B}(x,t) \coloneqq \{y \in \mathbb{R}^d : \|y - x\| \le t\}$, the BroxGD update step (4) can be equivalently written in the form

$$\boxed{x_{k+1} \in \operatorname*{argmin}_{z\in\mathcal{X}\cap\mathcal{B}(x_k,t_k)} \langle \nabla f(x_k), z\rangle,} \tag{5}$$

[2]Appendix A develops this viewpoint in detail.

Algorithm 1 Broximal gradient descent (BroxGD)

1: **Input:** starting point $x_0 \in \mathcal{X}$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Compute the gradient $g_k \coloneqq \nabla f(x_k)$
4: Choose a radius $t_k \geq 0$
5: Compute the next iterate as the solution of the local linear minimization problem

$$x_{k+1} \in \operatorname{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle g_k, z \rangle$$

6: **end for**

which more clearly reveals its projection-free nature. Instead of projections onto the constraint $\mathcal{X}$, BroxGD relies on linear minimization over the intersection of $\mathcal{X}$ and a local ball ("trust-region"). Note that without constraints and when $\nabla f(x_k) \neq 0$, BroxGD reduces to normalized gradient descent with stepsize t_k, or equivalently, to gradient descent with stepsize $t_k / \|\nabla f(x_k)\|$. We formalize the method in Algorithm 1.

1.3 Related work

Here we give an overview of closely related literature, with a broader discussion in Appendix B.

Our method is a *local projection-free first-order method*. In that sense, it is naturally positioned between Projected Gradient Descent (ProjGD) (Rosen, 1960) and Frank–Wolfe (FW) (Frank & Wolfe, 1956; Levitin & Polyak, 1966). ProjGD performs the update

$$x_{k+1} = \operatorname{Proj}_{\mathcal{X}}\big(x_k - \gamma_k \nabla f(x_k)\big),$$

where $\operatorname{Proj}_{\mathcal{X}}(x) \coloneqq \arg\min_{z \in \mathcal{X}} \|x - z\|$, and thus relies on a *projection oracle*. Under standard assumptions, it matches the standard convergence guarantees of unaccelerated unconstrained Gradient Descent (GD), without dependence on global geometric properties of $\mathcal{X}$, such as its diameter $\operatorname{diam} \mathcal{X} \coloneqq \sup_{x,y \in \mathcal{X}} \|x - y\|$, and naturally applies to unbounded constraints. Its main drawback, however, is that projections can be computationally expensive or even intractable for many constraints.

By contrast, classical Frank–Wolfe computes

$$s_k \in \operatorname{argmin}_{z \in \mathcal{X}} \langle \nabla f(x_k), z \rangle$$

and then updates x_k by moving toward s_k, thus relying on a *global linear minimization oracle* (see Appendix B.3, the original paper of (Frank & Wolfe, 1956) and the modern overview by (Jaggi, 2013)). Thus, FW avoids projections altogether. Classical FW is typically analyzed for problems of the form (1) under the assumptions that $\mathcal{X}$ is nonempty, convex, and compact, that f is convex and differentiable on $\mathcal{X}$. A large body of work shows that, if f is additionally smooth, one obtains the classical sublinear rate $f(x_K) - f(x_\star) = \mathcal{O}(1/K)$, which is generally unimprovable (Hazan, 2008; Lan, 2013; Jaggi, 2013). This smoothness assumption is often expressed via the FW curvature constant C_{FW} (see (38)), rather than a global gradient Lipschitz constant (Jaggi, 2013). Stronger guarantees, such as linear convergence, generally require substantially stronger assumptions, e.g, that the solution lies in the interior of $\mathcal{X}$ (Guélat & Marcotte, 1986), or that f is strongly convex together with favorable geometry of the feasible set (most often polyhedral structure) (Levitin & Polyak, 1966). Such rates are typically established not for vanilla FW, but for enhanced variants such such as away-step, pairwise, or fully corrective methods. Moreover, there are examples where no rate improvement is possible even when f is strongly convex (Lan, 2013). Overall, the simplicity of FW-type methods comes with well-known limitations (not shared by ProjGD), summarized in Table 1. This highlights a fundamental distinction between projection- and LMO-based methods.

It is also useful to distinguish our method from other linearly convergent projection-free variants, such as away-step FW (Wolfe, 1970), pairwise FW, and fully corrective FW methods (Lacoste-Julien & Jaggi, 2015). Those methods still rely on the *global* linear oracle, but add away/correction mechanisms to overcome the slow boundary behavior of vanilla FW; their analyses typically rely on polyhedral geometric constants such as pyramidal width (Lacoste-Julien & Jaggi, 2015). Our method instead enforces progress by restricting the search to a *local trust region*. In this sense, it is closer in spirit to localized FW methods than to away-step or pairwise FW.

1.4 CLOSELY RELATED WORK

To the best of our knowledge, the method we call BroxGD as well as local linear subproblems of this form were first studied by Ferris & Zavriev (1996) in their successive linear programming work. Garber & Hazan (2016) subsequently developed a relaxed local linear optimization oracle (LLOO) with an implementation for polytopes. We derived BroxGD and our initial convergence results independently of these works, from the BPM viewpoint described in Section 1.1. We became aware of Garber and Hazan while finalizing the first version of this work after all our results were already obtained, and of Ferris and Zavriev while preparing the present revision. Garber and Hazan also appear to have been unaware of Ferris and Zavriev's work, which is not cited in their paper. We fully acknowledge these prior works here. Having said that, our results are complementary, much more comprehensive, and have a different focus. Appendix B.7 explains in detail how our results differ from those of Ferris and Zavriev and Garber and Hazan, including the algorithms, assumptions, and convergence guarantees. We moved the discussion to the appendix because we need substantial space for the comparison.

2 SUMMARY OF CONTRIBUTIONS

We develop a comprehensive convergence theory for BroxGD in both convex and nonconvex regimes, with and without smoothness, and with and without strong convexity. See Table 1 for a summary of most of our results. We now briefly describe some of these contributions, mostly cast in contrast to what is known about FW and/or ProjGD.

▷ **GD-type guarantees without global geometric dependence.** We show that our method matches the convergence rates of ProjGD across three classical regimes: $\mathcal{O}(1/K)$ objective error in the smooth convex case (Theorem 3.2), $\mathcal{O}(1/\sqrt{K})$ objective error under bounded gradients (Theorem 3.2), and linear convergence of iterates under smoothness and strong convexity (Theorem 3.4). These guarantees hold *without requiring the constraint $\mathcal{X}$ to be bounded* and *without any dependence on global geometric quantities* such as the diameter of $\mathcal{X}$ or FW–style curvature constants (see (38)). The classical compact-domain FW bound depends on $\operatorname{diam}\mathcal{X}$. Directly replacing gradients by subgradients in vanilla FW can fail for non-smooth functions (Nesterov, 2017, Example 1), whereas BroxGD naturally applies in this regime and guarantees $f(\bar{x}_K) - f(x_\star) \le G\|x_0 - x_\star\|/\sqrt{K}$ for the average iterate $\bar{x}_K := \frac{1}{K}\sum_{k=0}^{K-1} x_k$; see also Appendix D.2.2. Convergence of objective values does not itself ensure convergence of the iterate sequence; examples for FW appear in Bolte et al. (2022).

▷ **A new perspective on projection-free optimization.** BroxGD can be viewed as a *generalization of* GD rather than a variant of FW (see Appendix C.8). In particular, when $\mathcal{X} = \mathbb{R}^d$, BroxGD reduces to

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k), \tag{6}$$

which is GD with stepsize $\gamma_k = t_k/\|\nabla f(x_k)\|$. When $\mathcal{X}$ is an affine subspace of $\mathbb{R}^d$, i.e., $\mathcal{X} = a + \mathcal{V}$, where $a \in \mathbb{R}^d$ and $\mathcal{V} \subseteq \mathbb{R}^d$ is a linear subspace, it reduces to GD for minimizing the restriction of f to $\mathcal{X}$. This is another way of seeing why the method inherits GD-type convergence behavior.

▷ **Extensions.** The main paper develops deterministic convex guarantees: practical fixed-horizon radii under bounded gradients or smoothness, radius backtracking, and a sharp contraction benchmark under smoothness and strong convexity. The appendix studies gradient-difference residuals under standard and generalized smoothness (Appendix D.3), and smooth nonconvex stationarity and projected-PŁ objective convergence (Appendix D.4); these results appear in Table 1. The latter two use projection-dependent radii. The table also includes Polyak-radius objective bounds for non-differentiable convex objectives (Appendix D.2.2) and smooth convex objectives (Theorem D.9). Stochastic extensions (Appendix F) are presented separately and are not included in the table.

▷ **Analysis of the oracle.** Just like FW, BroxGD relies on a linear minimization oracle only, and does not require access to a projection oracle. However, it replaces the global oracle over $\mathcal{X}$ with a local one over $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$, which can be both less or more involved, depending on context. We discuss the computational implications of this oracle and show that it can often be implemented efficiently in Appendix C.2.

Table 1: Comparison of FW and BroxGD under different regimes. Rates are grouped by the guarantee: practical objective convergence (A), solution-dependent convex benchmarks (B), gradient-difference residuals (C), and nonconvex stationarity or projected-PL convergence (D). Shaded rate rows are presented in the main paper; the remaining rate results are in the appendix. Appendix D follows this order, including all three Polyak-radius guarantees in group B. FW entries are upper bounds on compact domains; compare the stated metrics, which can differ across columns. Notation: $x_\star \in \mathcal{X}_\star$, $\Delta_k \coloneqq f(x_k) - f(x_\star)$, $R_k \coloneqq \|x_k - x_\star\|$, $\bar{x}_K = K^{-1}\sum_{k=0}^{K-1} x_k$, $M_0 = \|\nabla f(x_\star)\| + LR_0$, $Z_k \coloneqq \|\nabla f(x_k) - \nabla f(x_\star)\|$, $c_k \coloneqq L_0 + L_1\|\nabla f(x_k)\|$, $c_\star \coloneqq L_0 + L_1\|\nabla f(x_\star)\|$, $\mathcal{D}_\mathcal{X} \coloneqq \operatorname{diam}\mathcal{X}$, $R \geq R_0$ with $R > 0$, $C_0 \geq \Delta_0$ with $C_0 > 0$, $q = 1 - \frac{\mu}{4L}$, and $G_\gamma(x) \coloneqq \frac{1}{\gamma}\left(x - \operatorname{Proj}_\mathcal{X}\left(x - \gamma\nabla f(x)\right)\right)$.

Property	Frank–Wolfe	BroxGD	Practical radius	Result
Oracle	LMO for $\mathcal{X}$	LMO for $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$	—	—
Projection-free	✓	✓	—	—
Diameter-free guarantees on possibly unbounded sets	✗	✓	—	—
Reduces to GD if $\mathcal{X} = \mathbb{R}^d$	✗	✓	—	Prop. C.1
A. Practical objective guarantees: fixed schedules, then backtracking				
Fixed horizon G-gradient bounded & convex f	—[a]	$\min_{0\leq k\leq K}\Delta_k \overset{(13)}{\leq} \frac{GR}{\sqrt{K}}$ (with radii $t_k = \frac{\varepsilon}{\|\nabla f(x_k)\|}$, $\varepsilon = \frac{GR}{\sqrt{K}}$)	✓[i]	Thm. 3.2
Fixed horizon L–smooth & convex f	$\Delta_K = \mathcal{O}(\frac{L\mathcal{D}_\mathcal{X}^2}{K})$	$\min_{0\leq k\leq K}\Delta_k \overset{(15)}{\leq} \frac{LR^2}{2K}$ (with radii $t_k = \frac{R}{\sqrt{K}}$)	✓[i]	Thm. 3.2
Predetermined geometric L–smooth & μ–strongly convex f	$\Delta_K = \mathcal{O}(\frac{L\mathcal{D}_\mathcal{X}^2}{K})$	$\Delta_K \overset{(103)}{\leq} C_0 q^K$ (with radii $t_k = \sqrt{\frac{\mu C_0}{2L^2}}\, q^{\frac{k}{2}}$)	✓[i]	Thm. D.3
Radius backtracking L–smooth & convex f	$\Delta_K = \mathcal{O}(\frac{L\mathcal{D}_\mathcal{X}^2}{K})$	$\Delta_K \overset{(17)}{\leq} \frac{\max\{\Delta_0, \frac{4LR_{\text{lev}}^2}{\xi^2}\}}{K+1}$ ($t \leftarrow \xi t$, Algorithm 2)	✓[j]	Thm. 3.3
Radius backtracking L–smooth & μ–strongly convex f	$\Delta_K = \mathcal{O}(\frac{L\mathcal{D}_\mathcal{X}^2}{K})$	$\Delta_K \overset{(18)}{\leq} (1 + \frac{\xi^2\mu}{L})^{-K}\Delta_0$ ($t \leftarrow \xi t$, Algorithm 2)	✓[j]	Thm. 3.3
B. Solution-dependent convex benchmarks: objective error and distance				
Polyak radius G-gradient bounded & convex f	—[a]	$\min_{0\leq k\leq K-1}\Delta_k \overset{D.7}{\leq} \frac{GR_0}{\sqrt{K}}$ [h] (with radii $t_k = \frac{\Delta_k}{\|\nabla f(x_k)\|}$)	✗	Thm. D.7
Polyak radius G-subgradient bounded & convex f	—	$f(\bar{x}_K) - f_\star \overset{(134)}{\leq} \frac{GR_0}{\sqrt{K}}$ (with radii $t_k = \frac{\Delta_k}{\|g_k\|}$, $g_k \in \partial f(x_k)$)	✗	Thm. D.8
Polyak radius L–smooth & convex f	$\Delta_K = \mathcal{O}(\frac{L\mathcal{D}_\mathcal{X}^2}{K})$	$f(\bar{x}_K) - f_\star \overset{(137)}{\leq} \frac{M_0R_0}{\sqrt{K}}$ (with radii $t_k = \frac{\Delta_k}{\|\nabla f(x_k)\|}$)	✗	Thm. D.9
Distance-based radius L–smooth & μ–strongly convex f	$\Delta_K = \mathcal{O}\left(\frac{L\mathcal{D}_\mathcal{X}^2}{K}\right)$	$\|x_K - x_\star\|^2 \overset{3.4}{\leq} \left(\frac{L-\mu}{L+\mu}\right)^{2K} R_0^2$ [h] (with radii $t_k = \frac{2\sqrt{\mu L}}{L+\mu}R_k$)	✗	Thm. 3.4
C. Convex gradient-difference residual guarantees				
Gradient-difference radius L–smooth & convex[b] f	$\Delta_K = \mathcal{O}\left(\frac{L\mathcal{D}_\mathcal{X}^2}{K}\right)$	$\min_{0\leq k\leq K-1} Z_k{}^2 \overset{D.11}{\leq} \frac{L^2R_0^2}{K}$ [h] (with radii $t_k = \frac{Z_k}{L}$)	✗	Thm. D.11
Constant radius L–smooth & convex f	—	$\min_{0\leq k\leq K} Z_k{}^2 \overset{(154)}{\leq} \frac{L^2R^2}{K}$ (with radii $t_k = \frac{R}{\sqrt{K}}$)	✓[i]	Thm. D.14 Cor. D.3
Gradient-difference radius Globally (L_0, L_1)–smooth & convex f	$\Delta_K = \mathcal{O}\left(\frac{(L_0 + L_1 G_K^{\text{FW}})\mathcal{D}_\mathcal{X}^2}{K}\right)$ [c]	$\min_{0\leq k\leq K-1} Z_k{}^2 \overset{D.15}{\leq} \frac{4c_\star^2R_0^2}{K}$ [h, d] (with radii $t_k = \frac{Z_k(c_k+c_\star)}{2c_kc_\star}$)	✗	Thm. D.15
D. Beyond convexity: stationarity and projected-PL guarantees				
Gradient-mapping radius L–smooth & non-convex f	$\bar{g}_K = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ [e]	$\min_{0\leq k\leq K-1}\left\|G_{1/L}(x_k)\right\|^2 \overset{D.17}{\leq} \frac{2L\Delta_0}{K}$ [h] (with radii $t_k = \frac{1}{L}\left\|G_{1/L}(x_k)\right\|$)	✗	Thm. D.17
Gradient-mapping radius L–smooth & projected-PŁ[f] f	✗[g]	$\Delta_K \overset{D.19}{\leq} \left(1 - \frac{\mu}{L}\right)^K\Delta_0$ (with radii $t_k = \frac{1}{L}\left\|G_{1/L}(x_k)\right\|$)	✗	Thm. D.19

[a] Bounded gradients need not give finite curvature, so the classical curvature-based FW rate is unavailable. This does not imply nonconvergence: uniformly continuous gradients on compact domains allow FW convergence with suitable stepsizes (Xu, 2017).
[b] Theorem D.11 assumes *strict* convexity. However, it suffices to assume that $\nabla f(x) \neq \nabla f(y)\ \forall x, y \in \mathcal{X}, x \neq y$ (see Section 3.1).
[c] A direct FW guarantee under (L_0, L_1)–smoothness appears in Vyguzov & Stonyakin (2025). Here $G_K^{\text{FW}} \coloneqq \max_{0\leq k\leq K}\|\nabla f(x_k^{\text{FW}})\|$ bounds gradients along the FW trajectory, and is therefore similar in spirit to a bounded gradient assumption.
[d] For this row, convexity and asymmetric smoothness hold on all of $\mathbb{R}^d$ (Assumption D.2), not merely on $\mathcal{X}$. The displayed bound applies when $\sqrt{K} > 2L_1R_0$; see Theorem D.15.
[e] For a compact convex $\mathcal{X}$, the FW gap is defined by $g_{\text{FW}}(x) \coloneqq \max_{s\in\mathcal{X}}\langle\nabla f(x), x - s\rangle$. For L–smooth non-convex objectives, FW guarantees that the minimum FW gap $\bar{g}_K \coloneqq \min_{0\leq k\leq K-1} g_{\text{FW}}(x_k) = \mathcal{O}(\frac{1}{\sqrt{K}})$ (Lacoste-Julien, 2016). This uses a different stationarity measure and requires boundedness of $\mathcal{X}$. It is known that $g_{\text{FW}}(x) \geq \gamma\|G_\gamma(x)\|^2$. This converts the cited FW bound into an $\mathcal{O}(K^{-\frac{1}{2}})$ upper bound on the squared gradient mapping, versus our $\mathcal{O}(K^{-1})$ bound with a projection-dependent radius. It is not a lower bound on FW or a separation in FW-gap complexity.
[f] Here, projected-PŁ means that $\frac{1}{2}\left\|G_{1/L}(x)\right\|^2 \geq \mu\left(f(x) - f(x_\star)\right)$ for all $x \in \mathcal{X}$; see Assumption D.3 for more details.
[g] We are not aware of a direct Frank–Wolfe guarantee under projected-PŁ in the above form.
[h] Same as ProjGD (see Appendix E).
[i] Practical: the radius uses neither unknown solution quantities nor a projection, given certified bounds and exact BroxGD access (Appendix D.1). Horizon rules use R, K (and G in the bounded-gradient case); geometric radii use L, μ, C_0. The nonconvex and projected-PŁ radii require a projection and receive crosses. Dashes mean not applicable or no corresponding bound is listed. Practicality concerns the radius, not the cost of the exact oracle or observability of the residual.
[j] Backtracking uses a fixed input factor $\xi \in (0, 1)$, L, the current gradient, and exact local solves. Here K counts accepted steps; rejected trials cost additional oracle calls. The convex bound assumes $\operatorname{dist}(x_k, \mathcal{X}_\star) \leq R_{\text{lev}}$ for every accepted center; R_{lev} is not the initial-distance bound R. The strongly convex bound needs no such condition and no knowledge of μ to run the method. See Appendix D.1.3.

3 THEORY

Throughout this section, we impose the following assumption.

Assumption 3.1. The set $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty, closed, and convex; $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable; and $\mathcal{X}_\star := \operatorname{argmin}_{x\in\mathcal{X}} f(x)$ is nonempty.

3.1 THE GEOMETRY OF LOCALIZATION

We first show that a suitable radius makes a single BroxGD iteration decrease the distance to an optimal solution. This statement does *not* require f to be convex.

Theorem 3.1 (One-step geometry)**.** Under Assumption 3.1, fix $x_k \in \mathcal{X}$ and $x_\star \in \mathcal{X}_\star$. Suppose either of the following radius admissibility conditions holds, with its displayed denominator nonzero:

$$\text{Type I:} \quad 0 < t_k \le \frac{\langle \nabla f(x_k), x_k - x_\star\rangle}{\|\nabla f(x_k)\|}, \tag{7}$$

$$\text{Type II:} \quad 0 < t_k \le \frac{\langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle}{\|\nabla f(x_k) - \nabla f(x_\star)\|}. \tag{8}$$

Then the update (5) satisfies

$$\text{(i)} \quad t_k \le \|x_k - x_\star\|, \qquad \text{(ii)} \quad \|x_{k+1} - x_k\| = t_k, \tag{9}$$

$$\text{(iii)} \quad \|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - t_k^2. \tag{10}$$

We now offer a quick insight into why the distance decreases. Full proofs and denominator-degeneracy discussion appear in Appendices C.6.1 and C.6.2.

For a radius satisfying Type I admissibility, let us write $g = \nabla f(x_k)$ and $y = x_{k+1}$. Admissibility ensures $\langle g, y - x_\star\rangle \ge 0$: even the largest linear decrease available inside the ball cannot cross the hyperplane passing through $x_\star$ orthogonal to g. Optimality conditions then supply the relation $g+n+\lambda(y-x_k) = 0$, with $n \in N_{\mathcal{X}}(y)$ (the normal cone of $\mathcal{X}$ at y) and $\lambda \ge 0$. Since $\langle n, y-x_\star\rangle \ge 0$, a positive multiplier gives $\langle y - x_k, y - x_\star\rangle \le 0$ and a full-radius step. If $\lambda = 0$, equality in the admissibility and Cauchy–Schwarz bounds gives the same conclusions. The two cases establish these two geometric facts:

$$\|y - x_k\| = t_k, \qquad \langle y - x_k, y - x_\star\rangle \le 0. \tag{11}$$

Expanding the squared distance now gives

$$\begin{aligned}\|y - x_\star\|^2 &= \|x_k - x_\star\|^2 - \|y - x_k\|^2 + 2\langle y - x_k, y - x_\star\rangle \\ &\overset{(11)}{\le} \|x_k - x_\star\|^2 - t_k^2.\end{aligned}$$

Type II admissibility case uses the gradient difference together with the normal-cone condition at $x_\star$; the full argument follows similar steps.

The key conclusion of Theorem 3.1 is the squared-distance decrease: a radius can be justified through separation from a solution, without performing a projection!

3.2 PRACTICAL RADII WITH A FIXED ORACLE BUDGET

The result we present here uses the geometry established in Theorem 3.1 only while the requested objective accuracy has not yet been reached; the full argument is presented in Appendix D.1.

Theorem 3.2 (Practical fixed-horizon objective guarantees)**.** Let Assumption 3.1 hold, let f be convex, fix $x_\star \in \mathcal{X}_\star$, and let $R \ge R_0$ be a known positive upper bound on $R_0 := \|x_0 - x_\star\|$. Starting from x_0, run BroxGD for $K \ge 1$ iterations, and return the iterate x_K^{best} with the smallest objective value among $x_0, \dots, x_K$.

(i) **Bounded gradients.** If f has bounded gradients, i.e. $\|\nabla f(x)\| \le G$ for all $x \in \mathcal{X}$, and we use the radius rule

$$t_k = \frac{GR}{\sqrt{K}\,\|\nabla f(x_k)\|}, \tag{12}$$

then

$$f(x_K^{\text{best}}) - f_\star \le \frac{GR}{\sqrt{K}}. \tag{13}$$

If $\nabla f(x_k) = 0$ for some k, we stop and return x_k; the same guarantee holds.

(ii) **Smooth convex objectives.** If f has an L-Lipschitz gradient on an open convex neighborhood of $\mathcal{X}$, and we use the radius rule

$$t_k = \frac{R}{\sqrt{K}}, \tag{14}$$

then

$$f(x_K^{\text{best}}) - f_\star \le \frac{LR^2}{2K}. \tag{15}$$

Note that in the smooth case, the method does not need to know L to choose the radius. The two results are proved in Corollaries D.1 and D.2. They count local-oracle calls directly and require no boundedness of $\mathcal{X}$. These guarantees control objective error, whereas the solution-dependent rule in Theorem D.11 covering the L-smooth case controls a gradient-difference residual.

3.3 Radius backtracking

We now revisit the smooth convex regime already covered by part (ii) of Theorem 3.2. However, we seek to establish a result which does *not* require the final iteration horizon K to be known in advance. Moreover, we seek a method which does *not* need to know an upper bound on R_0. We solve this problem via a *radius-backtracking* strategy for BroxGD. Its acceptance test certifies objective decrease; the convergence proof uses endpoint optimality conditions rather than the radius admissibility conditions of Theorem 3.1. This modification is formalized as Algorithm 2.

In this new method, however, we do need L as an input. As a byproduct, we prove that, under strong convexity, Algorithm 2 converges linearly with the same order of condition-number dependence as ProjGD, for fixed ξ; the contraction factors differ.

Algorithm 2 BroxGD with radius backtracking

Input: $x_0 \in \mathcal{X}$, known $L > 0$, $\xi \in (0, 1)$, horizon K, exact local linear oracle
for $k = 0, \dots, K-1$ **do**
  Compute $a_k = \nabla f(x_k)$; if $a_k = 0$, return x_k
  Set $t = \|a_k\|/L$
  loop
    Compute $y \in \operatorname{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t)} \langle a_k, z - x_k \rangle$
    Set $s = y - x_k$ and $\mathcal{D} = -\langle a_k, s \rangle$. If $\mathcal{D} = 0$, return x_k.
    if $\mathcal{D} \ge L\|s\|^2$ **then**
      Set $x_{k+1} = y$ and exit the inner loop
    end if
    Set $t = \xi t$
  end loop
end for
Output: x_K

Theorem 3.3 (Practical backtracking and last-iterate rates)**.** Let Assumption 3.1 hold and suppose f is convex with a globally L-Lipschitz gradient, where $L > 0$ is known. Fix $\xi \in (0, 1)$. Then Algorithm 2 makes finitely many local LMO calls at each ball center x_k and either stops at a minimizer or accepts an objective-decreasing step. Freeze the sequence after optimal termination, whether the stopping test is a zero gradient or $\mathcal{D} = 0$.

If, for some $R_{\text{lev}} > 0$, the accepted centers satisfy

$$\operatorname{dist}(x_k, \mathcal{X}_\star) \le R_{\text{lev}} \qquad (k \ge 0), \tag{16}$$

then, for every $K \ge 0$,

$$\Delta_K \le \frac{\max\{\Delta_0, 4LR_{\text{lev}}^2/\xi^2\}}{K+1}. \tag{17}$$

If instead f is μ-strongly convex on $\mathcal{X}$, $\mu > 0$, then without (16),

$$\Delta_K \le \left(1 + \tfrac{\xi^2 \mu}{L}\right)^{-K} \Delta_0. \tag{18}$$

The method needs neither R_{lev} nor Δ_0 nor μ. Here K counts accepted steps; rejected trials incur additional local-oracle calls.

At x_k, we reuse $a_k = \nabla f(x_k)$ while trying radii starting from $\|a_k\|/L$ and shrinking by a fixed factor $\xi \in (0, 1)$, which is a hyper-parameter of the method. We accept a local step s when its linear-model decrease dominates the smoothness remainder:

$$-\langle a_k, s\rangle \ge L\|s\|^2. \tag{19}$$

The test uses only the knowledge of L, the current gradient, and the displacement returned by the local oracle. It does not require objective evaluations or a gradient at the trial point.

Theorems D.4 to D.6 prove this result. The proof bounds an analysis multiplier between L and L/ξ^2, and combines objective descent with a selected endpoint subgradient bound. It is a separate argument from Theorem 3.1. The convex distance condition concerns the entire accepted trajectory, not just R_0; for example, bounded distance to the solution set on the initial sublevel set suffices. Finite backtracking does not by itself give a uniform bound on the total number of trial calls.

3.4 A SHARP SOLUTION-DEPENDENT BENCHMARK

Practical radius selection and the sharpest distance contraction answer different questions. The following result identifies a contraction matching optimally tuned ProjGD, but its radius uses the unknown current solution distance.

Theorem 3.4 (Sharp strongly convex contraction)**.** Let Assumption 3.1 hold and suppose f is L-smooth and μ-strongly convex on $\mathcal{X}$, where $0 < \mu \le L$. For a fixed $x_\star \in \mathcal{X}_\star$, use

$$t_k = \theta\|x_k - x_\star\|, \qquad \theta = \tfrac{2\sqrt{\mu L}}{L+\mu}. \tag{20}$$

Freeze the sequence at $x_\star$. Then, for every $K \ge 0$,

$$\|x_K - x_\star\|^2 \le \left(\tfrac{L-\mu}{L+\mu}\right)^{2K} R_0^2. \tag{21}$$

Appendix D.2.4 gives a short proof under global assumptions and an affine-hull proof under the stated assumptions on $\mathcal{X}$. The factor in (21) matches the squared-distance contraction of ProjGD with stepsize $2/(L + \mu)$ (Schmidt, 2014). This is a benchmark for the local-oracle framework; Theorem 3.3 supplies an implementable linear-rate alternative, and Theorem D.3 gives predetermined geometric radii from certified initial bounds.

3.5 FURTHER RESULTS AND COMPUTATIONAL SCOPE

The appendix develops the remaining guarantees in Table 1, following the same four groups.

Practical schedules and convex benchmarks (A–B). Predetermined geometric radii give linear objective convergence from certified initial bounds (Theorem D.3). The three Polyak-radius results cover convex objectives with bounded gradients (Theorem D.7), nonsmooth convex objectives with bounded subgradients (Theorem D.8), and smooth convex objectives (Theorem D.9). They give $O(K^{-1/2})$ objective guarantees, for the best iterate in the first case and the averaged iterate in the latter two. Their radii use the optimal objective value. Target-accuracy refinements of the main fixed-horizon guarantees also appear in Appendix D.1.

Gradient residuals and nonconvex objectives (C–D). For smooth convex objectives, gradient-difference radii give an $O(K^{-1})$ bound on the best squared residual Z_k^2 (Theorem D.11); a constant radius gives a residual hitting-time guarantee and its fixed-horizon counterpart (Theorem D.14 and corollary D.3). Global (L_0, L_1)-smoothness admits a corresponding squared-residual rate (Theorem D.15). These control gradient differences, rather than objective error. Beyond convexity,

gradient-mapping radii yield an $O(K^{-1})$ best squared-stationarity bound (Theorem D.17) and linear objective convergence under a projected-PŁ condition (Theorem D.19). These radii require a projection, as indicated in Table 1.

Oracle structure and further extensions. The table's affine-subspace result gives an explicit local oracle and recovers a gradient step when $\mathcal{X} = \mathbb{R}^d$ (Proposition C.1). Stochastic guarantees are developed separately in Appendix F. Practicality of a radius does not imply that the exact local oracle is cheap. Oracle implementations and potential applications are discussed in Appendix H; Section 4 gives a computational illustration, with full protocols and further experiments in Appendix G.

4 Correlation-matrix experiments

We test weighted quadratic losses over correlation matrices $X \succeq 0$, $\mathrm{diag}(X) = \mathbf{1}$, at orders 128, 256, and 512, with ten seeds per size and three timing repetitions. BroxGD uses Theorem 3.2's fixed radius, $R = \sqrt{n(n-1)}$ and $K_{\mathrm{rad}} = 2000n/128$, reporting its best iterate. Its affine-ball solve is exact when a Cholesky check confirms feasibility. ProjGD and ProjGD-BB share this shortcut, but can require costly Dykstra projections. Timings include all inner work; Appendix G.3 gives the full protocol.

At relative objective error 10^{-2}, BroxGD beats both projected baselines on all 30 instances. Median paired speedups over the faster baseline are $6.0\times$, $6.2\times$, and $7.0\times$ with projection tolerance 10^{-10}, and $1.9\times$, $2.0\times$, and $2.2\times$ at 10^{-4}. Initial projections cost about $50\times$ and $14\times$ a local solve, respectively.

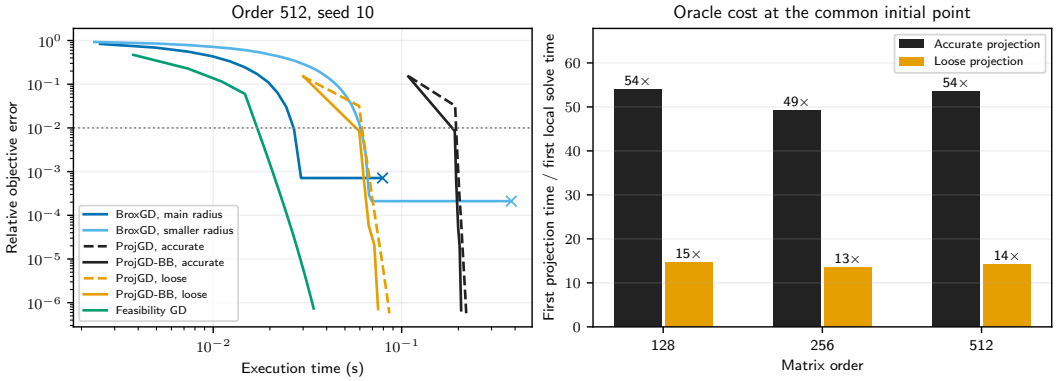


Figure 1: Left: order 512, seed 10; median timings over three repeats. Dashed/solid projected curves show ProjGD/ProjGD-BB. Crosses mark failed continuation of the affine-ball shortcut; the dotted line is the 10^{-2} target. Right: median initial projection/local-solve cost ratios over ten seeds.

This advantage is specific to low accuracy and the tested projected implementations. To test whether feasibility checks alone suffice, we include *Feasibility* GD: take the gradient with its diagonal removed, start with stepsize $1/L$, and halve it until a Cholesky check succeeds. This projection-free control is faster on 29 instances but stalls on one. The primary fixed radius reaches 10^{-3} error on 29 instances but 10^{-4} on none within the tested budget or before the shortcut fails. Further comparisons, including FW and unfavorable regimes, appear in Appendix G.

5 Conclusions

Our BroxGD framework may provide new insights into recently proposed LMO-based deep learning optimizers such as Muon, Scion and Gluon. In particular, in Appendix B.8 we show how intersecting spectral constraints with local trust-region balls leads to a damped LMO update that may be useful for neural network training. These insights are also of a preliminary nature, and are merely meant to hint at interesting future work directions. We invite the community to contribute to further development of this promising new direction.

AI USE STATEMENT

We used OpenAI Codex to assist with limited manuscript drafting and revision, formulation and revision of a few mathematical statements and proofs, experimental design, implementation and execution of the main synthetic numerical experiments, analysis of results, and preparation of figures and tables. The authors reviewed and verified all proofs (virtually all are human-generated; a couple are AI-assisted), experimental code, and reported results. Verification also included automated small-instance comparisons of numerical oracles against separate solver formulations, checks of feasibility and numerical optimality certificates, and consistency checks between saved experimental traces and reported summaries. These checks were implemented and run with AI assistance. The authors take responsibility for the final manuscript, including all AI-assisted text, mathematical claims, code, and experimental results.

REFERENCES

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3. (Cited on page 21)

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. (Cited on page 21)

Francis Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 175(1):419–459, 2019. (Cited on page 23)

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 21)

Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017. ISBN 1611974984. (Cited on page 22, 48, 66, 69, and 70)

Patrick Behr, Andre Guettler, and Felix Miebs. On portfolio optimization: Imposing the right constraints. *Journal of Banking & Finance*, 37(4):1232–1242, 2013. ISSN 0378-4266. (Cited on page 1)

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999. (Cited on page 1, 25, and 66)

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003. ISSN 1532-4435. (Cited on page 21)

Jérôme Bolte, Cyrille W. Combettes, and Edouard Pauwels. The iterates of the Frank-Wolfe algorithm may not converge. *arXiv preprint arXiv:2202.08711*, 2022. (Cited on page 4)

Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. (Cited on page 1 and 24)

Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods. *arXiv preprint arXiv:2211.14103*, 2022. (Cited on page 22 and 23)

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. (Cited on page 1)

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023. (Cited on page 1)

Sébastien Designolle, Gabriele Iommazzo, Mathieu Besançon, Sebastian Knebel, Patrick Gelß, and Sebastian Pokutta. Improved local models and new Bell inequalities via Frank-Wolfe algorithms. *Physical Review Research*, 5(4):043059, 2023. (Cited on page 23)

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 1)

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390191. (Cited on page 22)

Moran Feldman, Joseph Naor, and Roy Schwartz. A unified continuous greedy algorithm for submodular maximization. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pp. 570–579, 2011. doi: 10.1109/FOCS.2011.46. (Cited on page 23)

Michael C. Ferris and Sergei K. Zavriev. The linear convergence of a successive linear programming algorithm. Technical Report 96-12, Computer Sciences Department, University of Wisconsin–Madison, 1996. URL `https://pages.cs.wisc.edu/~ferris/techreports/96-12.pdf`. (Cited on page 4 and 25)

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956. (Cited on page 2, 3, and 22)

Dan Garber. Faster projection-free convex optimization over the spectrahedron. *Advances in Neural Information Processing Systems*, 29, 2016. (Cited on page 22)

Dan Garber and Elad Hazan. Faster rates for the Frank–Wolfe method over strongly-convex sets. In *Advances in Neural Information Processing Systems*, pp. 541–549, 2015. (Cited on page 23)

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, 2016. (Cited on page 4 and 26)

Gauthier Gidel, Fabian Pedregosa, and Simon Lacoste-Julien. Frank-Wolfe splitting via augmented Lagrangian method. In *International Conference on Artificial Intelligence and Statistics*, pp. 1456–1465. PMLR, 2018. (Cited on page 25)

Paul C. Gilmore and Ralph E. Gomory. A linear programming approach to the cutting stock problem i. *Oper Res*, 9, 01 1961. doi: 10.1287/opre.9.6.849. (Cited on page 2)

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. (Cited on page 21)

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025. (Cited on page 56)

Henry Gouk, Timothy M Hospedales, and Massimiliano Pontil. Distance-based regularisation of deep networks for fine-tuning. *arXiv preprint arXiv:2002.08253*, 2020. (Cited on page 21)

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021. (Cited on page 21)

Benjamin Grimmer and Ning Liu. Lower bounds for linear minimization oracle methods optimizing over strongly convex sets. *arXiv preprint arXiv:2602.22608*, 2026. (Cited on page 23)

Panagiotis D Grontas, Antonio Terpin, Efe C Balta, Raffaello D'Andrea, and John Lygeros. Pinet: Optimizing hard-constrained neural networks with orthogonal projection layers. *arXiv preprint arXiv:2508.10480*, 2025. (Cited on page 21)

László Grötschel, Martinand Lovász and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer Berlin, Heidelberg, 1993. doi: 10.1007/978-3-642-78240-4. (Cited on page 2)

Kaja Gruntkowska, Hanmin Li, Aadi Rane, and Peter Richtárik. The Ball-Proximal (=" Broximal") Point Method: a new algorithm, convergence theory, and applications. *arXiv preprint arXiv:2502.02002*, 2025. (Cited on page 1, 2, 18, 48, and 54)

Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's 'away step'. *Math. Program.*, 35 (1):110–119, May 1986. ISSN 0025-5610. (Cited on page 3)

Jannis Halbey, Daniel Deza, Max Zimmer, Christophe Roux, Bartolomeo Stellato, and Sebastian Pokutta. Lower bounds for Frank-Wolfe on strongly convex sets. *arXiv preprint arXiv:2602.04378*, 2026. (Cited on page 23)

Peter Hall and Brett Presnell. Density estimation under constraints. *Journal of Computational and Graphical Statistics*, 8(2):259–277, 1999. ISSN 10618600. (Cited on page 1)

Richard J. Hathaway. A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics*, 13(2):795 – 800, 1985. doi: 10.1214/aos/1176349557. (Cited on page 1)

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *LATIN 2008: Theoretical Informatics*, pp. 306–316, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. (Cited on page 3)

Deborah Hendrych, Mathieu Besançon, and Sebastian Pokutta. Solving the optimal experiment design problem with mixed-integer convex methods. *arXiv preprint arXiv:2312.11200*, 2023. (Cited on page 22)

Nicholas J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002. doi: 10.1093/imanum/22.3.329. (Cited on page 87)

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 427–435, 2013. (Cited on page 3, 23, and 54)

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. (Cited on page 27)

Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 253–268, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4. (Cited on page 23)

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. ISSN 0890-5401. (Cited on page 1)

Simon Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016. (Cited on page 5 and 23)

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank–Wolfe optimization variants. *Advances in Neural Information Processing Systems*, 28, 2015. (Cited on page 3)

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *International Conference on Machine Learning*, pp. 53–61. PMLR, 2013. (Cited on page 22)

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013. (Cited on page 3 and 23)

Claude Lemaréchal, Arkadi Nemirovski, and Yurii Nesterov. New variants of bundle methods. *Math. Program.*, 69:111–147, 07 1995. doi: 10.1007/BF01585555. (Cited on page 2)

Evgeny S. Levitin and Boris T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6:1–50, 12 1966. doi: 10.1016/0041-5553(66)90114-5. (Cited on page 2, 3, 22, 23, and 66)

Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on page 22)

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. (Cited on page 21)

John Mattingley and Stephen Boyd. Real-time convex optimization in signal processing. *Signal Processing Magazine, IEEE*, 27:50 – 61, 06 2010. doi: 10.1109/MSP.2010.936020. (Cited on page 1)

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965. (Cited on page 2 and 18)

Geoffrey Négiar, Gideon Dresdner, Alicia Tsai, Laurent El Ghaoui, Francesco Locatello, Robert Freund, and Fabian Pedregosa. Stochastic Frank-Wolfe for constrained finite-sum minimization. In *International Conference on Machine Learning*, pp. 7253–7262. PMLR, 2020. (Cited on page 22)

Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983. (Cited on page 1)

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009. (Cited on page 1)

Yurii Nesterov. Complexity bounds for primal-dual methods mini-mizing the model of objective function. , 2017. *Mathematical Programming*, 2017. (Cited on page 4 and 49)

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on page 22)

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994. doi: 10.1137/1.9781611970791. (Cited on page 24)

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer New York, NY, 2 edition, 2006. (Cited on page 22 and 25)

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. *arXiv preprint arXiv:2502.07529*, 2025. (Cited on page 27)

Sebastian Pokutta, Christoph Spiegel, and Max Zimmer. Deep neural network training with Frank-Wolfe. *arXiv preprint arXiv:2010.07243*, 2020. (Cited on page 1)

Boris T. Polyak. *Introduction to optimization*. New York, Optimization Software, 1987. (Cited on page 48)

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. (Cited on page 21)

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (bridging theory and practice of LMO-based optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025. (Cited on page 27)

Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints. In *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pp. 21–32. PMLR, 12 Dec 2020. (Cited on page 21)

J. B. Rosen. The gradient projection method for nonlinear programming. part i. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):181–217, 1960. (Cited on page 3 and 66)

Christophe Roux, Max Zimmer, Alexandre d'Aspremont, and Sebastian Pokutta. Don't be greedy, just relax! Pruning LLMs via Frank-Wolfe. *arXiv preprint arXiv:2510.13713*, 2025. (Cited on page 22)

Krzysztof E. Rutkowski. Closed-form expressions for projectors onto polyhedral sets in Hilbert spaces. *SIAM Journal on Optimization*, 27(3):1758–1771, 2017. (Cited on page 22)

Sara Sangalli, Ertunc Erdil, Andeas Hötker, Olivio Donati, and Ender Konukoglu. Constrained optimization to train neural networks on critical and under-represented classes. *Advances in Neural Information Processing Systems*, 34:25400–25411, 2021. (Cited on page 1)

Mark Schmidt. Convergence rate of Proximal–Gradient with a general step-size, 2014. (Cited on page 8)

Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Generalized conditional gradient with augmented Lagrangian for composite minimization. *SIAM Journal on Optimization*, 30(4):2687–2725, 2020. (Cited on page 25)

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 01 1996. ISSN 0035-9246. (Cited on page 21)

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning*, pp. 3570–3578. PMLR, 2017. (Cited on page 21)

AA Vyguzov and FS Stonyakin. Frank-Wolfe algorithms for (l0, l1)-smooth functions. *arXiv preprint arXiv:2510.16468*, 2025. (Cited on page 5 and 23)

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013. (Cited on page 22)

Philip Wolfe. *Integer and Nonlinear Programming*, chapter 1: Convergence Theory in Nonlinear Programming. North-Holland Publishing Company, 1970. (Cited on page 3 and 23)

Stephen J. Wright. *Primal-Dual Interior-Point Methods*. Society for Industrial and Applied Mathematics, 1997. doi: 10.1137/1.9781611971453. (Cited on page 24)

Hong-Kun Xu. Convergence analysis of the Frank–Wolfe algorithm and its generalization in banach spaces. *arXiv preprint arXiv:1710.07367*, 2017. (Cited on page 5 and 49)

Qisong Yang, Thiago Simão, Simon Tindemans, and Matthijs Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112:1–29, 06 2022. doi: 10.1007/s10994-022-06187-8. (Cited on page 1)

Adams Wei Yu, Hao Su, and Li Fei-Fei. Efficient Euclidean projections onto the intersection of norm balls. *arXiv preprint arXiv:1206.4638*, 2012. (Cited on page 22)

Alp Yurtsever, Olivier Fercoq, Francesco Locatello, and Volkan Cevher. A conditional gradient framework for composite convex minimizationwith applications to semidefinite programming. In *International Conference on Machine Learning*, 2018. (Cited on page 49)

Alp Yurtsever, Olivier Fercoq, and Volkan Cevher. A conditional-gradient-based augmented Lagrangian framework. In *International Conference on Machine Learning*, pp. 7272–7281. PMLR, 2019. (Cited on page 25)

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017. (Cited on page 21)

Yijun Zhong and Yi Shen. Stability of constrained optimization models for structured signal recovery. *arXiv preprint arXiv:2601.04849*, 2026. (Cited on page 1)

Appendix

CONTENTS

1 Introduction **1**

1.1 From proximal to broximal steps 2

1.2 Broximal gradient descent 2

1.3 Related work 3

1.4 Closely related work 4

2 Summary of contributions **4**

3 Theory **5**

3.1 The geometry of localization 6

3.2 Practical radii with a fixed oracle budget 6

3.3 Radius backtracking 7

3.4 A sharp solution-dependent benchmark 8

3.5 Further results and computational scope 8

4 Correlation-matrix experiments **9**

5 Conclusions **9**

A BroxGD as a forward–backward method with a broximal backward step **18**

A.1 Classical forward–backward splitting 18

A.2 From prox to brox 18

A.3 A forward–backward interpretation of BroxGD 19

A.4 Two variants of the broximal backward step 20

A.5 Comparison with classical ProjGD 20

A.6 A monotone inclusion viewpoint 20

B Related work **21**

B.1 Constrained optimization in machine learning 21

B.2 Projected Gradient Descent 21

B.3 Frank–Wolfe 22

B.4 Interior-point methods 24

B.5 Splitting and primal-dual approaches 25

B.6 Penalty approaches 25

B.7 Local linear optimization oracles and exact BroxGD 25

B.8 Modern motivation: LMO methods in deep learning 27

C Shared geometric tools **28**

C.1 Well-posedness of the algorithm 28

C.2 Examples of feasible sets admitting a closed-form broximal gradient step 28

C.3 Generic distance decomposition 30

C.4 Strict convexity 30

C.5 Reverse Cauchy–Schwarz inequality 31

C.6 One-step geometry and admissibility 33

C.6.1 Type-I admissibility case 33

C.6.2 Type-II admissibility case 35

C.7 Setting and separation for practical schedules 37

C.8 Affine-subspace oracle and connection to gradient descent 38

D Proofs of the convergence guarantees in Table 1 **40**

D.1 A. Practical objective guarantees 40

D.1.1 Fixed horizon with bounded gradients 40

D.1.2 Fixed horizon for smooth convex objectives 41

D.1.3 Predetermined geometric radii 42

D.1.4 Convex objective convergence with backtracking 46

D.1.5 Strongly convex objective convergence with backtracking 47

D.2 B. Solution-dependent convex benchmarks 48

D.2.1 Bounded-gradient Polyak radius 48

D.2.2 Bounded-subgradient Polyak radius 49

D.2.3 Smooth convex Polyak radius 50

D.2.4 Sharp strongly convex distance contraction 51

D.3 C. Gradient-difference residual guarantees 54

D.3.1 Smooth convex gradient-difference radius 54

D.3.2 Constant-radius residual guarantee 55

D.3.3 Generalized-smoothness gradient-difference radius 56

D.4 D. Beyond convexity 59

D.4.1 Smooth nonconvex stationarity 59

D.4.2 Projected-PL objective convergence 60

E Convergence theory of Projected Gradient Descent **63**

E.1 Smooth convex case 63

E.2 A negative result for Projected Gradient Descent 66

E.3 (L_0, L_1)–smooth convex case 68

E.4 Non-convex case 69

F Stochastic BroxGD: componentwise geometry and convergence **71**

F.1 Convex objective with bounded gradients 71
F.2 Smooth objectives and Type-II admissibility 75

G Experiments and numerical illustrations **79**
G.1 Finite-label structured quadratic: protocol and oracle costs 79
G.2 A numerical study across the application families 81
G.2.1 A selected fair-ranking example 86
G.3 Searching for a substantial projection-cost advantage 87
G.4 Problem setup 90
G.5 Geometry of the iterates 91
G.6 Gradient-difference convergence 91
G.7 Distance convergence and linear theory 92
G.8 Comparison with Projected Gradient Descent and Frank–Wolfe 93
G.9 Comparison with geometric radius schedules 94
G.10 Alternative constraint geometries 96
G.10.1 Shifted boxes 96
G.10.2 Three ball geometries 97
G.11 Discussion and scope 97

H Potential applications and local-oracle implementations **98**
H.1 A common template: locally inactive inequalities 98
H.2 Fair ranking, differentiable matching, and nonlinear transport 98
H.3 PSD-constrained metric, covariance, precision, and kernel learning 99
H.4 Occupancy measures for safe, fair, exploratory, or imitation RL 100
H.5 A common certificate for boundary-active atomic problems 101
H.6 Robust low-rank matrix learning over a nuclear-norm ball 102
H.7 Structured prediction over large products of simplices 102
H.8 What these examples establish 103

I Notation guide **104**
I.1 Core problem, iterates, and error measures 104
I.2 Geometry, regularity, and radius parameters 105
I.3 Backtracking, stochastic quantities, and application geometry 106

A BroxGD as a forward–backward method with a broximal backward step

In this section we explain that BroxGD can be viewed as a forward–backward scheme[3] for the composite problem

$$\min_{x\in\mathbb{R}^d} F(x) \coloneqq f(x) + g(x), \tag{22}$$

where

$$g(x) \coloneqq \iota_{\mathcal{X}}(x) \;\coloneqq\; \begin{cases} 0, & x \in \mathcal{X}, \\ +\infty, & x \notin \mathcal{X}. \end{cases} \tag{23}$$

Problem (22) is of course equivalent to the constrained problem

$$\min_{x\in\mathcal{X}} f(x).$$

The point of view developed below is useful for two reasons. First, it clarifies the relation between BroxGD and classical forward–backward / proximal-gradient methods. Second, it shows that BroxGD can be interpreted as a *hard-constrained* analogue of proximal gradient descent, in exactly the same spirit in which the broximal operator introduced by (Gruntkowska et al., 2025) is a hard-constrained analogue of the classical proximity operator (Moreau, 1965).

A.1 Classical forward–backward splitting

Let us first recall the usual forward–backward update for (22) in the case when f is differentiable and g is proper, closed, and convex. Given a stepsize $\gamma_k > 0$, the method takes the form

$$x_{k+1} = \operatorname{prox}_{\gamma_k g}\left(x_k - \gamma_k \nabla f(x_k)\right). \tag{24}$$

Equivalently, x_{k+1} is the minimizer of the regularized first-order model

$$x_{k+1} \in \operatorname*{argmin}_{z\in\mathbb{R}^d} \left\{ g(z) + \langle \nabla f(x_k), z - x_k\rangle + \frac{1}{2\gamma_k}\|z - x_k\|^2 \right\}. \tag{25}$$

Thus, the forward–backward method can be interpreted as follows:

- the *forward step* is the linearization of f at x_k,
- the *backward step* is the minimization of this linearized model regularized by the quadratic penalty $\frac{1}{2\gamma_k}\|z - x_k\|^2$.

In the special case (23), we have

$$\operatorname{prox}_{\gamma_k g}(y) = \operatorname{Proj}_{\mathcal{X}}(y),$$

and hence (24) reduces to Projected Gradient Descent:

$$x_{k+1} = \operatorname{Proj}_{\mathcal{X}}\left(x_k - \gamma_k \nabla f(x_k)\right).$$

A.2 From prox to brox

The broximal operator (Gruntkowska et al., 2025) replaces the quadratic penalty by a hard ball constraint. Given a proper function $h : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, a center $y \in \mathbb{R}^d$, and a radius $t > 0$, define

$$\operatorname{brox}_h^t(y) \;\coloneqq\; \operatorname*{argmin}_{z\in\mathcal{B}(y,t)} h(z), \tag{26}$$

where

$$\mathcal{B}(y,t) \coloneqq \left\{ z \in \mathbb{R}^d \;:\; \|z - y\| \le t \right\}.$$

[3] All results of this paper, including those in this section, can be formulated in the Hilbert space setting. We resort to $\mathbb{R}^d$ for simplicity of exposition only.

Hence, whereas the proximity operator solves the penalized problem

$$\operatorname*{argmin}_z \left\{ h(z) + \frac{1}{2\gamma}\|z-y\|^2 \right\},$$

the broximal operator solves the hard-constrained counterpart

$$\operatorname*{argmin}_z \{ h(z) \ : \ \|z-y\| \le t \}.$$

Thus, moving from the proximity to the broximity operator should be thought of as replacing a *soft quadratic penalty* by a *hard trust-region constraint*.

A.3 A FORWARD–BACKWARD INTERPRETATION OF BROXGD

We now apply this idea to the composite problem (22). At iteration k, define the first-order model of $F = f + g$ around x_k by

$$m_k(z) \coloneqq g(z) + \langle \nabla f(x_k), z - x_k \rangle . \tag{27}$$

The natural broximal forward–backward step associated with this model is

$$x_{k+1} \in \operatorname{brox}_{m_k}^{t_k}(x_k). \tag{28}$$

Using the definition (26), this means

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{B}(x_k,t_k)} \{ g(z) + \langle \nabla f(x_k), z - x_k \rangle \}. \tag{29}$$

Substituting $g = \iota_{\mathcal{X}}$, we obtain

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{B}(x_k,t_k)} \{ \iota_{\mathcal{X}}(z) + \langle \nabla f(x_k), z - x_k \rangle \}. \tag{30}$$

Since $\iota_{\mathcal{X}}(z) = +\infty$ outside $\mathcal{X}$, the feasible set in (30) collapses to $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$, and hence

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k,t_k)} \langle \nabla f(x_k), z - x_k \rangle . \tag{31}$$

Finally, since $-\langle \nabla f(x_k), x_k \rangle$ is constant with respect to z, (31) is equivalent to

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k,t_k)} \langle \nabla f(x_k), z \rangle . \tag{32}$$

But (32) is exactly the BroxGD update.

We have therefore proved the following interpretation.

Proposition A.1 (Forward–backward–brox interpretation of BroxGD)**.** Consider the constrained optimization problem

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

and rewrite it in composite form as

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

Then the BroxGD iteration

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k,t_k)} \langle \nabla f(x_k), z \rangle$$

is equivalent to the forward–backward step

$$x_{k+1} \in \operatorname{brox}_{m_k}^{t_k}(x_k), \qquad m_k(z) = \iota_{\mathcal{X}}(z) + \langle \nabla f(x_k), z - x_k \rangle .$$

In other words, BroxGD is obtained from classical forward–backward splitting by replacing the proximal backward step with a broximal backward step.

Proof. The equivalence follows immediately from (28)–(32). □

A.4 TWO VARIANTS OF THE BROXIMAL BACKWARD STEP

It is important to emphasize that the interpretation above is *not* the same as writing

$$x_{k+1} \in \operatorname{brox}_g^{t_k}\left(x_k - \gamma_k \nabla f(x_k)\right).$$

Indeed, this latter expression would minimize only g over a ball centered at the forward point $x_k - \gamma_k \nabla f(x_k)$, and hence would correspond to a direct replacement of $\operatorname{prox}_{\gamma_k g}(y)$ by $\operatorname{brox}_g^{t_k}(y)$ inside the classical formula (24). This is *not* what BroxGD does. Instead, BroxGD keeps the center of the ball at the current iterate x_k, linearizes the smooth part f, adds the non-smooth term g, and then applies the broximal operator to this linearized composite model:

$$x_{k+1} \in \operatorname{brox}_{g+\langle \nabla f(x_k), \cdot - x_k\rangle}^{t_k}(x_k).$$

Thus, the correct interpretation is:

> BroxGD is a forward–backward method in which the backward step is performed not on g itself, but on the *linearized composite model* $g + \langle \nabla f(x_k), \cdot - x_k\rangle$, using a broximal operator instead of a proximal one.

A.5 COMPARISON WITH CLASSICAL PROJGD

The distinction between ProjGD and BroxGD can now be expressed in one line.

For $g = \iota_{\mathcal{X}}$, projected gradient descent solves

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \iota_{\mathcal{X}}(z) + \langle \nabla f(x_k), z - x_k\rangle + \frac{1}{2\gamma_k}\|z - x_k\|^2 \right\}. \tag{33}$$

By contrast, BroxGD solves

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \iota_{\mathcal{X}}(z) + \langle \nabla f(x_k), z - x_k\rangle \ : \ \|z - x_k\| \le t_k \right\}. \tag{34}$$

Thus:

- ProjGD uses a *soft quadratic regularization*,
- BroxGD uses a *hard trust-region constraint*.

This is exactly analogous to the relation between the proximity operator and the broximal operator.

A.6 A MONOTONE INCLUSION VIEWPOINT

Throughout this subsection, assume additionally that f is convex. There is also a useful inclusion interpretation. The optimality condition for (22) is

$$0 \in \nabla f(x) + \partial g(x).$$

Since $g = \iota_{\mathcal{X}}$, we have $\partial g(x) = N_{\mathcal{X}}(x)$, and therefore the problem is equivalent to the monotone inclusion

$$0 \in \nabla f(x) + N_{\mathcal{X}}(x). \tag{35}$$

For classical forward–backward, the implicit step reads

$$0 \in \nabla f(x_k) + \partial g(x_{k+1}) + \frac{1}{\gamma_k}(x_{k+1} - x_k).$$

For BroxGD, the first-order optimality condition for (32) gives

$$0 \in \nabla f(x_k) + \partial g(x_{k+1}) + N_{\mathcal{B}(x_k, t_k)}(x_{k+1}). \tag{36}$$

Since $g = \iota_{\mathcal{X}}$, this becomes

$$0 \in \nabla f(x_k) + N_{\mathcal{X}}(x_{k+1}) + N_{\mathcal{B}(x_k, t_k)}(x_{k+1}). \tag{37}$$

Hence the Euclidean penalty term $\frac{1}{\gamma_k}(x_{k+1} - x_k)$ from proximal forward–backward is replaced by the normal cone to the trust-region ball, $N_{\mathcal{B}(x_k, t_k)}(x_{k+1})$.

B Related work

B.1 Constrained optimization in machine learning

Constrained optimization is a fundamental tool in machine learning, used to encode prior knowledge, enforce structure, and ensure desirable properties of learned models. Rather than relying solely on unconstrained objectives with regularization, many modern applications impose explicit constraints of the form

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

where the feasible set $\mathcal{X}$ captures domain-specific requirements. In this section, we provide a brief and non-exhaustive overview of key applications, focusing on representative examples rather than a complete survey.

Constraints are commonly used to impose structure on model parameters. Classical examples include sparsity constraints for feature selection (Tibshirani, 1996), low-rank constraints in matrix completion and representation learning (Recht et al., 2010), and simplex constraints in probabilistic models such as topic models (Blei et al., 2003). In deep learning, structured constraints are used to enforce sparsity, low-rank structure, or orthogonality in neural networks, improving training stability, generalization, and interpretability (Vorontsov et al., 2017; Bansal et al., 2018; Grontas et al., 2025).

A prominent example arises in adversarial robustness. In this setting, adversarial examples are generated by maximizing the loss over a norm-bounded perturbation set, typically an ℓ_p ball around a data point (Goodfellow et al., 2014; Madry et al., 2017). The constraint set encodes the allowable perturbations, and robust training then amounts to solving a constrained min–max optimization problem, where the outer minimization seeks model parameters that perform well under worst-case perturbations (Madry et al., 2017).

In sensitive applications, constraints are used to enforce fairness, safety, and reliability requirements (Zafar et al., 2017; Agarwal et al., 2018). For instance, one may impose constraints on prediction disparities across demographic groups, or enforce monotonicity and boundedness conditions in high-stakes decision systems. These constraints encode requirements that cannot be captured by the loss function alone and are increasingly important in real-world ML deployments.

Many ML models require parameters to lie in structured domains, leading naturally to constrained optimization formulations. A prominent example is the enforcement of orthogonality constraints on weight matrices, which can be formulated as optimization over the Stiefel manifold and has been used to stabilize training and improve generalization in deep networks (Absil et al., 2008; Bansal et al., 2018).

Another important class of constraints arises from enforcing Lipschitz continuity (Gouk et al., 2021), which is closely tied to robustness and stability. This leads to optimization problems with spectral or norm constraints on network weights, such as bounding operator norms (Gouk et al., 2020; Rosca et al., 2020).

In many of the settings above, the constraint set $\mathcal{X}$ is structured but complex: projections may be computationally expensive, while simpler primitives (e.g., linear minimization over $\mathcal{X}$) can be efficient. As a result, there is strong practical motivation for optimization methods that can exploit such structure without requiring costly projections. This motivates the study of alternative oracle models for constrained optimization that better align with the computational structure of ML applications. In particular, methods that avoid global projections while retaining strong convergence guarantees are of significant interest.

B.2 Projected Gradient Descent

Projection-based first-order methods are among the most classical and widely used approaches for solving constrained optimization problems of the form (1). They combine a gradient step with a projection onto the feasible set $\mathcal{X}$, typically defined via a Bregman divergence D_ϕ, i.e.,

$$\mathrm{Proj}_{\mathcal{X}}^{\phi}(x) \in \operatorname*{argmin}_{z \in \mathcal{X}} D_\phi(z, x).$$

In the Euclidean setting $\phi(x) = \frac{1}{2}\|x\|^2$, this reduces to the standard projection

$$\mathrm{Proj}_{\mathcal{X}}(x) = \arg\min_{z \in \mathcal{X}} \|x - z\|.$$

This leads to the well-known Projected Gradient Descent (ProjGD) iteration

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}\big(x_k - \gamma_k \nabla f(x_k)\big),$$

where $\gamma_k > 0$ is a stepsize, typically chosen as $\gamma_k \equiv 1/L$ for L–smooth convex objectives.

At a high level, ProjGD performs an unconstrained gradient step followed by a correction that enforces feasibility. This simple structure has made it a foundational method in constrained optimization, with extensive theoretical and algorithmic development; see, e.g., the classical references Nocedal & Wright (2006); Beck (2017); Nesterov (2018).

ProjGD enjoys convergence guarantees that match those of unconstrained gradient methods: $\mathcal{O}(1/\varepsilon)$ in the smooth convex case, $\mathcal{O}(\log(1/\varepsilon))$ in the smooth strongly convex case, $\mathcal{O}(1/\varepsilon^2)$ in the bounded-gradient convex case, and $\mathcal{O}(1/\varepsilon^2)$ for finding an ε-stationary point in the smooth non-convex case, measured by the projected gradient mapping G_γ. These are standard guarantees for unaccelerated projected methods; the objective-error and stationarity criteria are distinct.

A key implication is that constraints do not degrade the oracle complexity of the method: ProjGD achieves these guarantees without additional cost beyond projection operations. This makes it one of the few methods that cleanly separates optimization difficulty from feasibility constraints in the oracle complexity sense.

Despite its strong theoretical properties, the practical efficiency of ProjGD critically depends on the cost of computing the projection step. For simple constraint sets–such as norm balls, boxes, simplices, or other polyhedral structures–projections admit closed forms or can be computed efficiently (Duchi et al., 2008; Yu et al., 2012; Wang & Carreira-Perpinán, 2013; Rutkowski, 2017). However, for many modern applications, the projection itself requires solving a nontrivial optimization problem over $\mathcal{X}$. In such cases, each ProjGD iteration may be as expensive as solving a full auxiliary problem, making the projection step the computational bottleneck. This limitation is particularly pronounced when $\mathcal{X}$ is high-dimensional.

B.3 Frank–Wolfe

The main limitation of ProjGD is not its convergence rate, but its reliance on efficient projections. Its per-iteration cost can be prohibitive when projections are expensive. This has motivated a broad line of work on projection-free methods, which replace projections with different oracles. Their per-iteration cost depends on the relative difficulty of linear minimization and projection, and their convergence guarantees depend on the chosen variant and assumptions.

The Frank–Wolfe (FW) method (Frank & Wolfe, 1956; Levitin & Polyak, 1966), also known as the conditional gradient method (Braun et al., 2022), is a canonical example of such a projection-free first-order algorithm for constrained convex optimization. Unlike projected gradient methods, which require computing projections onto the feasible set $\mathcal{X}$, FW replaces this step with a single call per iteration to a Linear Minimization Oracle (LMO). Specifically, at iteration k, it computes

$$s_k \in \operatorname*{argmin}_{z \in \mathcal{X}} \langle \nabla f(x_k), z \rangle$$

and updates

$$x_{k+1} = x_k + \gamma_k (s_k - x_k),$$

where $\gamma_k \in [0, 1]$ is a stepsize. The LMO solves a linear problem over $\mathcal{X}$, which is often significantly cheaper than projection. As a result, FW is particularly well suited for large-scale problems where projections can be computationally prohibitive.

Due to this favorable structure, FW has found numerous applications across machine learning and optimization, including large-scale learning problems (Lacoste-Julien et al., 2013; Négiar et al., 2020), pruning large language models (Roux et al., 2025), matrix completion (Garber, 2016), optimal transport (Luise et al., 2019), optimal experiment design (Hendrych et al., 2023), image processing

(Joulin et al., 2014), submodular function maximization (Feldman et al., 2011; Bach, 2019), and quantum physics (Designolle et al., 2023).

From a theoretical perspective, FW is known to achieve a sublinear convergence rate in general. For smooth convex objectives on compact convex domains, the standard guarantee is an $\mathcal{O}(L\mathcal{D}_{\mathcal{X}}^2/K)$ objective error after K iterations (Jaggi, 2013). Oracle lower bounds concern worst-case problem classes and specified oracle models, rather than every instance or every variant (Lan, 2013). The smoothness assumption is often expressed through the FW curvature constant C_{FW} (Jaggi, 2013), defined as

$$C_{\mathsf{FW}} \coloneqq \sup_{\substack{x,s\in\mathcal{X},t\in[0,1],\\ y=x+t(s-x)}} \frac{2}{t^2}\left(f(y)-f(x)-\langle\nabla f(x),y-x\rangle\right), \tag{38}$$

rather than a global Lipschitz constant for the gradient. The same projection-free structure also extends to smooth non-convex objectives, but with a stationarity rather than primal-gap guarantee: over a compact convex set $\mathcal{X}$, the minimum FW gap satisfies $\bar{g}_K = \mathcal{O}\left(1/\sqrt{K}\right)$ (Lacoste-Julien, 2016).

More recently, Vyguzov & Stonyakin (2025) studied FW under the (L_0, L_1)–smoothness condition and proposed an (L_0, L_1)-aware short-step variant. For compact convex feasible sets, they proved an $\mathcal{O}(1/K)$ primal-gap guarantee in the general convex case, together with faster rates under additional assumptions. This extends FW analysis beyond the classical globally smooth setting. However, the result relies on a modified short-step rule adapted to the (L_0, L_1) geometry, rather than the standard curvature-based framework.

Faster convergence can only be achieved under additional structural assumptions. In particular, when the constraint set $\mathcal{X}$ is strongly convex, linear convergence is possible if the solution lies in the interior of $\mathcal{X}$ and the objective is strongly convex, or if the unconstrained minimizer lies outside $\mathcal{X}$ (Levitin & Polyak, 1966; Wolfe, 1970). These results highlight that the location of the optimizer plays a crucial role in the behavior of FW. It remained unclear whether the strong convexity of $\mathcal{X}$ alone can yield faster convergence rates without additional assumptions on the optimizer's location. This question was partially resolved by Garber & Hazan (2015), who showed that when both the objective function and the feasible set are strongly convex, FW achieves a rate of $\mathcal{O}(1/\sqrt{\varepsilon})$.

Recently, Halbey et al. (2026) proved that this rate is in fact optimal: for smooth and strongly convex objectives over smooth and strongly convex sets, FW with exact line search or short-step variants requires at least $\Omega(1/\sqrt{\varepsilon})$ iterations. This result establishes that the uniform convergence rate of $\Omega(1/\sqrt{\varepsilon})$ is tight with respect to ε, and further shows that the dependence on the optimizer's location is intrinsic rather than an artifact of earlier analyses. The lower bound applies to the stated FW variants and problem class; it does not rule out improvements by other algorithms or under additional assumptions. A concurrent work by Grimmer & Liu (2026) derived a similar lower bound of $\Omega(1/\sqrt{\varepsilon})$ for a broader class of LMO-based methods using a zero-chain construction. Their result covers a broader algorithm class, but only applies in the high-dimensional regime and its extension to smooth constraint sets is limited to more restrictive settings where the smoothness parameter scales with the target accuracy.

Beyond strong convexity, other structural assumptions can also lead to improved rates. For example, when $\mathcal{X}$ is a polytope, several FW variants, such as away-step, pairwise, and fully corrective methods, achieve faster convergence rates under suitable conditions. We refer to Braun et al. (2022) for a comprehensive overview of these variants and their theoretical guarantees.

For a modern perspective on lower bounds and their historical development, see Halbey et al. (2026); Grimmer & Liu (2026).

Frank–Wolfe curvature. We provide a simple example showing that bounded gradients on $\mathcal{X}$ do not guarantee finite Frank–Wolfe curvature.

Example B.1 (Bounded gradients do not imply finite Frank–Wolfe curvature)**.** *Consider the convex function $f : \mathcal{X} \to \mathbb{R}$ defined over $\mathcal{X} = [0, 1]$ by*

$$f(x) = x^{3/2}.$$

Then f is differentiable on $\mathcal{X}$ with gradient $\nabla f(x) = {}^3\sqrt{x}/2$. Hence, for all $x \in \mathcal{X}$,

$$\|\nabla f(x)\| \le \frac{3}{2},$$

so the bounded gradient condition (127) *holds with* $G = 3/2$.

We now show that the Frank–Wolfe curvature constant (38) *is infinite. Fix* $x = 0$ *and* $s = 1$*, and let* $y = x + t(s - x) = t$ *for* $t \in [0, 1]$*. Since the right derivative* $\nabla f(0) = 0$*, we have*

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = t^{3/2}.$$

Substituting into (38)*, we obtain*

$$\frac{2}{t^2} \left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right) = 2t^{-1/2}.$$

Taking $t \to 0$ *yields divergence, and therefore* $C_{\mathsf{FW}} = \infty$*, showing that bounded gradients alone do not control the second-order behavior captured by the Frank–Wolfe curvature constant.*

B.4 Interior-point methods

When additional structure of $\mathcal{X}$ is available, one can exploit it to design more specialized algorithms. A common setting is when $\mathcal{X}$ admits an explicit representation of the form

$$\mathcal{X} = \{x \in \mathbb{R}^d \ : \ \psi_i(x) \leq 0, \ i = 1, \ldots, m, Ax = b\}.$$

Interior-point methods approach such problems by solving a sequence of approximations to the original constrained problem, typically using Newton-type methods.

A canonical example is the barrier method (Nesterov & Nemirovskii, 1994; Boyd & Vandenberghe, 2004). The key idea is to replace the inequality constraints with a penalty term in the objective, yielding a problem to which Newton's method can be applied. To motivate this, consider rewriting the problem by incorporating the constraints via indicator functions

$$I_-(z) = \begin{cases} 0 & z \leq 0, \\ \infty & z > 0. \end{cases}$$

This leads to the equivalent formulation

$$\min_{x : Ax = b} \left\{ f(x) + \sum_{i=1}^{m} I_- \left(\psi_i(x) \right) \right\}.$$

While this removes the explicit constraints, the resulting objective is not differentiable, preventing the direct application of Newton's method. The barrier method replaces I_- with the approximation

$$\hat{I}_-(z) = \begin{cases} -\frac{1}{t} \log(-z) & z \leq 0, \\ \infty & z > 0, \end{cases}$$

where $t > 0$ controls the accuracy of the approximation. The function $\hat{I}_-$ is convex, nondecreasing, differentiable and closed, and becomes a better approximation of I_- as $t \to \infty$. This leads to the *logarithmic barrier function*

$$\phi(x) := \sum_{i=1}^{m} - \log(-\psi_i(x)),$$

and the associated problem

$$\min_{x : Ax = b} \left\{ t f(x) + \phi(x) \right\}.$$

Barrier methods approach (1) by approximately solving a sequence of such problems for an increasing sequence of scalars t, typically using second-order methods. This requires access to the first- and second-order derivatives of the functions ψ_i representing the constraints.

Barrier methods are only one class within interior-point methods. Another important class is primal–dual interior-point methods (Wright, 1997), which update primal and dual variables by solving perturbed KKT systems. For a comprehensive overview, see Boyd & Vandenberghe (2004).

B.5 SPLITTING AND PRIMAL-DUAL APPROACHES

Another closely related line of work studies projection-free methods for problems whose feasible region is described by several simpler constraints. In our case, even when linear minimization over $\mathcal{X}$ and over the ball $\mathcal{B}(x_k, t_k)$ are individually tractable, linear minimization over their intersection can be substantially harder. Splitting methods address this difficulty by decomposing a complicated feasible structure into simpler components.

In the Frank–Wolfe setting, this idea has been developed through augmented-Lagrangian and primal-dual formulations. For example, Gidel et al. (2018) consider problems in which the feasible region is represented through multiple convex sets coupled by a linear constraint. Instead of requiring a single LMO over the full intersection, their method calls LMOs over the separate component sets and enforces consistency through an augmented Lagrangian mechanism. Related primal-dual conditional-gradient methods, such as those of Yurtsever et al. (2019) and Silveti-Falls et al. (2020), similarly combine LMO-based primal updates with multiplier or proximal mechanisms for handling affine and composite constraints.

These methods are different from the BroxGD approach studied here. Splitting and primal-dual approaches avoid a difficult joint oracle and recover feasibility through consistency enforcing mechanisms, whereas our BroxGD keeps the intersection $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$ inside the oracle and therefore enforces local feasibility directly.

B.6 PENALTY APPROACHES

Penalty methods provide a related but conceptually distinct comparison point. Rather than decomposing the feasible region into simpler oracle calls, penalty methods replace hard constraints by additional objective terms that measure constraint violation (Nocedal & Wright, 2006; Bertsekas, 1999). For example, for equality constraints $h(x) = 0$ and inequality constraints $g_i(x) \le 0$, a quadratic penalty formulation considers a sequence of problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{\rho}{2} \|h(x)\|^2 + \frac{\rho}{2} \sum_i \left(\max\{0, g_i(x)\}\right)^2 \right\},$$

where $\rho > 0$ is a penalty parameter. As ρ increases, infeasible points become increasingly expensive, and solutions of the penalized problems are driven toward the feasible region. Exact penalty methods instead use non-smooth penalties, such as ℓ_1-type constraint violation terms, and can recover constrained solutions for sufficiently large finite penalty parameters under suitable assumptions.

From the perspective of BroxGD methods, a penalty based analogue would replace the exact local feasible subproblem by a softened model. Instead of solving

$$\min_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle,$$

one could consider a penalized surrogate of the form

$$\min_{z \in \mathbb{R}^d} \left\{ \langle \nabla f(x_k), z \rangle + \rho \operatorname{dist}^2(z, \mathcal{X}) + \rho \left(\max\{0, \|z - x_k\| - t_k\}\right)^2 \right\}.$$

Such a surrogate may be easier to optimize or decompose, but it does not enforce feasibility exactly for finite ρ. Thus, while penalty methods control feasibility through violation terms, BroxGD treats both $z \in \mathcal{X}$ and $z \in \mathcal{B}(x_k, t_k)$ as hard constraints of the linearized subproblem.

B.7 LOCAL LINEAR OPTIMIZATION ORACLES AND EXACT BROXGD

Shared oracle and projection connection. Our motivation is the BPM derivation in Section 1.1 and Appendix A. Here we place the resulting method alongside its antecedents. Ferris & Zavriev (1996) study linear minimization over a closed convex set intersected with a norm ball; their Euclidean subproblem is exactly (5). Their Theorem 2.10 relates its solutions to projected-gradient displacements. Thus the exact oracle and its projection connection are established ingredients.

Different algorithms: radius search and relaxed oracles. Ferris and Zavriev accept full local steps after an Armijo-type radius search using function-decrease tests, potentially solving several trial subproblems. Our analysis considers explicit radius schedules and, separately, observable backtracking. These algorithms share the local subproblem but use different radius rules and acceptance criteria.

Garber & Hazan (2016) introduced LLOO and a polytope implementation using one global LMO call plus a maintained convex decomposition. Their work also treats online, nonsmooth, and stochastic optimization; those settings are not exclusive to the present paper. For a center $x \in \mathcal{X}$, radius $t > 0$, and direction g, the abstraction requires an output p satisfying

$$p \in \mathcal{X}, \qquad \|p - x\| \le \rho t, \qquad \langle g, p\rangle \le \langle g, z\rangle \quad \forall z \in \mathcal{X} \cap \mathcal{B}(x,t), \qquad \rho \ge 1. \tag{39}$$

When $\rho = 1$, these conditions are equivalent to exact local minimization:

$$p \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle g, z\rangle. \tag{40}$$

For $\rho > 1$, (39) does not require minimizing over the enlarged ball; it specifies only feasibility, a distance bound, and comparison against the original local ball. Their strongly convex scheme uses damping $\alpha = \mu/(2L\rho^2)$ and obtains an objective bound $C_0 \exp(-\mu K/(4L\rho^2))$. Their construction has $\rho = \sqrt{d}\,\kappa_{\mathcal{P}}$, where $\kappa_{\mathcal{P}}$ denotes their polytope parameter $\mu(\mathcal{P})$, distinct from objective strong convexity. One LMO call does not eliminate the decomposition bookkeeping cost.

Different guarantees: asymptotic convergence and explicit radius bounds. Under a projected-gradient residual error bound and finiteness of stationary objective values on sublevel sets, Ferris and Zavriev's Theorem 4.5 gives asymptotic Q-linear objective convergence and convergence to a stationary point, allowing nonconvex objectives. Its iteration index counts accepted updates. Our results give explicit radius-dependent guarantees in the regimes summarized in Table 1; backtracking trial costs are accounted for separately in Appendix D.1.3. We do not claim that exact local linear minimization or linear convergence with adaptive radii originates here.

A central geometric statement in our analysis of (5) is the following: under the Type-I or Type-II admissibility conditions in Theorem 3.1, it satisfies

$$\|x_{k+1} - x_\star\|^2 \overset{(10)}{\le} \|x_k - x_\star\|^2 - t_k^2. \tag{41}$$

This decrease is relative to the selected minimizer for which admissibility holds and requires no polyhedral geometry. Exact local solves exist even on unbounded closed convex sets, whereas a global LMO can be unbounded below. The resulting convergence statements concern this stronger oracle model; they are not a comparison of wall-clock cost. The practical backtracking analysis uses objective descent and endpoint certificates, so we do not claim that every result in the paper follows from (41).

Why the relaxed axioms do not imply the same distance decrease. The following construction isolates the missing geometric implication, already on a compact rectangle and a smooth strongly convex quadratic.

Proposition B.1 (A relaxed LLOO step can increase distance under any damping)**.** For every $\rho > 1$, there exist a compact polytope $\mathcal{X}$, a smooth strongly convex f with unique minimizer $x_\star$, and a Type-I/II admissible query $(x, t, \nabla f(x))$ such that an output satisfying (39) yields $\|(1-\alpha)x + \alpha p - x_\star\| > \|x - x_\star\|$ for every $0 < \alpha \le 1$.

Proof. Fix $\rho > 1$, put $s = \sqrt{\rho^2 - 1}$, and choose $b > 1/s$. Define

$$\mathcal{X} = [-1, 1] \times [-s, b], \quad x = (0, 0), \quad t = 1, \quad x_\star = (-1, b),$$
$$f(z) = \tfrac{1}{2}(z - x_\star)^\top M (z - x_\star), \qquad M = \begin{pmatrix} 1 + b^2 & b \\ b & 1 \end{pmatrix}. \tag{42}$$

The leading principal minors of M are $1 + b^2 > 0$ and 1, so M is positive definite. Thus f is globally smooth and strongly convex, with unique minimizer $x_\star \in \mathcal{X}$. Moreover,

$$\nabla f(x) \overset{(42)}{=} (1, 0), \qquad \nabla f(x_\star) = 0, \qquad \frac{\langle \nabla f(x), x - x_\star\rangle}{\|\nabla f(x)\|} \overset{(42)}{=} 1 = t. \tag{43}$$

Hence both admissibility conditions hold. The exact solution is $z_{\mathrm{ex}} = (-1, 0)$, whose squared distance decreases by 1. In contrast, $p = (-1, -s) \in \mathcal{X}$ obeys

$$\|p - x\| = \sqrt{1 + s^2} = \rho t, \qquad \langle \nabla f(x), p \rangle \overset{(43)}{=} -1 \le \langle \nabla f(x), z \rangle \quad (z \in \mathcal{X} \cap \mathcal{B}(x, 1)). \tag{44}$$

Thus it satisfies (39). For $x^+ = (1 - \alpha)x + \alpha p$, direct expansion gives

$$\begin{aligned} \|x^+ - x_\star\|^2 - \|x - x_\star\|^2 &\overset{(42),(44)}{=} (1 - \alpha)^2 + (b + \alpha s)^2 - (1 + b^2) \\ &= 2\alpha(bs - 1) + \alpha^2 \rho^2 > 0. \end{aligned} \tag{45}$$

For $0 < \alpha \le 1$, the damped point is feasible by convexity. A valid oracle can return this p at the specified query and exact local minimizers at other queries; the example therefore respects the oracle abstraction. □

The proposition rules out a universal positive distance-decrease guarantee from the relaxed axioms alone, including one obtained merely by changing constants as a function of ρ. It does not rule out objective convergence for damped relaxed oracles, the Garber–Hazan guarantees, or stronger conclusions for particular oracle selections. The example's conditioning and domain depend on ρ; it is not a fixed-condition-number complexity lower bound.

Damping an exact local solution is different. Let z be the exact oracle point at an admissible query (x, t), and let $x^+ = (1 - \alpha)x + \alpha z$, $0 < \alpha \le 1$. The squared-norm identity and Theorem 3.1 give

$$\begin{aligned} \|x^+ - x_\star\|^2 &= (1 - \alpha)\|x - x_\star\|^2 + \alpha\|z - x_\star\|^2 - \alpha(1 - \alpha)\|z - x\|^2 \\ &\overset{(10),(9)}{\le} \|x - x_\star\|^2 - \alpha(2 - \alpha)t^2. \end{aligned} \tag{46}$$

Thus damping preserves the exact-oracle distance mechanism with a reduced coefficient. What fails in the counterexample is the geometric control supplied by an exact point inside the prescribed ball, not damping itself.

B.8 Modern motivation: LMO methods in deep learning

Although the projection-versus-LMO dichotomy has been known for decades, recent developments in large-scale machine learning have renewed interest in LMO-based optimization. Several modern optimizers for deep learning explicitly rely on LMOs, including the Muon optimizer (Jordan et al., 2024) and related methods such as Scion (Pethick et al., 2025) and Gluon (Riabinin et al., 2025). Scion is closely related to FW through its use of LMO over a norm ball. In the constrained variant, the update takes the form

$$s_k \in \operatorname*{argmin}_{z \in \mathcal{X}} \langle d_k, z \rangle, \qquad x_{k+1} = (1 - \gamma_k)x_k + \gamma_k s_k,$$

which coincides with a stochastic or momentum version of FW over the feasible set $\mathcal{X}$. In particular, when $d_k = \nabla f(x_k)$, this reduces exactly to one step of the classical FW update over $\mathcal{X}$. These developments have sparked a resurgence of interest in FW-type and projection-free methods more broadly. This renewed practical relevance motivates revisiting the theoretical foundations of LMO-based methods and addressing the classical limitations of the FW framework.

C SHARED GEOMETRIC TOOLS

C.1 WELL-POSEDNESS OF THE ALGORITHM

We now establish well-posedness of the BroxGD iteration.

Theorem C.1 (Well-posedness)**.** Let Assumption 3.1 hold. For every $x_k \in \mathcal{X}$ and every $t_k \geq 0$, the set

$$\mathcal{X} \cap \mathcal{B}(x_k, t_k)$$

is nonempty, closed, and bounded. Consequently, the optimization problem

$$\min_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle$$

admits at least one solution.

Proof. Since $x_k \in \mathcal{X}$ and $x_k \in \mathcal{B}(x_k, t_k)$, the feasible set is nonempty. It is closed as the intersection of two closed sets, and bounded because it is contained in the ball $\mathcal{B}(x_k, t_k)$. Since the objective is continuous and the feasible set is nonempty and compact, a minimizer exists. □

C.2 EXAMPLES OF FEASIBLE SETS ADMITTING A CLOSED-FORM BROXIMAL GRADIENT STEP

In this section, we list several examples of convex sets $\mathcal{X} \subseteq \mathbb{R}^d$ for which the local linear minimization problem

$$\min_{z \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle g, z \rangle \tag{47}$$

admits a closed-form solution, assuming throughout that $x \in \mathcal{X}$, $t > 0$, and $g \neq 0$. Let $u \coloneqq {}^{g}/_{\|g\|}$. Geometrically, problem (47) asks for the feasible point within distance at most t from x whose displacement from x has the largest component in the direction $-u$. The most transparent families are affine sets, one-dimensional convex sets, Euclidean balls, and orthogonal products of these. These examples illustrate that the local linear minimization oracle required by Equation (5) is not merely an abstract object, but can be evaluated explicitly in a number of relevant cases.

1. Whole space. Let $\mathcal{X} = \mathbb{R}^d$. Then the unique solution is simply the point on the boundary of the ball in direction $-u$:

$$z^\star = x - tu.$$

2. Singleton. Let $\mathcal{X} = \{c\}$. Since $x \in \mathcal{X}$, we necessarily have $x = c$, and therefore $z^\star = c$.

3. Affine subspace. Let $\mathcal{X} = a + \mathcal{V}$, where $\mathcal{V} \subseteq \mathbb{R}^d$ is a linear subspace, and let $\mathrm{Proj}_{\mathcal{V}}$ denote the orthogonal projector onto $\mathcal{V}$. Since feasible displacements from x must lie in $\mathcal{V}$, the problem reduces to minimizing the linear form over the Euclidean ball in $\mathcal{V}$. Hence

$$z^\star = \begin{cases} x - t\dfrac{\mathrm{Proj}_{\mathcal{V}} g}{\|\mathrm{Proj}_{\mathcal{V}} g\|}, & \mathrm{Proj}_{\mathcal{V}} g \neq 0, \\ x, & \mathrm{Proj}_{\mathcal{V}} g = 0. \end{cases}$$

(see Appendix C.8).

4. Hyperplane. Let $\mathcal{X} = \{z \in \mathbb{R}^d : \ a^\top z = b\}$, $a \neq 0$. This is a special case of an affine subspace. If we define the component of the gradient tangent to the hyperplane by $g_\top \coloneqq g - \frac{a^\top g}{\|a\|^2} a$, then

$$z^\star = \begin{cases} x - t\,\dfrac{g_\top}{\|g_\top\|}, & g_\top \neq 0, \\ x, & g_\top = 0. \end{cases}$$

5. Affine line. Let $\mathcal{X} = a + \mathrm{span}\{v\}$, $v \neq 0$. Denote $\hat{v} \coloneqq v/\|v\|$. Since feasible displacements are one-dimensional, the solution is obtained by moving to one endpoint of the feasible segment:

$$z^\star = \begin{cases} x - t\,\hat{v}, & \langle g, \hat{v}\rangle > 0, \\ x + t\,\hat{v}, & \langle g, \hat{v}\rangle < 0, \\ \text{any point in } \mathcal{X} \cap \mathcal{B}(x,t), & \langle g, \hat{v}\rangle = 0. \end{cases}$$

6. Ray. Let $\mathcal{X} = a + \{\alpha v :\ \alpha \geq 0\}$, $\|v\| = 1$. Write $x = a + \alpha_x v$, $\alpha_x \geq 0$. Then feasible points in $\mathcal{X} \cap \mathcal{B}(x,t)$ are of the form $a + \alpha v$, where $\alpha \in [\max\{0, \alpha_x - t\},\ \alpha_x + t]$. Hence

$$z^\star = a + \alpha^\star v,$$

where

$$\alpha^\star = \begin{cases} \max\{0, \alpha_x - t\}, & \langle g, v\rangle > 0, \\ \alpha_x + t, & \langle g, v\rangle < 0, \\ \text{any feasible } \alpha, & \langle g, v\rangle = 0. \end{cases}$$

7. Line segment. Let $\mathcal{X} = [a,b] \coloneqq \{(1-\lambda)a + \lambda b :\ \lambda \in [0,1]\}$, suppose $x = (1-\lambda_x)a + \lambda_x b$. Then feasible points are parameterized by $\lambda \in [0,1] \cap \left[\lambda_x - \frac{t}{\|b-a\|},\ \lambda_x + \frac{t}{\|b-a\|}\right]$. Thus

$$z^\star = (1-\lambda^\star)a + \lambda^\star b,$$

where $\lambda^\star$ is the left or right endpoint of this interval depending on the sign of $\langle g, b-a\rangle$:

$$\lambda^\star = \begin{cases} \text{left endpoint}, & \langle g, b-a\rangle > 0, \\ \text{right endpoint}, & \langle g, b-a\rangle < 0, \\ \text{any feasible } \lambda, & \langle g, b-a\rangle = 0. \end{cases}$$

8. Euclidean ball. Let $\mathcal{X} = \mathcal{B}(c,R) \coloneqq \{z \in \mathbb{R}^d :\ \|z - c\| \leq R\}$. The local problem is

$$\min_{z \in \mathcal{B}(c,R) \cap \mathcal{B}(x,t)} \langle g, z\rangle, \qquad u \coloneqq \frac{g}{\|g\|}.$$

There are three cases:

- If the minimizer over $\mathcal{B}(x,t)$ remains in $\mathcal{X}$, namely if $\|x - tu - c\| \leq R$, then
$$z^\star = x - tu.$$
- If $x - tu \notin \mathcal{X}$, but the minimizer over $\mathcal{X}$ lies in $\mathcal{B}(x,t)$, namely if $\|c - Ru - x\| \leq t$, then
$$z^\star = c - Ru.$$
- In the remaining case, neither of the two single-ball minimizers is feasible for the other constraint. Hence, both constraints are active at the optimum, and $z^\star$ lies on the intersection $\|z - x\| = t$ and $\|z - c\| = R$. In this case $x \neq c$, so the following quantities are well defined:
$$d \coloneqq c - x, \quad \rho \coloneqq \|d\|, \quad e_1 \coloneqq \frac{d}{\rho}, \quad \alpha \coloneqq \frac{\rho^2 + t^2 - R^2}{2\rho}, \quad s \coloneqq \sqrt{t^2 - \alpha^2}.$$
Let $p \coloneqq u - (u^\top e_1)e_1$ be the component of u orthogonal to e_1. If $p \neq 0$, then
$$z^\star = x + \alpha e_1 - s\,\frac{p}{\|p\|}.$$
If $p = 0$, then every point in the sphere intersection is optimal. In the degenerate case $s = 0$, this point is unique.

9. Slab between two parallel hyperplanes. Let $\mathcal{X} = \{z \in \mathbb{R}^d : \ell \le a^\top z \le r\}$, where $a \neq 0$ and $\ell \le r$, and suppose $x \in \mathcal{X}$. If $g = 0$ or $t = 0$, choose $z^\star = x$. Otherwise set $u = g/\|g\|$ and $z_{\mathrm{ball}} = x - tu$. If $z_{\mathrm{ball}} \in \mathcal{X}$, it is the solution. If it violates the slab, define the violated boundary by

$$b = \begin{cases} \ell, & a^\top z_{\mathrm{ball}} < \ell, \\ r, & a^\top z_{\mathrm{ball}} > r. \end{cases} \tag{48}$$

The intersection of the original ball with this boundary hyperplane has center c and radius s, where

$$c = x + \frac{b - a^\top x}{\|a\|^2} a, \qquad s = \sqrt{t^2 - \frac{(b - a^\top x)^2}{\|a\|^2}}, \qquad g_\top = g - \frac{a^\top g}{\|a\|^2} a. \tag{49}$$

The square root is real because the segment from x to z_{ball} crosses the hyperplane inside the ball. An exact oracle selection is

$$z^\star = \begin{cases} c - s\dfrac{g_\top}{\|g_\top\|}, & g_\top \neq 0, \\ c, & g_\top = 0. \end{cases} \tag{50}$$

To justify the active boundary, any slab-feasible point off that boundary can be moved toward z_{ball} until it reaches the boundary, staying in the ball and slab and strictly decreasing the linear objective. Here strict decrease follows because z_{ball} is the unique minimizer over the ball when $g \neq 0$ and $t > 0$. For a point $z = c + v$ on the boundary, $a^\top v = 0$, and orthogonality gives

$$\|z - x\|^2 \overset{(49)}{=} \frac{(b - a^\top x)^2}{\|a\|^2} + \|v\|^2, \qquad \langle g, z\rangle \overset{(49)}{=} \langle g, c\rangle + \langle g_\top, v\rangle. \tag{51}$$

Thus feasibility is equivalent to $\|v\| \le s$, and Cauchy–Schwarz proves (50). If $g_\top = 0$, every point of this cross-section is optimal; the formula selects its center. It also covers $\ell = r$ and a zero cross-section radius.

C.3 Generic distance decomposition

The following simple lemma plays an important role in the convergence proof.

Lemma C.1 (Generic decomposition)**.** For any three points $x_+, x, z \in \mathbb{R}^d$,

$$\|x_+ - z\|^2 = \|x - z\|^2 - 2\langle x - x_+, x_+ - z\rangle - \|x - x_+\|^2.$$

In particular,

$$\|x_{k+1} - x_\star\|^2 = \|x_k - x_\star\|^2 - 2\langle x_k - x_{k+1},\, x_{k+1} - x_\star\rangle - \|x_k - x_{k+1}\|^2. \tag{52}$$

Proof. Observe that $x - z = (x - x_+) + (x_+ - z)$. Hence, by expanding the squared norm, we get

$$\|x - z\|^2 = \|(x - x_+) + (x_+ - z)\|^2 = \|x - x_+\|^2 + \|x_+ - z\|^2 + 2\langle x - x_+, x_+ - z\rangle.$$

Rearranging this identity yields

$$\|x_+ - z\|^2 = \|x - z\|^2 - 2\langle x - x_+, x_+ - z\rangle - \|x - x_+\|^2,$$

which proves the first claim. The second identity follows immediately by substituting

$$x_+ = x_{k+1}, \qquad x = x_k, \qquad z = x_\star. \qquad \square$$

C.4 Strict convexity

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function. Recall that f is *strictly convex* if for every pair of distinct points $x, y \in \operatorname{dom} f$ and every $\lambda \in (0, 1)$, we have

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda) f(y).$$

Lemma C.2 (Strict convexity implies strict monotonicity of the gradient)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strictly convex on a convex set $\mathcal{X} \subseteq \mathbb{R}^d$. Then

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle > 0, \qquad \nabla f(x) \neq \nabla f(y)$$

for any two distinct points $x, y \in \mathcal{X}$.

Proof. Since f is convex and differentiable, the first-order inequality holds at every point:

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x \rangle, \tag{53}$$

and

$$f(x) \geq f(y) + \langle \nabla f(y),\, x - y \rangle.$$

We claim that if $x \neq y$, then both inequalities are in fact strict:

$$f(y) > f(x) + \langle \nabla f(x),\, y - x \rangle, \tag{54}$$

$$f(x) > f(y) + \langle \nabla f(y),\, x - y \rangle. \tag{55}$$

Indeed, suppose for contradiction that equality holds in (53), i.e., $f(y) = f(x) + \langle \nabla f(x), y - x \rangle$. Let $z_t \coloneqq (1-t)x + ty$, $t \in [0,1]$ By convexity of f,

$$f(z_t) \leq (1-t)f(x) + tf(y).$$

Using the assumed equality, the right-hand side becomes

$$(1-t)f(x) + t\big(f(x) + \langle \nabla f(x), y - x \rangle\big) = f(x) + t\langle \nabla f(x), y - x \rangle.$$

On the other hand, applying (53) with z_t in place of y, we get

$$f(z_t) \geq f(x) + \langle \nabla f(x), z_t - x \rangle = f(x) + t\langle \nabla f(x), y - x \rangle.$$

Hence

$$f(z_t) = f(x) + t\langle \nabla f(x), y - x \rangle = (1-t)f(x) + tf(y) \qquad \forall t \in [0,1].$$

Thus f is affine on the segment $[x, y]$, which contradicts strict convexity since $x \neq y$. Therefore (54) holds. The proof of (55) is identical. Now add (54) and (55):

$$f(y) + f(x) > f(x) + \langle \nabla f(x), y - x \rangle + f(y) + \langle \nabla f(y), x - y \rangle.$$

Canceling $f(x) + f(y)$ from both sides yields

$$0 > \langle \nabla f(x), y - x \rangle + \langle \nabla f(y), x - y \rangle.$$

Rearranging gives $\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0$, which proves the claim. □

C.5 Reverse Cauchy–Schwarz inequality

The Cauchy–Schwarz inequality says that

$$\langle u, v \rangle \leq \|u\| \|v\|$$

for any vectors $u, v \in \mathbb{R}^d$. In general, Cauchy–Schwarz does not admit a reverse form. However, when u and v are special, a reverse form of the Cauchy–Schwarz inequality holds. In particular, such a reverse inequality holds for

$$v = x - y, \qquad u = \nabla f(x) - \nabla f(y),$$

where f is L–smooth and μ–strongly convex, stated and proved below. It plays an important role in our convergence proof of BroxGD in the smooth and strongly convex regime.

Lemma C.3. Let $U \subseteq \mathbb{R}^d$ be open and convex. If $f : U \to \mathbb{R}$ is differentiable, L–smooth, and μ–strongly convex, with $0 < \mu \leq L$, then

$$\theta \, \|\nabla f(x) - \nabla f(y)\| \, \|x - y\| \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle, \qquad x, y \in U, \tag{56}$$

where $\theta = 2\sqrt{\mu L}/(L + \mu)$.

Proof. Fix $x, y \in U$ and put $d = x - y$. The case $d = 0$ is immediate. The compact segment $[y, x]$ has a convex tubular neighborhood whose closure lies in U. Convolve f there with a nonnegative smooth unit-mass kernel of sufficiently small support, obtaining f_ε. Averaging translates preserves μ-strong convexity and L-smoothness, so

$$\mu I \preceq A_\varepsilon := \int_0^1 \nabla^2 f_\varepsilon(y + sd)\, ds \preceq LI, \qquad h_\varepsilon := \nabla f_\varepsilon(x) - \nabla f_\varepsilon(y) = A_\varepsilon d. \tag{57}$$

The spectral theorem gives

$$(L + \mu)\langle h_\varepsilon, d\rangle - \|h_\varepsilon\|^2 - \mu L\|d\|^2 \overset{(57)}{=} d^\top (A_\varepsilon - \mu I)(LI - A_\varepsilon)d \overset{(57)}{\geq} 0.$$

Since ∇f is continuous, the mollified gradients converge to it as $\varepsilon \downarrow 0$. Thus

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\|\nabla f(x) - \nabla f(y)\|^2 + \mu L\|x - y\|^2}{L + \mu}. \tag{58}$$

Writing $h = \nabla f(x) - \nabla f(y)$, the arithmetic–geometric mean inequality yields

$$\langle h, d\rangle \overset{(58)}{\geq} \frac{\|h\|^2 + \mu L\|d\|^2}{L + \mu} \geq \frac{2\sqrt{\mu L}}{L + \mu}\|h\|\,\|d\|.$$

This proves (56), including $L = \mu$. □

C.6 One-step geometry and admissibility

For the reader's convenience, we restate the Theorem 3.1 here.

Theorem 3.1 (One-step geometry)**.** Under Assumption 3.1, fix $x_k \in \mathcal{X}$ and $x_\star \in \mathcal{X}_\star$. Suppose either of the following radius admissibility conditions holds, with its displayed denominator nonzero:

$$\text{Type I:} \quad 0 < t_k \leq \frac{\langle \nabla f(x_k), x_k - x_\star \rangle}{\|\nabla f(x_k)\|}, \tag{7}$$

$$\text{Type II:} \quad 0 < t_k \leq \frac{\langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle}{\|\nabla f(x_k) - \nabla f(x_\star)\|}. \tag{8}$$

Then the update (5) satisfies

$$\text{(i)} \quad t_k \leq \|x_k - x_\star\|, \qquad \text{(ii)} \quad \|x_{k+1} - x_k\| = t_k, \tag{9}$$

$$\text{(iii)} \quad \|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2. \tag{10}$$

Admissibility and interpretation. The key inequality (10) shows that each step yields a *Fejér-type decrease* with respect to the solution set $\mathcal{X}_\star$. Note that (9) follows from (10). However, while (9) is a simple consequence of Cauchy–Schwarz, the proof of (10) is more involved. So, while not necessary from a mathematical perspective, we decided to state and prove (9) separately to make it clearer and intuitive why (9) holds.

A subtle point concerns the admissibility conditions in Theorem 3.1, which involve ratios of the form (7)-(8). These expressions require that the corresponding gradients (or gradient differences) are nonzero. When the denominator in (7) vanishes, the situation can be treated separately: if $\nabla f(x_k) = 0$, then x_k is a first-order stationary point, which implies optimality when f is convex. This highlights an interesting structural feature of the framework. When $\nabla f(x_\star) = 0$, the constraint set plays no active role at the solution, since the unconstrained minimizer already lies in $\mathcal{X}$. In this case, the problem effectively reduces to unconstrained optimization, and one could in principle discard the constraint and run vanilla GD instead. Nevertheless, even when the constraint is inactive at optimality, it may still provide algorithmic benefit: prior knowledge that $x_\star \in \mathcal{X}$ can be exploited to improve performance. More generally, our method can leverage such structural information by inheriting the improved conditioning of the restricted problem. For instance, when $\mathcal{X}$ is an affine subspace of $\mathbb{R}^d$, Algorithm 1 reduces to GD on that subspace, and its convergence is governed by the corresponding subspace condition number rather than the ambient one. This leads to acceleration whenever the objective is better conditioned along the constraint set. For an affine constraint, projected gradient steps also reduce to gradient steps on the subspace; this restricted conditioning is therefore shared by the two methods.

The Type-II admissibility condition (8) is more delicate. Since it involves gradient differences, the denominator can vanish without implying optimality, and hence such a condition cannot be used as a stopping criterion without additional assumptions. To rule out such degeneracies, one must assume certain identifiability assumptions (e.g., strict convexity; see Lemma C.2), which guarantee that equality of gradients implies equality of points.

C.6.1 Type-I admissibility case

Proof. By the radius condition (7), and the Cauchy–Schwarz inequality, we have

$$t_k \overset{(7)}{\leq} \frac{\langle \nabla f(x_k), x_k - x_\star \rangle}{\|\nabla f(x_k)\|} \leq \frac{\|\nabla f(x_k)\| \, \|x_k - x_\star\|}{\|\nabla f(x_k)\|} = \|x_k - x_\star\|,$$

establishing claim *(i)*. Let us now establish claims *(ii)* and *(iii)*. Fix $k \geq 0$, and for the purposes of this proof, let us use the shorter notation $\mathcal{B}_k \coloneqq \mathcal{B}(x_k, t_k)$.

Step 1: A favorable half-space inequality. Using the trivial decomposition

$$x_{k+1} - x_\star = (x_k - x_\star) + (x_{k+1} - x_k),$$

we can write

$$\langle \nabla f(x_k), x_{k+1} - x_\star \rangle = \langle \nabla f(x_k), x_k - x_\star \rangle + \langle \nabla f(x_k), x_{k+1} - x_k \rangle. \tag{59}$$

Since $x_{k+1} \in \mathcal{B}_k$, we have $\|x_{k+1} - x_k\| \le t_k$. Using the Cauchy–Schwarz inequality, we can therefore bound

$$\begin{aligned} \langle \nabla f(x_k), x_{k+1} - x_k \rangle &\ge -\|\nabla f(x_k)\| \, \|x_{k+1} - x_k\| \\ &\ge -\|\nabla f(x_k)\| \, t_k. \end{aligned} \tag{60}$$

Plugging this bound back into (59), and using the radius bound (7), we get

$$\begin{aligned} \langle \nabla f(x_k), x_{k+1} - x_\star \rangle &\overset{(59)+(60)}{\ge} \langle \nabla f(x_k), x_k - x_\star \rangle - \|\nabla f(x_k)\| \, t_k \\ &\overset{(7)}{\ge} 0. \end{aligned} \tag{61}$$

Step 2: KKT conditions for a constrained linear optimization problem. Since x_{k+1} minimizes the linear functional $z \mapsto \langle \nabla f(x_k), z \rangle$ over the nonempty closed convex set $\mathcal{X} \cap \mathcal{B}_k$, we have

$$0 \in \nabla f(x_k) + N_{\mathcal{X}}(x_{k+1}) + N_{\mathcal{B}_k}(x_{k+1}),$$

where $N_{\mathcal{C}}(z)$ denotes the normal cone of a set $\mathcal{C}$ at a point z. Hence there exist

$$n_k \in N_{\mathcal{X}}(x_{k+1}), \qquad \lambda_k \ge 0$$

such that

$$\nabla f(x_k) + n_k + \lambda_k (x_{k+1} - x_k) = 0. \tag{62}$$

Step 3: Deductions from the KKT conditions. Taking the inner product of the zero vector from (62) with $x_{k+1} - x_\star$ gives the identity

$$\langle \nabla f(x_k), x_{k+1} - x_\star \rangle + \langle n_k, x_{k+1} - x_\star \rangle + \lambda_k \langle x_{k+1} - x_k, x_{k+1} - x_\star \rangle = 0. \tag{63}$$

Note that due to (61), the first term in (63) is nonnegative. We shall now argue that the second term is nonnegative, too. Indeed, since $n_k \in N_{\mathcal{X}}(x_{k+1})$ and $x_\star \in \mathcal{X}$, we must have

$$\langle n_k, x_{k+1} - x_\star \rangle \ge 0. \tag{64}$$

Since the sum of the three terms is zero, the third term in (63) must be nonpositive, which can be equivalently written as

$$\lambda_k \langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle \ge 0. \tag{65}$$

Step 4: Two cases. We now split the argument into two cases.

Case 1: $\lambda_k > 0$. Then x_{k+1} lies on the boundary of the ball $\mathcal{B}_k$, and thus

$$\|x_{k+1} - x_k\| = t_k. \tag{66}$$

So, identity *(ii)* holds in this case. Moreover, (65) implies that

$$\langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle \ge 0. \tag{67}$$

Therefore

$$\begin{aligned} \|x_{k+1} - x_\star\|^2 &= \|x_k - x_\star - (x_k - x_{k+1})\|^2 \\ &\overset{(52)}{=} \|x_k - x_\star\|^2 - 2\langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle - \|x_k - x_{k+1}\|^2 \\ &\overset{(66)+(67)}{\le} \|x_k - x_\star\|^2 - t_k^2. \end{aligned}$$

So, inequality *(iii)* holds in this case.

Case 2: $\lambda_k = 0$**.** Then (62) becomes $\nabla f(x_k) + n_k = 0$, which means that

$$0 \in \nabla f(x_k) + N_{\mathcal{X}}(x_{k+1}).$$

Therefore, x_{k+1} minimizes the linear function $z \mapsto \langle \nabla f(x_k), z\rangle$ over all of $\mathcal{X}$. In particular, it must be the case that $\langle \nabla f(x_k), x_{k+1} - x_\star\rangle \le 0$. Combined with (61), this yields

$$\langle \nabla f(x_k), x_{k+1} - x_\star\rangle = 0. \tag{68}$$

Likewise, from $\nabla f(x_k) + n_k = 0$ and the normal-cone inequality $\langle n_k, x_{k+1} - x_\star\rangle \ge 0$ (recall (64)), we also get

$$\langle n_k, x_{k+1} - x_\star\rangle = 0.$$

Further, expanding (68), we obtain

$$0 = \langle \nabla f(x_k), x_k - x_\star\rangle + \langle \nabla f(x_k), x_{k+1} - x_k\rangle.$$

Using (7) and the Cauchy–Schwarz inequality, we arrive at

$$\begin{aligned} 0 &= \langle \nabla f(x_k), x_k - x_\star\rangle + \langle \nabla f(x_k), x_{k+1} - x_k\rangle \\ &\overset{(7)}{\ge} \|\nabla f(x_k)\|\, t_k - \|\nabla f(x_k)\|\, \|x_{k+1} - x_k\|. \end{aligned}$$

Hence $\|x_{k+1} - x_k\| \ge t_k$. However, since $x_{k+1} \in \mathcal{B}_k$, we also have $\|x_{k+1} - x_k\| \le t_k$. Therefore,

$$\|x_{k+1} - x_k\| = t_k, \tag{69}$$

and hence identity *(ii)* holds in this case also.

Moreover, equality in Cauchy–Schwarz implies that $x_{k+1} - x_k$ is a multiple of $-\nabla f(x_k)$. Since $\|x_{k+1} - x_k\| = t_k$, we must therefore have

$$x_{k+1} - x_k = -\, t_k\, \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}.$$

Since $\langle \nabla f(x_k), x_{k+1} - x_\star\rangle = 0$ (see (68)), it follows that

$$\langle x_k - x_{k+1}, x_{k+1} - x_\star\rangle = 0. \tag{70}$$

Therefore,

$$\begin{aligned} \|x_{k+1} - x_\star\|^2 &\overset{(52)}{=} \|x_k - x_\star\|^2 - 2\langle x_k - x_{k+1}, x_{k+1} - x_\star\rangle - \|x_k - x_{k+1}\|^2 \\ &\overset{(69)+(70)}{=} \|x_k - x_\star\|^2 - t_k^2. \end{aligned}$$

So, inequality *(iii)* holds in this case also. □

C.6.2 Type-II admissibility case

Proof. By the radius condition (8), and the Cauchy–Schwarz inequality, we have

$$t_k \overset{(8)}{\le} \frac{\langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle}{\|\nabla f(x_k) - \nabla f(x_\star)\|} \le \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|\, \|x_k - x_\star\|}{\|\nabla f(x_k) - \nabla f(x_\star)\|} = \|x_k - x_\star\|,$$

establishing claim *(i)*. Let us now establish claims *(ii)* and *(iii)*. Fix $k \ge 0$, and for the purposes of this proof, let us use the shorter notation $\mathcal{B}_k := \mathcal{B}(x_k, t_k)$.

Step 1: Perturbed monotonicity. Using the trivial decomposition

$$x_{k+1} - x_\star = (x_k - x_\star) + (x_{k+1} - x_k),$$

we can write

$$\begin{aligned} &\langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_\star\rangle \\ &= \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle + \langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_k\rangle. \end{aligned} \tag{71}$$

Since $x_{k+1} \in \mathcal{B}_k$, we have $\|x_{k+1} - x_k\| \leq t_k$. Using the Cauchy–Schwarz inequality, we can therefore bound

$$\begin{aligned} \langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_k \rangle &\geq -\|\nabla f(x_k) - \nabla f(x_\star)\| \, \|x_{k+1} - x_k\| \\ &\geq -\|\nabla f(x_k) - \nabla f(x_\star)\| \, t_k. \end{aligned} \tag{72}$$

Plugging this bound back into (71), and using the radius bound (8), we get

$$\begin{aligned} \langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_\star \rangle &\overset{(71),(72)}{\geq} \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle \\ &\quad - \|\nabla f(x_k) - \nabla f(x_\star)\| t_k \\ &\overset{(8)}{\geq} 0. \end{aligned} \tag{73}$$

Step 2: Applying the first-order optimality condition. Since $x_\star$ minimizes f over $\mathcal{X}$, the first-order optimality condition for (1) gives $\langle \nabla f(x_\star), z - x_\star \rangle \geq 0$, for all $z \in \mathcal{X}$. Since $x_{k+1} \in \mathcal{X}$, this implies

$$\langle \nabla f(x_\star), x_{k+1} - x_\star \rangle \geq 0. \tag{74}$$

Step 3: Combining Steps 1 and 2. Adding (73) and (74), we obtain

$$\langle \nabla f(x_k), x_{k+1} - x_\star \rangle = \langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_\star \rangle + \langle \nabla f(x_\star), x_{k+1} - x_\star \rangle \geq 0. \tag{75}$$

Step 4: KKT conditions for a constrained linear optimization problem. Since x_{k+1} minimizes the linear functional $z \mapsto \langle \nabla f(x_k), z \rangle$ over the nonempty closed convex set $\mathcal{X} \cap \mathcal{B}_k$, we have

$$0 \in \nabla f(x_k) + N_{\mathcal{X}}(x_{k+1}) + N_{\mathcal{B}_k}(x_{k+1}),$$

where $N_{\mathcal{C}}(z)$ denotes the normal cone of set $\mathcal{C}$ at point z . Hence there exist

$$n_k \in N_{\mathcal{X}}(x_{k+1}), \qquad \lambda_k \geq 0$$

such that

$$\nabla f(x_k) + n_k + \lambda_k (x_{k+1} - x_k) = 0. \tag{76}$$

Step 5: Deductions from the KKT conditions. Taking the inner product of the zero vector from (76) with $x_{k+1} - x_\star$ gives the identity

$$\langle \nabla f(x_k), x_{k+1} - x_\star \rangle + \langle n_k, x_{k+1} - x_\star \rangle + \lambda_k \langle x_{k+1} - x_k, x_{k+1} - x_\star \rangle = 0. \tag{77}$$

Note that due to (75), the first term in (77) is nonnegative. We shall now argue that the second term is nonnegative, too. Indeed, since $n_k \in N_{\mathcal{X}}(x_{k+1})$ and $x_\star \in \mathcal{X}$, we must have

$$\langle n_k, x_{k+1} - x_\star \rangle \geq 0. \tag{78}$$

Since the sum of the three terms is zero, the third term in (77) must be nonpositive, which can be equivalently written as

$$\lambda_k \langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle \geq 0. \tag{79}$$

Step 6: Two cases. We now split the argument into two cases.

Case 1: $\lambda_k > 0$**.** Then x_{k+1} lies on the boundary of the ball $\mathcal{B}_k$, and thus

$$\|x_{k+1} - x_k\| = t_k. \tag{80}$$

So, identity *(ii)* holds in this case. Moreover, (79) implies that

$$\langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle \geq 0. \tag{81}$$

Therefore

$$\begin{aligned} \|x_{k+1} - x_\star\|^2 &= \|x_k - x_\star - (x_k - x_{k+1})\|^2 \\ &\overset{(52)}{=} \|x_k - x_\star\|^2 - 2\langle x_k - x_{k+1}, x_{k+1} - x_\star \rangle - \|x_k - x_{k+1}\|^2 \\ &\overset{(80)+(81)}{\leq} \|x_k - x_\star\|^2 - t_k^2. \end{aligned}$$

So, inequality *(iii)* holds in this case.

Case 2: $\lambda_k = 0$**.** Then (76) becomes $\nabla f(x_k) + n_k = 0$, which means that

$$0 \in \nabla f(x_k) + N_{\mathcal{X}}(x_{k+1}).$$

Therefore, x_{k+1} minimizes the linear function $z \mapsto \langle \nabla f(x_k), z\rangle$ over all of $\mathcal{X}$. In particular, it must be the case that $\langle \nabla f(x_k), x_{k+1} - x_\star\rangle \le 0$. Combined with (75), this yields

$$\langle \nabla f(x_k), x_{k+1} - x_\star\rangle = 0. \tag{82}$$

Likewise, from $\nabla f(x_k) + n_k = 0$ and the normal-cone inequality $\langle n_k, x_{k+1} - x_\star\rangle \ge 0$ (recall (78)), we also get $\langle n_k, x_{k+1} - x_\star\rangle = 0$. Further,

$$0 \overset{(82)}{=} \langle \nabla f(x_k), x_{k+1} - x_\star\rangle = \langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_\star\rangle + \langle \nabla f(x_\star), x_{k+1} - x_\star\rangle.$$

Both terms on the right are nonnegative by (73) and (74), so each must vanish. In particular,

$$\langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_\star\rangle = 0. \tag{83}$$

Expanding (83), using (8) and the Cauchy–Schwarz inequality, we arrive at

$$\begin{aligned} 0 &\overset{(83)}{=} \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle + \langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_k\rangle \\ &\overset{(8)}{\ge} \|\nabla f(x_k) - \nabla f(x_\star)\|\, t_k - \|\nabla f(x_k) - \nabla f(x_\star)\|\, \|x_{k+1} - x_k\|. \end{aligned}$$

Since we assume that $\nabla f(x_k) \ne \nabla f(x_\star)$, this implies $\|x_{k+1} - x_k\| \ge t_k$. However, since $x_{k+1} \in \mathcal{B}_k$, we also have $\|x_{k+1} - x_k\| \le t_k$. Therefore,

$$\|x_{k+1} - x_k\| = t_k, \tag{84}$$

and hence identity *(ii)* holds in this case also. Moreover, equality in Cauchy–Schwarz implies that $x_{k+1} - x_k$ is a multiple of $\nabla f(x_\star) - \nabla f(x_k)$. Since $\|x_{k+1} - x_k\| = t_k$, we must therefore have

$$x_{k+1} - x_k = -\, t_k\, \frac{\nabla f(x_k) - \nabla f(x_\star)}{\|\nabla f(x_k) - \nabla f(x_\star)\|}.$$

Since $\langle \nabla f(x_k) - \nabla f(x_\star), x_{k+1} - x_\star\rangle = 0$ (see (83)), it follows that

$$\langle x_k - x_{k+1}, x_{k+1} - x_\star\rangle = 0. \tag{85}$$

Therefore,

$$\begin{aligned} \|x_{k+1} - x_\star\|^2 &\overset{(52)}{=} \|x_k - x_\star\|^2 - 2\langle x_k - x_{k+1}, x_{k+1} - x_\star\rangle - \|x_k - x_{k+1}\|^2 \\ &\overset{(84)+(85)}{=} \|x_k - x_\star\|^2 - t_k^2. \end{aligned}$$

So, inequality *(iii)* holds in this case also. □

C.7 Setting and separation for practical schedules

Assumption C.1 (Setting for the practical schedules)**.** The set $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty, closed, and convex. The function f is convex and differentiable on an open convex set containing $\mathcal{X}$, and has a constrained minimizer $x_\star \in \mathcal{X}$. All norms in this appendix are Euclidean. Write $f_\star = f(x_\star)$, $\Delta_k = f(x_k) - f_\star$, and $R_0 = \|x_0 - x_\star\|$, where $x_0 \in \mathcal{X}$. For a specified positive radius t_k, use

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z\rangle. \tag{86}$$

The oracle may return any minimizer. The feasible local set is compact and nonempty, so the update exists.

The next lemma isolates the geometry needed for the two accuracy-based rules. Its hypothesis will hold until an accurate iterate has appeared; it need not hold afterwards.

Lemma C.4 (Distance decrease from a separating linear objective)**.** Let $\mathcal{X}$ be nonempty, closed, and convex, let $x, u \in \mathcal{X}$, $t > 0$, and

$$z \in \operatorname*{argmin}_{y \in \mathcal{X} \cap \mathcal{B}(x,t)} \langle g, y\rangle. \tag{87}$$

If $\langle g, u - z\rangle < 0$, then

$$\|z - x\| = t, \qquad \|z - u\|^2 \le \|x - u\|^2 - t^2. \tag{88}$$

Proof. If $\|z - x\| < t$, a sufficiently short segment from z toward u stays in the ball and in $\mathcal{X}$, and strictly decreases the linear objective, contradicting (87). Thus $\|z-x\| = t$. Define the closed convex set $H = \{y \in \mathcal{X} : \langle g, y - z\rangle \le 0\}$. It contains both u and z. No $y \in H$ can satisfy $\|y - x\| < t$: if its linear objective is strictly smaller than that of z, it contradicts (87) directly; if equal, a short move from y toward u gives that contradiction. Consequently z is a nearest point in H to x. For $s \in [0, 1]$, the point $z + s(u - z)$ belongs to H, so minimality gives

$$0 \le \|z + s(u - z) - x\|^2 - \|z - x\|^2 = 2s\langle z - x, u - z\rangle + s^2\|u - z\|^2.$$

Divide by $s > 0$ and let s decrease to zero. Hence $\langle x - z, u - z\rangle \le 0$, and the Euclidean norm identity yields

$$\|x - u\|^2 = \|x - z\|^2 + \|z - u\|^2 + 2\langle x - z, z - u\rangle \ge t^2 + \|z - u\|^2.$$

This proves (88). Projection is used only to interpret this proof, not as an algorithmic operation. □

C.8 Affine-subspace oracle and connection to gradient descent

Assume that the feasible set is an affine subspace of $\mathbb{R}^d$, i.e., $\mathcal{X} = a + \mathcal{V}$, where $a \in \mathbb{R}^d$ and $\mathcal{V} \subseteq \mathbb{R}^d$ is a linear subspace. We have the following result.

Proposition C.1 (Explicit form of BroxGD on an affine subspace)**.** Assume that $\mathcal{X} = a + \mathcal{V}$, where $a \in \mathbb{R}^d$ and $\mathcal{V} \subseteq \mathbb{R}^d$ is a linear subspace. Let $\mathrm{Proj}_{\mathcal{V}} : \mathbb{R}^d \to \mathcal{V}$ denote the orthogonal projector onto $\mathcal{V}$. Consider one step of Algorithm 1, i.e.,

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z\rangle, \tag{89}$$

where $x_k \in \mathcal{X}$ and $t_k > 0$. Impose the tie-breaking rule $x_{k+1} = x_k$ whenever $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) = 0$. Then

$$x_{k+1} = \begin{cases} x_k - t_k \dfrac{\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)}{\|\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)\|}, & \text{if } \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \neq 0, \\ x_k, & \text{if } \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) = 0. \end{cases}$$

In particular, if $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \neq 0$, then the step is unique.

Proof. Since $x_k \in \mathcal{X} = a + \mathcal{V}$, every point $z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)$ can be written uniquely in the form $z = x_k + s$, $s \in \mathcal{V}$, $\|s\| \le t_k$. Therefore the subproblem (89) can be written as

$$\min_{s \in \mathcal{V},\ \|s\| \le t_k} \langle \nabla f(x_k), x_k + s\rangle .$$

Since the term $\langle \nabla f(x_k), x_k\rangle$ does not depend on s, this is the same as

$$\min_{s \in \mathcal{V},\ \|s\| \le t_k} \langle \nabla f(x_k), s\rangle .$$

Now, for every $s \in \mathcal{V}$, we have

$$\langle \nabla f(x_k), s\rangle = \langle \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k), s\rangle ,$$

because $s \in \mathcal{V}$ and $\nabla f(x_k) - \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \in \mathcal{V}^\perp$. Hence the problem further reduces to

$$\min_{s \in \mathcal{V},\ \|s\| \le t_k} \langle \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k), s\rangle .$$

If $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) = 0$, then the objective is identically zero over the feasible set. Hence every $s \in \mathcal{V}$ satisfying $\|s\| \le t_k$ is optimal. Equivalently, every point in $x_k + \{s \in \mathcal{V} : \|s\| \le t_k\}$ is a solution of (89). The imposed tie-breaking rule selects $s = 0$, hence $x_{k+1} = x_k$. Assume now that $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \neq 0$. Then by the Cauchy–Schwarz inequality,

$$\langle \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k), s\rangle \ge -\|\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)\|\|s\| \ge -\|\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)\| t_k$$

for every feasible s. Equality holds if and only if

$$\|s\| = t_k \qquad \text{and} \qquad s = -t_k \frac{\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)}{\|\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)\|}.$$

Therefore,

$$x_{k+1} = x_k - t_k \frac{\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)}{\|\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)\|}.$$

□

The above proposition shows that, when $\mathcal{X}$ is an affine subspace, the method in Algorithm 1 coincides with normalized GD on the restriction of f to that affine subspace whenever the projected gradient is nonzero. At points where the projected gradient vanishes, the imposed tie-breaking rule gives $x_{k+1} = x_k$, agreeing with the stationary GD update. Indeed, consider the restricted problem

$$\min_{x \in a+\mathcal{V}} f(x).$$

The tangent space of the feasible set is $\mathcal{V}$, and the gradient of the restricted objective is precisely the projection of the ambient gradient onto this tangent space:

$$\nabla_{\mathcal{X}} f(x) = \mathrm{Proj}_{\mathcal{V}} \nabla f(x).$$

Therefore, whenever $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \neq 0$, the update from Proposition C.1 can be rewritten as

$$x_{k+1} = x_k - t_k \frac{\nabla_{\mathcal{X}} f(x_k)}{\|\nabla_{\mathcal{X}} f(x_k)\|}.$$

Thus the method moves in exactly the same direction as GD for minimizing f over the affine subspace $\mathcal{X}$, namely the negative projected gradient direction, but with step length prescribed directly by the radius t_k. For comparison, ordinary GD on the restricted problem would take the form

$$x_{k+1}^{\mathsf{GD}} = x_k - \gamma_k \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) = x_k - \gamma_k \nabla_{\mathcal{X}} f(x_k),$$

for some scalar stepsize $\gamma_k > 0$. Hence, whenever $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \neq 0$, the two methods differ only in how the step length is parameterized. In fact, the update from Algorithm 1 is exactly a GD step with effective stepsize

$$\gamma_k = \frac{t_k}{\|\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k)\|} = \frac{t_k}{\|\nabla_{\mathcal{X}} f(x_k)\|}.$$

This discussion can be summarized formally as follows.

Remark C.2 (Connection with GD on the affine subspace)**.** *When $\mathcal{X} = a + \mathcal{V}$ is an affine subspace, under the tie-breaking rule of Proposition C.1, Algorithm 1 coincides with normalized* GD *applied to the restriction of f to $\mathcal{X}$. More precisely, if $\mathrm{Proj}_{\mathcal{V}} \nabla f(x_k) \neq 0$, then*

$$x_{k+1} = x_k - t_k \frac{\nabla_{\mathcal{X}} f(x_k)}{\|\nabla_{\mathcal{X}} f(x_k)\|}, \qquad \nabla_{\mathcal{X}} f(x_k) = \mathrm{Proj}_{\mathcal{V}} \nabla f(x_k).$$

Equivalently, the same step can be written as an ordinary GD *step*

$$x_{k+1} = x_k - \gamma_k \nabla_{\mathcal{X}} f(x_k), \qquad \gamma_k = \frac{t_k}{\|\nabla_{\mathcal{X}} f(x_k)\|}.$$

Hence, in the affine subspace case, Algorithm 1 is exactly GD *on the restricted problem $f|_{\mathcal{X}}$, expressed in trust-region form.*

Finally, the unconstrained case appears as the special case $\mathcal{V} = \mathbb{R}^d$, in which $\mathrm{Proj}_{\mathcal{V}} = \mathrm{Id}$ is the identity mapping. When $\nabla f(x_k) \neq 0$, the formula reduces to

$$x_{k+1} = x_k - t_k \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|},$$

which is precisely normalized GD in the ambient space. When $\nabla f(x_k) = 0$, the tie-breaking rule keeps $x_{k+1} = x_k$.

D PROOFS OF THE CONVERGENCE GUARANTEES IN TABLE 1

The four subsections and their numbered entries follow the corresponding groups and rate rows of Table 1. Group B places the bounded-gradient, bounded-subgradient, and smooth convex Polyak results together, before the strongly convex distance benchmark. Shared geometric tools appear in Appendix C. Main-text theorems are restated with their original numbers; additional formulations and supporting proofs are placed beside the corresponding row.

D.1 A. PRACTICAL OBJECTIVE GUARANTEES

This subsection follows group A of Table 1: fixed schedules first, then radius backtracking. We restate the combined main-text horizon theorem before giving the two cases and their proofs in table order. The shared setting is Assumption C.1.

Theorem 3.2 (Practical fixed-horizon objective guarantees)**.** Let Assumption 3.1 hold, let f be convex, fix $x_\star \in \mathcal{X}_\star$, and let $R \geq R_0$ be a known positive upper bound on $R_0 := \|x_0 - x_\star\|$. Starting from x_0, run BroxGD for $K \geq 1$ iterations, and return the iterate x_K^{best} with the smallest objective value among $x_0, \ldots, x_K$.

(i) **Bounded gradients.** If f has bounded gradients, i.e. $\|\nabla f(x)\| \leq G$ for all $x \in \mathcal{X}$, and we use the radius rule

$$t_k = \frac{GR}{\sqrt{K}\,\|\nabla f(x_k)\|}, \tag{12}$$

then

$$f(x_K^{\mathrm{best}}) - f_\star \leq \frac{GR}{\sqrt{K}}. \tag{13}$$

If $\nabla f(x_k) = 0$ for some k, we stop and return x_k; the same guarantee holds.

(ii) **Smooth convex objectives.** If f has an L-Lipschitz gradient on an open convex neighborhood of $\mathcal{X}$, and we use the radius rule

$$t_k = \frac{R}{\sqrt{K}}, \tag{14}$$

then

$$f(x_K^{\mathrm{best}}) - f_\star \leq \frac{LR^2}{2K}. \tag{15}$$

D.1.1 FIXED HORIZON WITH BOUNDED GRADIENTS

Corollary D.1 (Fixed-horizon bounded-gradient schedule)**.** Under the assumptions of Theorem D.1, let $R \geq R_0$ be known, $R > 0$, and let $K \geq 1$ be the number of oracle calls. Set

$$\varepsilon = \frac{GR}{\sqrt{K}}, \qquad t_k = \frac{GR}{\sqrt{K}\,\|\nabla f(x_k)\|}. \tag{90}$$

Return an iterate x_K^{best} of smallest objective value among $x_0, \ldots, x_K$, or the point at which a zero gradient is detected. Then

$$f(x_K^{\mathrm{best}}) - f_\star \leq \frac{GR}{\sqrt{K}}. \tag{91}$$

We establish the target-accuracy result below, then derive this fixed-horizon guarantee.

Theorem D.1 (Target accuracy without the optimal value)**.** Let Assumption C.1 hold and suppose

$$\|\nabla f(x)\| \le G \qquad (x \in \mathcal{X}), \qquad G > 0. \tag{92}$$

Given $\varepsilon > 0$, use

$$t_k = \frac{\varepsilon}{\|\nabla f(x_k)\|} \tag{93}$$

whenever the gradient is nonzero; otherwise stop and return x_k, which is optimal. At every iteration with $\Delta_k > \varepsilon$,

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - \frac{\varepsilon^2}{G^2}. \tag{94}$$

For every integer $K > G^2 R_0^2/\varepsilon^2$, the best of $x_0, \dots, x_{K-1}$ has gap at most ε. With the endpoint included, the same conclusion holds for every integer $K \ge G^2 R_0^2/\varepsilon^2$:

$$\min_{0 \le k \le K} \Delta_k \le \varepsilon. \tag{95}$$

Early termination satisfies both conclusions.

Proof. Let $g_k = \nabla f(x_k)$. A zero gradient certifies optimality by convexity. Otherwise, convexity, Cauchy–Schwarz, and feasibility give

$$\begin{aligned}\langle g_k, x_\star - x_{k+1}\rangle &= \langle g_k, x_\star - x_k\rangle + \langle g_k, x_k - x_{k+1}\rangle \\ &\overset{(86)}{\le} -\Delta_k + \|g_k\| t_k \overset{(93)}{=} -\Delta_k + \varepsilon.\end{aligned}$$

If $\Delta_k > \varepsilon$, this expression is negative, so Lemma C.4 applies with $u = x_\star$:

$$\|x_{k+1} - x_\star\|^2 \overset{(88)}{\le} \|x_k - x_\star\|^2 - t_k^2 \overset{(93),(92)}{\le} \|x_k - x_\star\|^2 - \varepsilon^2/G^2.$$

If all queried iterates $x_0, \dots, x_{K-1}$ have gap greater than ε, summing gives

$$0 \le \|x_K - x_\star\|^2 \overset{(94)}{\le} R_0^2 - K\varepsilon^2/G^2. \tag{96}$$

For $K > G^2 R_0^2/\varepsilon^2$, this is impossible. At equality, it forces $x_K = x_\star$, which proves the endpoint-inclusive assertion as well. This handles the distinction between strict and non-strict iteration budgets. □

Proof of Corollary D.1. The choice of ε gives

$$\frac{G^2 R_0^2}{\varepsilon^2} \overset{(90)}{=} K \frac{R_0^2}{R^2} \le K.$$

Apply the endpoint-inclusive part of Theorem D.1; the resulting bound is

$$f(x_K^{\mathrm{best}}) - f_\star \overset{(95)}{\le} \varepsilon \overset{(90)}{=} GR/\sqrt{K}.$$

If $R = 0$ is certified, $x_0 = x_\star$ and no call is needed. □

To run the target-accuracy rule one needs only ε and the observed gradient norm. A predetermined budget additionally needs G and a bound on R_0; for example, $\lfloor G^2 R^2/\varepsilon^2 \rfloor + 1$ calls suffice even when only the queried points are retained. These statements do not assert last-iterate accuracy after the target has first been reached, or convergence to arbitrary precision for a single fixed positive ε.

D.1.2 Fixed horizon for smooth convex objectives

Corollary D.2 (Fixed-horizon smooth convex schedule)**.** Under the assumptions of Theorem D.2, let $R \ge R_0$, $R > 0$, and $K \ge 1$. Use $t_k = R/\sqrt{K}$ for K calls and return the best of $x_0, \dots, x_K$. Then

$$f(x_K^{\mathrm{best}}) - f_\star \le \frac{LR^2}{2K}. \tag{97}$$

We establish the target-accuracy result below, then derive this fixed-horizon guarantee.

Theorem D.2 (Constant radius and objective accuracy)**.** Let Assumption C.1 hold. Suppose f has an L-Lipschitz gradient on its open convex domain, with known $L > 0$. Fix $\varepsilon > 0$ and use

$$t_k \equiv t = \sqrt{\frac{2\varepsilon}{L}}. \tag{98}$$

For every integer $K \geq LR_0^2/(2\varepsilon)$, the exact full-step iterates satisfy

$$\min_{0 \leq k \leq K} \{f(x_k) - f_\star\} \leq \varepsilon. \tag{99}$$

Equivalently, a radius $t = \sqrt{2/L}\, c$ with $c > 0$ achieves best-iterate gap at most c^2 after any integer $K \geq LR_0^2/(2c^2)$.

Proof. If $\Delta_0 \leq \varepsilon$ there is nothing to prove. At any subsequent step, smoothness and convexity imply

$$\begin{aligned}\langle \nabla f(x_k), x_\star - x_{k+1} \rangle &\leq f_\star - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\ &\leq -\Delta_{k+1} + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\overset{(86),(98)}{\leq} -\Delta_{k+1} + \varepsilon. \end{aligned} \tag{100}$$

If $\Delta_{k+1} > \varepsilon$, (100) is strictly negative, so Lemma C.4 gives

$$\|x_{k+1} - x_\star\|^2 \overset{(88)}{\leq} \|x_k - x_\star\|^2 - t^2. \tag{101}$$

Suppose, for contradiction, that every iterate $x_0, \ldots, x_K$ has gap greater than ε. Summing over all K steps yields

$$0 \leq \|x_K - x_\star\|^2 \overset{(101)}{\leq} R_0^2 - Kt^2 \overset{(98)}{=} R_0^2 - 2K\varepsilon/L \leq 0.$$

If the right side is negative there is a contradiction; if it is zero, $x_K = x_\star$, again contradicting the assumed gap. The case $K = 0$ permitted by the budget has $R_0 = 0$ and is immediate. This proves (99). Substituting $\varepsilon = c^2$ proves the equivalent formulation. □

Proof of Corollary D.2. Set $\varepsilon = LR^2/(2K)$. Then

$$\sqrt{2\varepsilon/L} = R/\sqrt{K}, \qquad LR_0^2/(2\varepsilon) = KR_0^2/R^2 \leq K.$$

The hypotheses and budget of Theorem D.2 are satisfied, and

$$f(x_K^{\mathrm{best}}) - f_\star \overset{(99)}{\leq} \varepsilon = LR^2/(2K).$$

□

The target-accuracy radius needs only L and ε to run; a certified stopping budget additionally needs a bound on R_0. In the fixed-horizon formulation, the radius itself uses R and K, while L enters only the accuracy guarantee. No strict convexity is needed. The constant-radius conclusion controls objective values, rather than only the gradient-difference residual. The distance argument is Euclidean, and no arbitrary-norm extension is claimed here. Both accuracy-based schedules allow objective increases after a good point is reached, which is why the output is the best observed iterate.

D.1.3 PREDETERMINED GEOMETRIC RADII

Theorem D.3 (Geometric radii from certified bounds)**.** Let Assumption C.1 hold. Suppose f is L-smooth and μ-strongly convex on $\mathcal{X}$, with known valid bounds $0 < \mu \le L$; conservative upper and lower bounds may be used as L and μ, respectively. Let $C_0 > 0$ satisfy $\Delta_0 \le C_0$. Set

$$\alpha = \frac{\mu}{2L}, \qquad q = 1 - \frac{\mu}{4L}, \qquad t_k = \alpha\sqrt{\frac{2C_0}{\mu}}\, q^{k/2}. \tag{102}$$

Then the iterates of (86) satisfy, for every $k \ge 0$,

$$\Delta_k \le C_0 q^k, \qquad \|x_k - x_\star\|^2 \le \frac{2C_0}{\mu} q^k. \tag{103}$$

Thus, for $0 < \varepsilon < C_0$, it suffices to take

$$K \ge \left\lceil \frac{\log(C_0/\varepsilon)}{-\log q} \right\rceil \tag{104}$$

to ensure $f(x_K) - f_\star \le \varepsilon$.

Proof. Constrained first-order optimality gives $\langle \nabla f(x_\star), x - x_\star \rangle \ge 0$ for $x \in \mathcal{X}$. Strong convexity therefore implies

$$f(x) - f_\star \ge \frac{\mu}{2}\|x - x_\star\|^2 \qquad (x \in \mathcal{X}). \tag{105}$$

We prove the objective bound by induction. It holds at $k = 0$. Assuming $\Delta_k \le C_0 q^k$, define $y_k = (1-\alpha)x_k + \alpha x_\star$. Since $0 < \alpha \le 1/2$, this point belongs to $\mathcal{X}$, and

$$\|y_k - x_k\| = \alpha\|x_k - x_\star\| \overset{(105)}{\le} \alpha\sqrt{2\Delta_k/\mu} \le \alpha\sqrt{2C_0 q^k/\mu} \overset{(102)}{=} t_k.$$

The comparator y_k is thus feasible for the local oracle. By oracle optimality followed by convexity,

$$\langle \nabla f(x_k), x_{k+1} - x_k \rangle \overset{(86)}{\le} \alpha\langle \nabla f(x_k), x_\star - x_k \rangle \le -\alpha\Delta_k. \tag{106}$$

Smoothness and feasibility of the update give

$$\begin{aligned}
\Delta_{k+1} &\le \Delta_k + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\overset{(106),(86)}{\le} (1-\alpha)\Delta_k + \frac{L}{2}t_k^2 \\
&\overset{(102)}{\le} \left(1 - \frac{\mu}{2L} + \frac{\mu}{4L}\right) C_0 q^k \overset{(102)}{=} C_0 q^{k+1}.
\end{aligned}$$

The penultimate step also uses the induction hypothesis and $1 - \alpha \ge 0$. This closes the induction. The distance bound follows from (105). Finally,

$$\Delta_K \overset{(103)}{\le} C_0 q^K \overset{(104)}{\le} \varepsilon.$$

□

A known lower bound $\ell \le f_\star$ gives the valid initial bound $C_0 = f(x_0) - \ell$ whenever this is positive. If it is zero, x_0 is already optimal and no step is needed. For a nonnegative loss one can use $\ell = 0$. The schedule needs L, μ, C_0, but no current optimality gap or distance; its guaranteed contraction is weaker than that of the idealized radius in Theorem 3.4. Even at an optimal iterate an arbitrary exact oracle may move; the induction above still applies.

Shared ingredients for the backtracking guarantees The next two table rows share one algorithm and one certificate argument. We restate Theorem 3.3, establish those ingredients once, then prove its convex and strongly convex cases consecutively.

Theorem 3.3 (Practical backtracking and last-iterate rates)**.** Let Assumption 3.1 hold and suppose f is convex with a globally L-Lipschitz gradient, where $L > 0$ is known. Fix $\xi \in (0, 1)$. Then Algorithm 2 makes finitely many local LMO calls at each ball center x_k and either stops at a minimizer or accepts an objective-decreasing step. Freeze the sequence after optimal termination, whether the stopping test is a zero gradient or $\mathcal{D} = 0$.

If, for some $R_{\mathrm{lev}} > 0$, the accepted centers satisfy

$$\operatorname{dist}(x_k, \mathcal{X}_\star) \le R_{\mathrm{lev}} \qquad (k \ge 0), \tag{16}$$

then, for every $K \ge 0$,

$$\Delta_K \le \frac{\max\{\Delta_0, 4LR_{\mathrm{lev}}^2/\xi^2\}}{K+1}. \tag{17}$$

If instead f is μ-strongly convex on $\mathcal{X}$, $\mu > 0$, then without (16),

$$\Delta_K \le \left(1 + \frac{\xi^2\mu}{L}\right)^{-K} \Delta_0. \tag{18}$$

Unlike the fixed schedules above, backtracking selects its radius from an observable test. Here we instead choose the radius by an observable acceptance test. The method needs only a known smoothness bound and an exact local linear oracle. In particular, neither an initial objective gap nor a solution-distance bound is an algorithm input. The guarantee concerns the last accepted iterate, but a single accepted step may require several local-oracle calls.

Assumption D.1 (Setting for radius backtracking)**.** The set $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty, closed, and convex. The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, its minimum over $\mathcal{X}$ is attained, and a known constant $L > 0$ satisfies

$$\|\nabla f(u) - \nabla f(v)\| \le L\|u - v\| \qquad (u, v \in \mathbb{R}^d). \tag{107}$$

Start at $x_0 \in \mathcal{X}$ and use an exact local linear minimization oracle. Write $\mathcal{X}_\star = \operatorname{argmin}_{x\in\mathcal{X}} f(x)$ and $\Delta_k = f(x_k) - f_\star$.

Radius policy. Fix an input contraction factor $\xi \in (0, 1)$. At x_k, compute $a_k = \nabla f(x_k)$ and stop if $a_k = 0$. Otherwise initialize $t_{k,0} = \|a_k\|/L$. For $j = 0, 1, \ldots$, compute

$$\begin{aligned} t_{k,j} &= \xi^j \frac{\|a_k\|}{L}, \qquad y_{k,j} \in \operatorname*{argmin}_{z\in\mathcal{X}\cap\mathcal{B}(x_k,t_{k,j})} \langle a_k, z - x_k\rangle, \\ s_{k,j} &= y_{k,j} - x_k, \qquad \mathcal{D}_{k,j} = -\langle a_k, s_{k,j}\rangle \ge 0. \end{aligned} \tag{108}$$

A zero value of $\mathcal{D}_{k,j}$ certifies optimality of x_k. Otherwise accept the first trial satisfying

$$\mathcal{D}_{k,j} \ge L\|s_{k,j}\|^2, \tag{109}$$

and set $x_{k+1} = y_{k,j}$. If the test fails, multiply the radius by ξ; halving corresponds to $\xi = 1/2$. The test uses the attained step length, which need not equal the trial radius. It requires no projection, objective evaluation, gradient at the trial point, or multiplier estimate.

The pseudocode is given in Algorithm 2 in the main text. The certificates and rate proofs below establish Theorem 3.3.

We first prove the endpoint estimate that keeps the constants sharp. It is a consequence of convex smoothness, independent of the local oracle.

Lemma D.1 (Smooth endpoint estimate)**.** Under Assumption D.1, for $s = y - x$, $v = \nabla f(y) - \nabla f(x)$, and $\tau \ge 0$,

$$\|v - \tau s\| \le \max\{\tau, L - \tau\}\|s\| \le \max\{L, \tau\}\|s\|. \tag{110}$$

Proof. **Step 1: Derive cocoercivity.** Convexity and (107) imply, respectively,

$$\begin{aligned} f(u) &\ge f(w) + \langle \nabla f(w), u - w\rangle, \\ f(u) &\le f(w) + \langle \nabla f(w), u - w\rangle + \frac{L}{2}\|u - w\|^2 \qquad (u, w \in \mathbb{R}^d). \end{aligned} \tag{111}$$

For completeness, the second inequality follows by integrating the gradient along the segment from w to u and applying (107). Define $h(z) = f(z) - \langle \nabla f(x), z\rangle$. Its gradient vanishes at x, so x is a global minimizer. Applying the upper model to h at y gives

$$h(x) \le h(y - v/L) \overset{(111)}{\le} h(y) - \frac{\|v\|^2}{2L}.$$

Thus $f(y) - f(x) - \langle \nabla f(x), s\rangle \ge \|v\|^2/(2L)$. Interchanging x, y and adding gives

$$\langle v, s\rangle \ge \frac{\|v\|^2}{L}. \tag{112}$$

Step 2: Center and shift the gradient difference. Expanding the square yields

$$\left\| v - \frac{L}{2} s \right\|^2 = \|v\|^2 - L\langle v, s\rangle + \frac{L^2}{4}\|s\|^2 \overset{(112)}{\le} \frac{L^2}{4}\|s\|^2. \tag{113}$$

The triangle inequality therefore gives

$$\begin{aligned} \|v - \tau s\| &\le \left\|v - \tfrac{L}{2}s\right\| + \left|\tfrac{L}{2} - \tau\right| \|s\| \\ &\overset{(113)}{\le} \left(\tfrac{L}{2} + \left|\tfrac{L}{2} - \tau\right|\right) \|s\| \\ &= \max\{\tau, L - \tau\}\|s\| \le \max\{L, \tau\}\|s\|. \end{aligned}$$

The identity follows by considering $\tau \le L/2$ and $\tau \ge L/2$. This also covers $s = 0$. □

Theorem D.4 (Finite backtracking and accepted-step certificates)**.** Under Assumption D.1, fix $\xi \in (0, 1)$ in Algorithm 2. The algorithm uses finitely many oracle calls at each center and either stops at a minimizer or accepts a step. Define $F = f + \iota_{\mathcal{X}}$, where $\iota_{\mathcal{X}}$ is the convex indicator. Each accepted step $s_k = x_{k+1} - x_k$ admits $g_{k+1} \in \partial F(x_{k+1})$ and an analysis multiplier λ_k such that

$$L \le \lambda_k \le L/\xi^2, \qquad f(x_k) - f(x_{k+1}) \ge \frac{\lambda_k}{2}\|s_k\|^2, \qquad \|g_{k+1}\| \le \lambda_k \|s_k\|. \tag{114}$$

Neither g_{k+1} nor λ_k is needed to run the algorithm.

Proof. **Step 1: Justify stopping and finite backtracking.** Fix a center x, set $a = \nabla f(x)$, and write $y, s, t, \mathcal{D}$ for a trial. Compactness of $\mathcal{X} \cap \mathcal{B}(x, t)$ guarantees an oracle solution. If $a = 0$, convexity certifies optimality. If $\mathcal{D} = 0$, then $\langle a, z - x\rangle \ge 0$ for all locally feasible z. For any $u \in \mathcal{X} \setminus \{x\}$, choose $\theta = \min\{1, t/\|u - x\|\} > 0$. The point $z = x + \theta(u - x)$ is locally feasible. Dividing $\langle a, z - x\rangle \ge 0$ by θ gives $\langle a, u - x\rangle \ge 0$. This also holds at $u = x$, so convexity proves optimality over $\mathcal{X}$.

At a nonoptimal center choose $x_\star \in \mathcal{X}_\star$ and set $d = \|x - x_\star\| > 0$, $\Delta = f(x) - f_\star > 0$. For $0 < t \le d$, the feasible segment comparator $x + (t/d)(x_\star - x)$ gives

$$\mathcal{D} \overset{(108)}{\ge} \frac{t}{d}\langle a, x - x_\star\rangle \overset{(111)}{\ge} \frac{t\Delta}{d}. \tag{115}$$

If also $t \le \Delta/(Ld)$, then

$$\mathcal{D} \overset{(115)}{\ge} \frac{t\Delta}{d} \ge Lt^2 \overset{(108)}{\ge} L\|s\|^2.$$

Since $\xi^j \|a\|/L \to 0$, repeated contraction reaches this positive threshold in finitely many trials. The unknown d, Δ occur only in this argument.

Step 2: Identify the normal vector and ball multiplier. At an accepted nonterminal trial, $\mathcal{D} > 0$, hence $r = \|s\| > 0$. The ball indicator is continuous at $x \in \mathcal{X}$, since $t > 0$. The convex normal-cone sum rule therefore yields $n \in N_{\mathcal{X}}(y)$ and $\tau \ge 0$ satisfying

$$a + n + \tau s = 0, \qquad \tau(r^2 - t^2) = 0. \tag{116}$$

Here $N_{\mathcal{X}}(y) = \{n : \langle n, z-y\rangle \le 0 \text{ for all } z \in \mathcal{X}\}$. Taking $z = x$ shows $\langle n, s\rangle \ge 0$, and consequently

$$\mathcal{D} \overset{(108),(116)}{=} \tau r^2 + \langle n, s\rangle \ge \tau r^2. \tag{117}$$

Set

$$\lambda = \max\{L, \tau\}. \tag{118}$$

This analysis multiplier can differ from the oracle's ball multiplier.

Step 3: Establish the two certificates. Acceptance and (117) give $\mathcal{D} \ge \lambda r^2$. Thus

$$\begin{aligned} f(x) - f(y) &\overset{(111),(108)}{\ge} \mathcal{D} - \tfrac{L}{2} r^2 \\ &\overset{(109),(117),(118)}{\ge} (\lambda - L/2) r^2 \overset{(118)}{\ge} \tfrac{\lambda}{2} r^2. \end{aligned}$$

With $v = \nabla f(y) - \nabla f(x)$, select

$$g = \nabla f(y) + n \overset{(116)}{=} v - \tau s \in \partial F(y). \tag{119}$$

The endpoint lemma now gives

$$\|g\| \overset{(119)}{=} \|v - \tau s\| \overset{(110)}{\le} \max\{L, \tau\} r \overset{(118)}{=} \lambda r.$$

Step 4: Bound the analysis multiplier. If $\tau = 0$, then $\lambda = L$. If the initial radius is accepted and $\tau > 0$, complementary slackness gives $r = t = \|a\|/L$. Hence

$$\tau r^2 \overset{(117)}{\le} \mathcal{D} \overset{(108)}{\le} \|a\| r, \qquad \tau \le \frac{\|a\|}{r} = L.$$

If instead a previous trial was rejected, its radius was t/ξ. Denote its decrease and step by $\widetilde{\mathcal{D}}, \widetilde{s}$. Nesting of the feasible balls and rejection imply

$$\mathcal{D} \le \widetilde{\mathcal{D}} \overset{(109)}{<} L\|\widetilde{s}\|^2 \overset{(108)}{\le} L t^2/\xi^2. \tag{120}$$

For $\tau > 0$, (116) gives $r = t > 0$. Combining (117) with (120) and dividing by t^2 gives $\tau < L/\xi^2$. All cases yield $L \le \lambda \le L/\xi^2$, completing the proof. □

D.1.4 Convex objective convergence with backtracking

Theorem D.5 (Convex last-iterate objective rate)**.** Under Assumption D.1, use Algorithm 2 with fixed $\xi \in (0, 1)$ and freeze the sequence after optimal termination. For the convex rate assume that, for some $R_{\mathrm{lev}} > 0$, the accepted centers satisfy

$$\operatorname{dist}(x_k, \mathcal{X}_\star) \le R_{\mathrm{lev}} \qquad (k \ge 0). \tag{121}$$

Then, for every $K \ge 0$,

$$\Delta_K \le \frac{\max\{\Delta_0, 4LR_{\mathrm{lev}}^2/\xi^2\}}{K+1}. \tag{122}$$

The algorithm does not need R_{lev} or Δ_0. The index K counts accepted steps.

Proof. **Step 1: Eliminate the step length.** The certificates and $\lambda_k \le L/\xi^2$ imply

$$\Delta_k - \Delta_{k+1} \overset{(114)}{\ge} \frac{\|g_{k+1}\|^2}{2\lambda_k} \overset{(114)}{\ge} \frac{\xi^2 \|g_{k+1}\|^2}{2L}. \tag{123}$$

Step 2: Obtain the convex recurrence. The solution set is nonempty, closed, and convex. Choose its nearest point z to x_{k+1}. The subgradient inequality and Cauchy–Schwarz give

$$\Delta_{k+1} \le \langle g_{k+1}, x_{k+1} - z\rangle \overset{(121)}{\le} R_{\mathrm{lev}} \|g_{k+1}\|.$$

Using (123), define $a = \xi^2/(2LR_{\text{lev}}^2) > 0$ to obtain

$$\Delta_{k+1} + a\Delta_{k+1}^2 \le \Delta_k. \tag{124}$$

Step 3: Solve the recurrence explicitly. Set $A = \max\{\Delta_0, 2/a\}$ and $w_k = A/(k+1)$. Since $A > 0$,

$$\frac{w_k - w_{k+1}}{w_{k+1}^2} = \frac{k+2}{A(k+1)} \le \frac{2}{A} \le a.$$

Thus $w_k \le w_{k+1} + aw_{k+1}^2$. The function $H(u) = u + au^2$ is increasing for $u \ge 0$. Starting from $\Delta_0 \le w_0$, induction gives

$$H(\Delta_{k+1}) \overset{(124)}{\le} \Delta_k \le w_k \le H(w_{k+1}),$$

so $\Delta_{k+1} \le w_{k+1}$. Since $2/a = 4LR_{\text{lev}}^2/\xi^2$, this proves (122).

□

D.1.5 Strongly convex objective convergence with backtracking

Theorem D.6 (Strongly convex last-iterate objective rate)**.** Under Assumption D.1, use Algorithm 2 with fixed $\xi \in (0,1)$ and freeze the sequence after optimal termination. Suppose f is μ-strongly convex on $\mathcal{X}$, meaning

$$f(u) \ge f(y) + \langle \nabla f(y), u - y\rangle + \frac{\mu}{2}\|u - y\|^2 \qquad (u, y \in \mathcal{X}), \quad \mu > 0. \tag{125}$$

Without (121),

$$\Delta_K \le \left(1 + \frac{\xi^2\mu}{L}\right)^{-K} \Delta_0. \tag{126}$$

The algorithm needs neither R_{lev} nor μ nor Δ_0. The index K counts accepted steps, not all trial oracle calls.

Proof. **Step 1: Use strong convexity instead of a distance bound.** For $y = x_{k+1}$ and $g = g_{k+1} = \nabla f(y) + n$, normality gives $\langle n, x_\star - y\rangle \le 0$. Thus

$$\begin{aligned} f_\star &\overset{(125)}{\ge} f(y) + \langle g, x_\star - y\rangle + \tfrac{\mu}{2}\|x_\star - y\|^2 \\ &\ge f(y) - \tfrac{\|g\|^2}{2\mu}. \end{aligned}$$

The last step completes the square, so $\|g_{k+1}\|^2 \ge 2\mu\Delta_{k+1}$. Substituting into (123) yields

$$\Delta_k - \Delta_{k+1} \overset{(123)}{\ge} \frac{\xi^2\mu}{L}\Delta_{k+1}, \qquad \Delta_{k+1} \le \left(1 + \frac{\xi^2\mu}{L}\right)^{-1} \Delta_k.$$

Multiplying the one-step factors proves (126). All inequalities remain valid for a frozen optimal continuation, whose objective gaps are zero. □

Interpretation and oracle cost. The distance assumption in (121) is an analysis condition, not an input. For example, it holds if the initial sublevel set has bounded distance to the solution set, because accepted steps decrease the objective. It must not be silently replaced by the initial distance R_0: the proof above establishes objective descent, not Fejér monotonicity. The strong-convexity rate needs no such distance assumption and no knowledge of μ to choose radii. Each outer iteration reuses one gradient, while every rejected trial incurs another exact local-oracle solve. Finite backtracking alone does not assert a uniform bound on their total number. Approximate solves require separate error analysis. In the unconstrained case, the initial trial is the gradient step with stepsize $1/L$ and passes the test with equality.

D.2 B. SOLUTION-DEPENDENT CONVEX BENCHMARKS

As in group B of Table 1, we first give Polyak-radius objective guarantees for bounded gradients, bounded subgradients, and smooth convex objectives, then the sharp distance-contraction benchmark. These radii use solution-dependent quantities.

D.2.1 BOUNDED-GRADIENT POLYAK RADIUS

Theorem D.7 (Rate under bounded gradients)**.** Let Assumption 3.1 hold, let f be convex, and assume that there exists $G > 0$ such that

$$\|\nabla f(x)\| \le G \qquad \forall x \in \mathcal{X}. \tag{127}$$

Let $\{x_k\}_{k\ge 0}$ be the iterates generated by BroxGD, where $x_0 \in \mathcal{X}$. Further, let $x_\star \in \mathcal{X}_\star$, and choose the "Polyak radius" whenever $f(x_k) > f(x_\star)$:

$$t_k := \frac{f(x_k) - f(x_\star)}{\|\nabla f(x_k)\|}. \tag{128}$$

If $f(x_k) = f(x_\star)$, set $t_k = 0$ and freeze the sequence at x_k. This includes every constrained minimizer, not only the selected $x_\star$. Then for every $K \ge 1$,

$$\frac{1}{K}\sum_{k=0}^{K-1}(f(x_k) - f(x_\star))^2 \le \frac{G^2\|x_0 - x_\star\|^2}{K}. \tag{129}$$

Moreover, the average iterate $\bar{x}_K := \frac{1}{K}\sum_{k=0}^{K-1} x_k$ satisfies

$$f(\bar{x}_K) - f(x_\star) \le \frac{G\|x_0 - x_\star\|}{\sqrt{K}}. \tag{130}$$

Proof. Write $\Delta_k = f(x_k) - f(x_\star) \ge 0$. If $\Delta_k = 0$, the frozen step has $t_k = 0$ and $x_{k+1} = x_k$. If $\Delta_k > 0$, convexity implies $\nabla f(x_k) \ne 0$ and

$$0 < t_k \overset{(128)}{=} \frac{\Delta_k}{\|\nabla f(x_k)\|} \le \frac{\langle \nabla f(x_k), x_k - x_\star\rangle}{\|\nabla f(x_k)\|}.$$

Thus the Type-I condition of Theorem 3.1 holds at every positive-gap iterate. At such an iterate,

$$\begin{aligned}\|x_{k+1} - x_\star\|^2 &\overset{(10)}{\le} \|x_k - x_\star\|^2 - t_k^2 \\ &\overset{(128),(127)}{\le} \|x_k - x_\star\|^2 - \Delta_k^2/G^2.\end{aligned}$$

The resulting inequality also holds with equality at a frozen zero-gap iterate. Summing over $k = 0, \dots, K-1$ and dropping the nonnegative final squared distance gives

$$\sum_{k=0}^{K-1}\Delta_k^2 \le G^2\|x_0 - x_\star\|^2. \tag{131}$$

Dividing by K proves (129). Convexity and Cauchy–Schwarz then give

$$0 \le f(\bar{x}_K) - f(x_\star) \le \frac{1}{K}\sum_{k=0}^{K-1}\Delta_k \le \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k^2} \overset{(131)}{\le} \frac{G\|x_0 - x_\star\|}{\sqrt{K}},$$

which proves (130), including trajectories that reach any minimizer early. □

Theorem D.7 establishes an $\mathcal{O}(1/\sqrt{K})$ convergence rate in function value suboptimality. The adaptive radius in (128) can be interpreted as a normalized Polyak-type step (Polyak, 1987), which reduces to the classical Polyak stepsize in the unconstrained setting (the standard Polyak rule $(f(x_k)-f(x_\star))/\|\nabla f(x_k)\|^2$ corresponds exactly to (128) when $\mathcal{X} = \mathbb{R}^d$ (Gruntkowska et al., 2025, Remark F.5)). The bound (130) matches the rate achieved by ProjGD with Polyak stepsizes (Beck, 2017,

Theorem 8.13). Bounded gradients do not ensure finite classical FW curvature (Example B.1), so they do not imply the standard curvature-based objective rate. Infinite curvature does not preclude FW convergence: for example, uniformly continuous gradients on compact convex domains suffice with suitable stepsizes (Xu, 2017). This bounded gradient regime is also the most relevant one for modern deep learning, where global smoothness or strong convexity assumptions are rarely realistic.

D.2.2 Bounded-subgradient Polyak radius

Assumption 3.1 requires the function f to be convex and differentiable. Under this assumption, we then consider three settings: (i) smooth objectives, (ii) strongly convex and smooth objectives, and (iii) objectives with bounded gradients. The first two explicitly require smoothness, and hence differentiability of f.

In this section, we focus on the last setting and remove the differentiability assumption. Direct substitution of subgradients into vanilla FW can fail to converge in this setting (Nesterov, 2017, Example 1). Nesterov's counter-example is simple:

$$d = 2, \quad f(u, v) = \max\{u, v\}, \quad \mathcal{X} = \{(u, v) \ : \ u^2 + v^2 \leq 1\}.$$

See also (Yurtsever et al., 2018, Appendix 3).

We make a slight modification to Algorithm 1. At iteration k, if $0 \in \partial f(x_k)$, then x_k is already a minimizer of f, and the algorithm terminates. Otherwise, the method selects a subgradient $g_k \in \partial f(x_k)$ and defines the next iterate as

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle g_k, z \rangle . \tag{132}$$

We now show that the bounded gradient analysis extends directly to this non-smooth setting.

Proposition D.1 (Well-posedness in the non-smooth case)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. For every $x_k \in \mathcal{X}$, every $t_k \geq 0$, and every choice of subgradient $g_k \in \partial f(x_k)$, the set $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$ is nonempty, closed, and bounded. Consequently, the optimization problem

$$\min_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle g_k, z \rangle$$

admits at least one solution.

Proof. The proof is identical to that of Theorem C.1. □

Proposition D.2 (Type-I admissibility with subgradients)**.** Let $x_k \in \mathcal{X}$ and $x_\star \in \mathcal{X}_\star$. Assume that

$$g_k \in \partial f(x_k), \qquad 0 \notin \partial f(x_k), \qquad 0 < t_k \leq \frac{\langle g_k, x_k - x_\star \rangle}{\|g_k\|}.$$

Define

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle g_k, z \rangle .$$

Then the conclusions of the Type-I part of Theorem 3.1 remain valid, namely,

$$\|x_k - x_\star\| \geq t_k, \qquad \|x_{k+1} - x_k\| = t_k, \qquad \|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2.$$

Proof. The proof is the same as that of the Type-I admissibility part in Theorem 3.1, after replacing $\nabla f(x_k)$ by g_k. Indeed, the argument uses only the admissibility condition, the optimality conditions of the local linear minimization subproblem, and the geometry of the ball constraint. No differentiability of f is needed. □

Theorem D.8 (Rate under bounded subgradients)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, and let $\{x_k\}_{k\geq 0}$ be generated by the modified BroxGD method in (132), with $x_0 \in \mathcal{X}$. Assume that the subgradients of f are bounded on $\mathcal{X}$, i.e.,

$$\|g\| \leq G \qquad \forall x \in \mathcal{X},\ \forall g \in \partial f(x),$$

for some constant $G > 0$. Let $x_\star \in \mathcal{X}_\star$, and assume that whenever $0 \notin \partial f(x_k)$, the radius is chosen as

$$t_k := \frac{f(x_k) - f(x_\star)}{\|g_k\|}, \qquad g_k \in \partial f(x_k).$$

At every zero-gap iterate, set $t_k = 0$ and freeze all subsequent iterates. In particular, this applies if $0 \in \partial f(x_k)$. Thus the averages are defined for every integer $K \geq 1$. Then, for every $K \geq 1$,

$$\frac{1}{K}\sum_{k=0}^{K-1} (f(x_k) - f(x_\star))^2 \leq \frac{G^2\|x_0 - x_\star\|^2}{K}. \tag{133}$$

Moreover, the average iterate $\bar{x}_K := \frac{1}{K}\sum_{k=0}^{K-1} x_k$ satisfies

$$f(\bar{x}_K) - f(x_\star) \leq \frac{G\|x_0 - x_\star\|}{\sqrt{K}}. \tag{134}$$

Proof. Fix $k \geq 0$. If $f(x_k) = f(x_\star)$, the sequence is frozen and the distance recurrence below holds with equality. Otherwise the gap is positive, so every $g_k \in \partial f(x_k)$ is nonzero. We establish the recurrence for this positive-gap case. Since $g_k \in \partial f(x_k)$, the subgradient inequality gives $f(x_k) - f(x_\star) \leq \langle g_k, x_k - x_\star\rangle$. Dividing by $\|g_k\|$, we obtain

$$t_k = \frac{f(x_k) - f(x_\star)}{\|g_k\|} \leq \frac{\langle g_k, x_k - x_\star\rangle}{\|g_k\|}.$$

Hence the Type-I admissibility condition in Proposition D.2 is satisfied. Therefore,

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2.$$

Substituting the definition of t_k yields

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - \frac{(f(x_k) - f(x_\star))^2}{\|g_k\|^2}.$$

Using the bounded subgradients assumption, we get

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - \frac{(f(x_k) - f(x_\star))^2}{G^2}.$$

This last inequality holds also at zero-gap iterates. Summing this inequality for $k = 0, 1, \ldots, K-1$, we obtain after rearranging,

$$\frac{1}{K}\sum_{k=0}^{K-1} (f(x_k) - f(x_\star))^2 \leq \frac{G^2\|x_0 - x_\star\|^2}{K}.$$

This proves (133). The bound (134) for the average iterate follows exactly as in the proof of Theorem D.7, by Jensen's inequality. □

D.2.3 Smooth convex Polyak radius

Theorem D.9 (Polyak radius in the L-smooth convex regime)**.** Let Assumption 3.1 hold, and assume that f is convex and L–smooth on an open set containing $\mathcal{X}$, with $L > 0$. Let $\{x_k\}_{k\geq 0}$ be the iterates generated by BroxGD, where $x_0 \in \mathcal{X}$, and let $x_\star \in \mathcal{X}_\star$. Whenever $f(x_k) > f(x_\star)$, choose the Polyak radius

$$t_k := \frac{f(x_k) - f(x_\star)}{\|\nabla f(x_k)\|}. \tag{135}$$

If $f(x_k) = f(x_\star)$, set $t_k = 0$ and freeze the sequence at x_k, including when x_k is a minimizer different from the selected $x_\star$.

Define

$$R_0 \coloneqq \|x_0 - x_\star\|, \qquad M_0 \coloneqq \|\nabla f(x_\star)\| + LR_0.$$

Then, for every $K \geq 1$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x_k) - f(x_\star)\right)^2 \leq \frac{{M_0}^2 {R_0}^2}{K}. \tag{136}$$

Moreover, the average iterate

$$\bar{x}_K \coloneqq \frac{1}{K}\sum_{k=0}^{K-1} x_k$$

satisfies

$$f(\bar{x}_K) - f(x_\star) \leq \frac{M_0 R_0}{\sqrt{K}}. \tag{137}$$

Proof. Write $\Delta_k = f(x_k) - f(x_\star) \geq 0$. If $\Delta_k = 0$, then $t_k = 0$ and $x_{k+1} = x_k$. Otherwise convexity ensures $\nabla f(x_k) \neq 0$ and

$$0 < t_k \overset{(135)}{=} \frac{\Delta_k}{\|\nabla f(x_k)\|} \leq \frac{\langle \nabla f(x_k), x_k - x_\star\rangle}{\|\nabla f(x_k)\|}.$$

The Type-I part of Theorem 3.1 therefore applies at positive-gap iterates. Together with the frozen zero-gap case, it gives, for every $k \geq 0$,

$$\|x_{k+1} - x_\star\|^2 \overset{(10)}{\leq} \|x_k - x_\star\|^2 - t_k^2. \tag{138}$$

In the zero-gap case this relation is simply equality and requires no invocation of the positive-radius theorem. Iterating yields $\|x_k - x_\star\| \leq R_0$; hence smoothness implies

$$\|\nabla f(x_k)\| \leq \|\nabla f(x_\star)\| + L\|x_k - x_\star\| \overset{(138)}{\leq} \|\nabla f(x_\star)\| + LR_0 = M_0. \tag{139}$$

If $M_0 = 0$, then $R_0 = 0$ because $L > 0$, and the sequence starts at a minimizer and remains there. Both claimed bounds are zero. Assume henceforth $M_0 > 0$. At every positive-gap iterate,

$$t_k^2 \overset{(135)}{=} \frac{\Delta_k^2}{\|\nabla f(x_k)\|^2} \overset{(139)}{\geq} \frac{\Delta_k^2}{{M_0}^2}.$$

The inequality $t_k^2 \geq \Delta_k^2/{M_0}^2$ also holds at zero-gap iterates, with both sides zero. Summing the distance inequality therefore gives

$$\frac{1}{{M_0}^2}\sum_{k=0}^{K-1}\Delta_k^2 \leq \sum_{k=0}^{K-1} t_k^2 \overset{(138)}{\leq} {R_0}^2,$$

which proves (136). Finally,

$$0 \leq f(\bar{x}_K) - f(x_\star) \leq \frac{1}{K}\sum_{k=0}^{K-1}\Delta_k \leq \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k^2} \overset{(136)}{\leq} \frac{M_0 R_0}{\sqrt{K}},$$

by convexity and Cauchy–Schwarz. This proves (137). □

Remark D.10. *Theorem D.9 gives an $\mathcal{O}(1/\sqrt{K})$ objective-error guarantee for the Polyak radius. The fixed-horizon schedule in Corollary D.2 instead gives an $\mathcal{O}(1/K)$ best-iterate objective bound. This comparison concerns the available guarantees: the weaker Polyak analysis does not prove that this radius cannot achieve a faster rate. The $\mathcal{O}(1/K)$ bound in Theorem D.11 controls a squared gradient-difference residual and is a different metric.*

D.2.4 Sharp strongly convex distance contraction

For convenience, we restate Theorem 3.4 here.

Theorem 3.4 (Sharp strongly convex contraction)**.** Let Assumption 3.1 hold and suppose f is L-smooth and μ-strongly convex on $\mathcal{X}$, where $0 < \mu \le L$. For a fixed $x_\star \in \mathcal{X}_\star$, use

$$t_k = \theta\|x_k - x_\star\|, \qquad \theta = \frac{2\sqrt{\mu L}}{L+\mu}. \tag{20}$$

Freeze the sequence at $x_\star$. Then, for every $K \ge 0$,

$$\|x_K - x_\star\|^2 \le \left(\frac{L-\mu}{L+\mu}\right)^{2K} R_0^2. \tag{21}$$

We give two proofs. The first is shorter but assumes that f is L-smooth and μ-strongly convex on all of $\mathbb{R}^d$. The second proves the theorem under its stated assumptions on $\mathcal{X}$ alone, including lower-dimensional feasible sets.

Proof under global smoothness and strong convexity. At an iterate $x_k \ne x_\star$, set $d_k = x_k - x_\star$ and $w_k = \nabla f(x_k) - \nabla f(x_\star)$. Strong convexity gives $\langle w_k, d_k\rangle \ge \mu\|d_k\|^2 > 0$, so $w_k \ne 0$. The global assumptions allow a direct application of Lemma C.3:

$$t_k \overset{(20)}{=} \theta\|d_k\| \overset{(56)}{\le} \frac{\langle w_k, d_k\rangle}{\|w_k\|}.$$

This is the Type-II radius condition of Theorem 3.1. Therefore

$$\begin{aligned}\|x_{k+1} - x_\star\|^2 &\overset{(10)}{\le} \|x_k - x_\star\|^2 - t_k^2 \\ &\overset{(20)}{=} \left(\frac{L-\mu}{L+\mu}\right)^2 \|x_k - x_\star\|^2.\end{aligned}$$

The same contraction holds for the frozen continuation at $x_\star$. Iterating proves (21). At $K = 0$ the bound is the initial identity; if $L = \mu$, the method reaches $x_\star$ in at most one step. □

Proof under the stated assumptions on $\mathcal{X}$. **Step 1: Restrict gradients to the affine hull.** Write $\mathrm{aff}(\mathcal{X}) = a + V$, where V is a linear subspace, and let P_V be its orthogonal projector. If $V = \{0\}$, the feasible set is a singleton and the conclusion is immediate. Otherwise define the intrinsic gradient

$$b(x) = P_V \nabla f(x) \qquad (x \in \mathcal{X}). \tag{140}$$

For $x, y \in \mathcal{X}$, the displacement $x - y$ lies in V. Hence

$$\langle b(y), x - y\rangle = \langle \nabla f(y), x - y\rangle, \qquad \|b(x) - b(y)\| \le L\|x - y\|. \tag{141}$$

The second inequality uses nonexpansiveness of P_V and the assumed smoothness on $\mathcal{X}$. Thus the restriction of f to $\mathcal{X}$, viewed in $a + V$, retains the same strong-convexity and smoothness constants. Moreover, replacing $\nabla f(x_k)$ by $b(x_k)$ leaves every local linear minimizer unchanged, since their inner products with feasible displacements coincide by (141). No projection is required by the algorithm; P_V is used only in the proof.

Step 2: Establish interpolation, including at the relative boundary. We claim that, with $d = x - y$ and $w = b(x) - b(y)$,

$$\langle w, d\rangle \ge \frac{\|w\|^2 + \mu L\|d\|^2}{L+\mu} \qquad (x, y \in \mathcal{X}). \tag{142}$$

Here it is essential to use intrinsic gradients: the analogous assertion for ambient gradient differences need not hold under assumptions imposed only on a lower-dimensional $\mathcal{X}$.

First take x, y in the relative interior $U = \mathrm{ri}(\mathcal{X})$, which is open and convex in $a + V$. The segment $[x, y]$ has a neighborhood contained in U. In orthonormal coordinates on V, convolve the restricted function with a nonnegative smooth kernel of integral one and sufficiently small support. Denote the resulting smooth function by f_ε. On a neighborhood of this segment, averaging preserves μ-strong convexity and the L-Lipschitz gradient bound in (141). Therefore

$$\mu I_V \preceq \nabla_V^2 f_\varepsilon(z) \preceq L I_V, \qquad H_\varepsilon = \int_0^1 \nabla_V^2 f_\varepsilon(y + sd)\, ds, \qquad \mu I_V \preceq H_\varepsilon \preceq L I_V. \tag{143}$$

The fundamental theorem of calculus gives

$$w_\varepsilon = \nabla_V f_\varepsilon(x) - \nabla_V f_\varepsilon(y) = H_\varepsilon d. \tag{144}$$

Every eigenvalue λ of H_ε lies in $[\mu, L]$, so $(\lambda - \mu)(L - \lambda) \geq 0$. Consequently,

$$(L+\mu)\langle w_\varepsilon, d\rangle - \|w_\varepsilon\|^2 - \mu L\|d\|^2 \overset{(144)}{=} \langle d, (H_\varepsilon - \mu I_V)(LI_V - H_\varepsilon)d\rangle \overset{(143)}{\geq} 0.$$

The intrinsic gradient is continuous, so the mollified gradients converge to it as $\varepsilon \downarrow 0$. Passing to the limit proves (142) on U.

For arbitrary $x, y \in \mathcal{X}$, fix $c \in U$ and take $x_\eta = (1-\eta)x + \eta c$ and $y_\eta = (1-\eta)y + \eta c$, where $0 < \eta < 1$. Both points belong to U. Apply the inequality just proved to x_η, y_η and let $\eta \downarrow 0$. Continuity from (141) proves (142) on all of $\mathcal{X}$.

Step 3: Apply the one-step geometry intrinsically. If $x_k = x_\star$, the prescribed frozen continuation satisfies the claim. Otherwise set $d_k = x_k - x_\star$ and $w_k = b(x_k) - b(x_\star)$. Adding the two strong-convexity inequalities gives

$$\langle w_k, d_k\rangle \overset{(141)}{=} \langle \nabla f(x_k) - \nabla f(x_\star), d_k\rangle \geq \mu\|d_k\|^2 > 0,$$

so $w_k \neq 0$. Arithmetic–geometric mean and the intrinsic interpolation inequality imply

$$t_k \overset{(20)}{=} \theta\|d_k\| \leq \frac{\|w_k\|^2 + \mu L\|d_k\|^2}{(L+\mu)\|w_k\|} \overset{(142)}{\leq} \frac{\langle w_k, d_k\rangle}{\|w_k\|}. \tag{145}$$

Identify $a + V$ isometrically with $\mathbb{R}^{\dim V}$. The feasible set remains closed and convex, the local oracle is unchanged by Step 1, and the restricted differentiable objective has minimizer $x_\star$. Thus (145) is precisely the Type-II condition of Theorem 3.1 in these coordinates. It follows that

$$\|x_{k+1} - x_\star\|^2 \overset{(10)}{\leq} \|x_k - x_\star\|^2 - t_k^2 \overset{(20)}{=} \left(\frac{L-\mu}{L+\mu}\right)^2 \|x_k - x_\star\|^2.$$

Iteration proves (21). The case $K = 0$ is the initial identity. When $L = \mu$, the one-step factor is zero, so the method reaches $x_\star$ in at most one step and then stays there; the $K = 0$ bound is understood separately. □

D.3 C. GRADIENT-DIFFERENCE RESIDUAL GUARANTEES

These results control gradient differences rather than objective gaps. The order matches group C of Table 1.

D.3.1 SMOOTH CONVEX GRADIENT-DIFFERENCE RADIUS

Theorem D.11 (Rate under L–smoothness)**.** Let Assumption 3.1 hold, and assume that f is L–smooth and strictly convex, with $L > 0$. If $\{x_k\}_{k\geq 0}$ are the iterates generated by Algorithm 1 with radii

$$t_k = \tfrac{\|\nabla f(x_k)-\nabla f(x_\star)\|}{L}, \tag{146}$$

where $x_0 \in \mathcal{X}$ and $x_\star \in \mathcal{X}_\star$, then, for every $K \geq 1$,

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{K}. \tag{147}$$

Proof. If $\nabla f(x_k) = \nabla f(x_\star)$, then $t_k = 0$ and $x_{k+1} = x_k$, so the distance recurrence below holds with equality. Otherwise the residual is nonzero. By cocoercivity,

$$\frac{1}{L}\|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle,$$

so the choice $t_k = \frac{\|\nabla f(x_k)-\nabla f(x_\star)\|}{L}$ satisfies radius condition (8). Hence Theorem 3.1 gives

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|^2}{L^2}.$$

The recurrence therefore holds in both cases. Summing this for $k = 0, \ldots, K-1$ yields

$$\|x_K - x_\star\|^2 \leq \|x_0 - x_\star\|^2 - \frac{1}{L^2}\sum_{k=0}^{K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2.$$

Since $\|x_k - x_\star\|^2 \geq 0$, we obtain

$$\sum_{k=0}^{K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq L^2\|x_0 - x_\star\|^2,$$

which proves (147). Dividing by K and taking the minimum gives

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{L^2\|x_0 - x_\star\|^2}{K}.$$

□

Remark D.12. *The proof of Theorem D.11 relies on the Type-II admissibility condition in* (8)*. Consequently, we must ensure that the denominator in* (8) *does not vanish whenever this condition is invoked. The zero-residual branch is handled directly by the singleton local ball, without invoking that condition. Strict convexity additionally ensures that a zero residual identifies $x_k = x_\star$; more generally, gradient injectivity on $\mathcal{X}$ suffices for this identification.*

Remark D.13. *If $\nabla f(x_\star) = 0$, the radius choice in* (146) *simplifies to $t_k = \|\nabla f(x_k)\|/L$. When $\mathcal{X} = \mathbb{R}^d$, this translates into the standard* GD *stepsize $\gamma_k = t_k/\|\nabla f(x_k)\| = 1/L$ (see* (6) *and Gruntkowska et al. (2025, Theorem 3.2)).*

Theorem D.11 bounds the best squared gradient-difference residual by $L^2R_0^2/K$. This matches the residual upper bound proved for ProjGD in Appendix E; it is not an objective-error bound. The counterexample in Appendix E.2 rules out replacing L^2 by L uniformly in this residual estimate, but does not establish optimality of the dependence on K.

For an objective-error comparison, the fixed-horizon BroxGD schedule in Corollary D.2 gives a best-iterate bound of order LR_0^2/K when its certified distance bound equals R_0. Classical FW on a compact domain gives a last-iterate objective bound of order $L\mathcal{D}_{\mathcal{X}}^2/K$ (Jaggi, 2013). These concern the same objective metric, but use different iterates, inputs, and oracles: the local oracle optimizes over a ball intersection, whereas FW uses a global LMO. Neither comparison alone establishes an oracle-complexity lower bound or a computational advantage.

D.3.2 CONSTANT-RADIUS RESIDUAL GUARANTEE

The same constant-radius algorithm also admits a guarantee for the gradient-difference residual. This complements the objective guarantee in Theorem D.2; the two conclusions concern different quantities. The radius is implementable, but the residual itself involves the unknown gradient at a minimizer and is generally not an observable stopping criterion.

Theorem D.14 (Constant-radius residual guarantee)**.** Let Assumption C.1 hold, and suppose that f has an L-Lipschitz gradient on its open convex domain, with $L > 0$. Fix a constant radius

$$t_k \equiv t > 0. \tag{148}$$

Write $v_k = \nabla f(x_k) - \nabla f(x_\star)$. If $\|v_k\| \geq Lt$, the radius is Type-II admissible in (8). At every such iteration (and, more generally, whenever that admissibility condition holds),

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t^2. \tag{149}$$

For every integer $K > R_0^2/t^2$,

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k) - \nabla f(x_\star)\| < Lt. \tag{150}$$

With the endpoint included, every integer $K \geq R_0^2/t^2$ satisfies

$$\min_{0\leq k\leq K} \|\nabla f(x_k) - \nabla f(x_\star)\| < Lt. \tag{151}$$

Proof. Convexity and smoothness imply cocoercivity:

$$\frac{1}{L}\|v_k\|^2 \leq \langle v_k, x_k - x_\star\rangle. \tag{152}$$

If $\|v_k\| \geq Lt$, then $\|v_k\| > 0$, and

$$t \leq \frac{\|v_k\|}{L} \overset{(152)}{\leq} \frac{\langle v_k, x_k - x_\star\rangle}{\|v_k\|}.$$

This verifies Type-II admissibility. The master theorem then gives

$$\|x_{k+1} - x_\star\|^2 \overset{(10)}{\leq} \|x_k - x_\star\|^2 - t_k^2 \overset{(148)}{=} \|x_k - x_\star\|^2 - t^2,$$

proving (149). If the residual is at least Lt at all queried points $x_0, \ldots, x_{K-1}$, summation yields

$$0 \leq \|x_K - x_\star\|^2 \overset{(149)}{\leq} R_0^2 - Kt^2. \tag{153}$$

For $K > R_0^2/t^2$ the right side is negative, a contradiction proving (150). At equality, (153) forces $x_K = x_\star$, whose residual is zero and hence strictly less than Lt. This proves (151); if $K = 0$ is permitted by the budget, $R_0 = 0$ and the assertion is immediate. No strict convexity is needed: a zero residual already meets the stated target, irrespective of whether it identifies a unique minimizer. □

Corollary D.3 (Fixed-horizon residual guarantee)**.** Under the hypotheses of Theorem D.14, let $R \geq R_0$, $R > 0$, and $K \geq 1$. Using $t = R/\sqrt{K}$ for K oracle calls gives

$$\min_{0\leq k\leq K} \|\nabla f(x_k) - \nabla f(x_\star)\| < \frac{LR}{\sqrt{K}}. \tag{154}$$

Proof. The chosen radius satisfies $R_0^2/t^2 = KR_0^2/R^2 \leq K$. Therefore,

$$\min_{0\leq k\leq K} \|\nabla f(x_k) - \nabla f(x_\star)\| \overset{(151)}{<} Lt = LR/\sqrt{K}.$$

□

This formalizes the fixed-horizon interpretation while retaining the endpoint required at equality in the budget. If $R = 0$ is certified, the starting point is already a minimizer. Although Corollaries D.2 and D.3 use the same radius, their good iterates need not coincide: selecting the smallest observed objective certifies the objective bound, but need not select the point with smallest gradient-difference residual.

D.3.3 GENERALIZED-SMOOTHNESS GRADIENT-DIFFERENCE RADIUS

We now consider the generalized smoothness regime in place of the standard L–smoothness assumption. Throughout this section, $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable on $\mathbb{R}^d$, and the following smoothness condition holds globally. The feasible set $\mathcal{X}$ may still be unbounded or lower-dimensional.

Assumption D.2. The differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is globally asymmetrically (L_0, L_1)–smooth: there exist constants $L_0 > 0$, $L_1 \geq 0$ such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1\|\nabla f(y)\|)\,\|x - y\|.$$

We refer the readers to Gorbunov et al. (2025) for further details on this assumption.

Global convexity and Assumption D.2 imply the following cocoercivity inequality. Requiring the smoothness bound only for pairs in $\mathcal{X}$ would not suffice for this ambient-gradient conclusion.

Lemma D.2 (Lemma 2.2 of Gorbunov et al. (2025))**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex on $\mathbb{R}^d$ and satisfy the global smoothness condition in Assumption D.2. For any $x, y \in \mathbb{R}^d$, we have

$$\frac{1}{2}\left(\frac{\|\nabla f(x) - \nabla f(y)\|^2}{L_0 + L_1\|\nabla f(y)\|} + \frac{\|\nabla f(x) - \nabla f(y)\|^2}{L_0 + L_1\|\nabla f(x)\|}\right) \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle\,.$$

Theorem D.15 (Rate under asymmetric (L_0, L_1)–smoothness)**.** Let Assumptions 3.1 and D.2 hold, with f convex on $\mathbb{R}^d$ and the asymmetric smoothness bound valid for all pairs in $\mathbb{R}^d$. Let $\{x_k\}_{k\geq 0}$ be the iterates generated by Algorithm 1. Assume that the radii are chosen as

$$t_k := \begin{cases} \dfrac{1}{2}\left(\dfrac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \dfrac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|}\right), & \nabla f(x_k) \neq \nabla f(x_\star), \\ 0, & \nabla f(x_k) = \nabla f(x_\star). \end{cases}$$

Then, for every $K \geq 1$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{1}{2}\left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|}\right)\right)^2 \leq \frac{\|x_0 - x_\star\|^2}{K},$$

and hence

$$\min_{0\leq k\leq K-1}\left(\frac{1}{2}\left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|}\right)\right)^2 \leq \frac{\|x_0 - x_\star\|^2}{K}.$$

Moreover, if $\sqrt{K} > L_1\|x_0 - x_\star\|$, then

$$\min_{0\leq k\leq K-1}\|\nabla f(x_k) - \nabla f(x_\star)\| \leq \frac{(L_0 + L_1\|\nabla f(x_\star)\|)\,\|x_0 - x_\star\|}{\sqrt{K} - L_1\|x_0 - x_\star\|}.$$

Thus, if $K \geq 4L_1^2\|x_0 - x_\star\|^2$, then

$$\min_{0\leq k\leq K-1}\|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{4(L_0 + L_1\|\nabla f(x_\star)\|)^2\|x_0 - x_\star\|^2}{K}.$$

Proof. Fix $k \geq 0$. If $\nabla f(x_k) = \nabla f(x_\star)$, then $t_k = 0$ and $x_{k+1} = x_k$, so the distance-decrease inequality below holds with equality. We prove it for nonzero residuals next and then sum the resulting recurrence over all iterations. Assume now that $\nabla f(x_k) \neq \nabla f(x_\star)$. By Lemma D.2 with $x = x_k$ and $y = x_\star$, we have

$$\frac{1}{2}\left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|^2}{L_0 + L_1\|\nabla f(x_\star)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|^2}{L_0 + L_1\|\nabla f(x_k)\|}\right) \leq \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle\,.$$

Dividing both sides by $\|\nabla f(x_k) - \nabla f(x_\star)\|$, we obtain

$$t_k \leq \frac{\langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star\rangle}{\|\nabla f(x_k) - \nabla f(x_\star)\|}.$$

Hence the Type-II admissibility condition of Theorem 3.1 is satisfied. Therefore,

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - t_k^2.$$

Summing this inequality for $k = 0, 1, \dots, K-1$, rearranging and substituting the definition of t_k yield

$$\|x_K - x_\star\|^2 + \sum_{k=0}^{K-1} \left(\frac{1}{2} \left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|} \right) \right)^2 \le \|x_0 - x_\star\|^2.$$

Dropping the nonnegative term $\|x_K - x_\star\|^2$ and dividing by K, we obtain

$$\frac{1}{K} \sum_{k=0}^{K-1} \left(\frac{1}{2} \left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|} \right) \right)^2 \le \frac{\|x_0 - x_\star\|^2}{K}.$$

Now we apply $\min_{0 \le k \le K-1} a_k \le \frac{1}{K} \sum_{k=0}^{K-1} a_k$, with

$$a_k := \left(\frac{1}{2} \left(\frac{1}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{1}{L_0 + L_1\|\nabla f(x_\star)\|} \right) \cdot \|\nabla f(x_k) - \nabla f(x_\star)\| \right)^2,$$

we have

$$\min_{0 \le k \le K-1} \left(\frac{1}{2} \left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|} \right) \right)^2 \le \frac{\|x_0 - x_\star\|^2}{K}.$$

To interpret the result, notice that by the triangle inequality, we have

$$L_0 + L_1\|\nabla f(x_k)\| \le L_0 + L_1\|\nabla f(x_\star)\| + L_1\|\nabla f(x_k) - \nabla f(x_\star)\|,$$
$$L_0 + L_1\|\nabla f(x_\star)\| \le L_0 + L_1\|\nabla f(x_k)\| + L_1\|\nabla f(x_k) - \nabla f(x_\star)\|.$$

We can thus lower bound

$$\frac{1}{2} \left(\frac{1}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{1}{L_0 + L_1\|\nabla f(x_\star)\|} \right) \ge \frac{1}{L_0 + L_1\|\nabla f(x_\star)\| + L_1\|\nabla f(x_k) - \nabla f(x_\star)\|}.$$

Therefore,

$$\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\| + L_1\|\nabla f(x_k) - \nabla f(x_\star)\|} \le \frac{1}{2} \left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|} \right).$$

Hence

$$\min_{0 \le k \le K-1} \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\| + L_1\|\nabla f(x_k) - \nabla f(x_\star)\|} \le \frac{\|x_0 - x_\star\|}{\sqrt{K}}.$$

Now fix an index $k \in \{0, \dots, K-1\}$ attaining the minimum above, and suppose that $\sqrt{K} > L_1\|x_0 - x_\star\|$. Then

$$\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\| + L_1\|\nabla f(x_k) - \nabla f(x_\star)\|} \le \frac{\|x_0 - x_\star\|}{\sqrt{K}}.$$

Rearranging gives

$$\left(1 - \frac{L_1\|x_0 - x_\star\|}{\sqrt{K}} \right) \|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{(L_0 + L_1\|\nabla f(x_\star)\|)\,\|x_0 - x_\star\|}{\sqrt{K}}.$$

Since $1 - \frac{L_1\|x_0 - x_\star\|}{\sqrt{K}} > 0$, we conclude that

$$\|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{(L_0 + L_1\|\nabla f(x_\star)\|)\,\|x_0 - x_\star\|}{\sqrt{K} - L_1\|x_0 - x_\star\|}.$$

Therefore,

$$\min_{0\le k\le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{(L_0 + L_1\|\nabla f(x_\star)\|)\,\|x_0 - x_\star\|}{\sqrt{K} - L_1\|x_0 - x_\star\|}.$$

Finally, if $K \ge 4L_1^2\|x_0 - x_\star\|^2$, then

$$\sqrt{K} - L_1\|x_0 - x_\star\| \ge \frac{\sqrt{K}}{2},$$

and hence

$$\min_{0\le k\le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{2\,(L_0 + L_1\|\nabla f(x_\star)\|)\,\|x_0 - x_\star\|}{\sqrt{K}}.$$

Squaring both sides gives

$$\min_{0\le k\le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \le \frac{4\,(L_0 + L_1\|\nabla f(x_\star)\|)^2\,\|x_0 - x_\star\|^2}{K}.$$

□

Remark D.16. *Theorem D.15 is the natural analogue of Theorem D.11 under asymmetric (L_0, L_1)–smoothness. In the standard L–smooth case, the coefficient in front of $\|\nabla f(x_k) - \nabla f(x_\star)\|$ is the constant $1/L$, and one recovers the bound*

$$\min_{0\le k\le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{L_0\|x_0 - x_\star\|}{\sqrt{K}}.$$

D.4 D. BEYOND CONVEXITY

Group D of Table 1 first gives stationarity without convexity, then objective convergence under projected PŁ. In both rows, $G_\gamma(x) = \gamma^{-1}(x - \mathrm{Proj}_{\mathcal{X}}(x - \gamma\nabla f(x)))$ is the projected gradient mapping.

D.4.1 SMOOTH NONCONVEX STATIONARITY

Theorem D.17 (Sublinear convergence in the smooth non-convex case)**.** Let Assumption 3.1 hold without f being convex. Let f be L–smooth on an open set containing $\mathcal{X}$. Assume that the constrained minimum $f_\star = \min_{x\in\mathcal{X}} f(x)$ is attained. Fix some $\gamma \in (0, 1/L]$ and let $\{x_k\}_{k\geq 0}$ be generated by

$$x_{k+1} \in \operatorname*{argmin}_{z\in\mathcal{X}\cap\mathcal{B}(x_k,t_k)} \langle \nabla f(x_k), z\rangle,$$

with radii

$$t_k := \gamma\|G_\gamma(x_k)\|.$$

Then, for every $k \geq 0$,

$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{2}\|G_\gamma(x_k)\|^2.$$

Consequently, for every $K \geq 1$,

$$\min_{0\leq k\leq K-1} \|G_\gamma(x_k)\|^2 \leq \frac{2\,(f(x_0) - f_\star)}{\gamma K}.$$

In particular, if $\gamma = 1/L$, then

$$\min_{0\leq k\leq K-1} \left\|G_{1/L}(x_k)\right\|^2 \leq \frac{2L\,(f(x_0) - f_\star)}{K}.$$

Proof. Fix $k \geq 0$, and for simplicity, define

$$y_k := \mathrm{Proj}_{\mathcal{X}}\left(x_k - \gamma\nabla f(x_k)\right).$$

Then, by the definition of G_γ,

$$y_k = x_k - \gamma G_\gamma(x_k), \qquad \|y_k - x_k\| = \gamma\|G_\gamma(x_k)\| = t_k.$$

Hence $y_k \in \mathcal{X} \cap \mathcal{B}_k$, and by the optimality of x_{k+1} for the local linear minimization problem,

$$\langle \nabla f(x_k), x_{k+1}\rangle \leq \langle \nabla f(x_k), y_k\rangle.$$

Equivalently, $\langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq \langle \nabla f(x_k), y_k - x_k\rangle$. By the optimality condition of projection onto the closed convex set $\mathcal{X}$ for $y_k = \mathrm{Proj}_{\mathcal{X}}\left(x_k - \gamma\nabla f(x_k)\right)$, we have

$$\langle y_k - (x_k - \gamma\nabla f(x_k)), z - y_k\rangle \geq 0, \qquad \forall z \in \mathcal{X}.$$

Choosing $z = x_k \in \mathcal{X}$, we obtain

$$\langle y_k - x_k + \gamma\nabla f(x_k), x_k - y_k\rangle \geq 0.$$

Expanding this inequality gives

$$\langle \nabla f(x_k), y_k - x_k\rangle \leq -\frac{1}{\gamma}\|y_k - x_k\|^2 = -\gamma\|G_\gamma(x_k)\|^2.$$

Therefore,

$$\langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq -\gamma\|G_\gamma(x_k)\|^2.$$

Since f is L–smooth, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Because $x_{k+1} \in \mathcal{B}_k$, we also have

$$\|x_{k+1} - x_k\| \le t_k = \gamma \|G_\gamma(x_k)\|.$$

Substituting the previous two estimates yields

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) - \gamma \|G_\gamma(x_k)\|^2 + \frac{L\gamma^2}{2} \|G_\gamma(x_k)\|^2 \\ &= f(x_k) - \gamma \left(1 - \frac{L\gamma}{2}\right) \|G_\gamma(x_k)\|^2. \end{aligned}$$

Since $\gamma \le 1/L$, we have $1 - L\gamma/2 \ge 1/2$, and thus

$$f(x_{k+1}) \le f(x_k) - \frac{\gamma}{2} \|G_\gamma(x_k)\|^2.$$

Summing this inequality for $k = 0, 1, \dots, K-1$, we obtain

$$\frac{\gamma}{2} \sum_{k=0}^{K-1} \|G_\gamma(x_k)\|^2 \le f(x_0) - f(x_K) \le f(x_0) - f_\star.$$

Hence

$$\sum_{k=0}^{K-1} \|G_\gamma(x_k)\|^2 \le \frac{2\,(f(x_0) - f_\star)}{\gamma}.$$

Dividing by K, we get

$$\min_{0 \le k \le K-1} \|G_\gamma(x_k)\|^2 \le \frac{1}{K} \sum_{k=0}^{K-1} \|G_\gamma(x_k)\|^2 \le \frac{2\,(f(x_0) - f_\star)}{\gamma K}.$$

The last claim follows by setting $\gamma = 1/L$. □

Remark D.18 (Interpretation of the non-convex guarantee)**.** *The conclusion of Theorem D.17 is a rate to approximate first order stationarity, measured by the projected gradient mapping. The proof of the theorem shows that*

$$\sum_{k=0}^{\infty} \|G_\gamma(x_k)\|^2 < \infty,$$

and hence

$$\lim_{k \to \infty} \|G_\gamma(x_k)\| = 0.$$

Therefore, if $x_{k_j} \to \bar{x}$ along some subsequence, then by continuity of ∇f and of the Euclidean projection onto a closed convex set,

$$G_\gamma(\bar{x}) = \lim_{j \to \infty} G_\gamma(x_{k_j}) = 0.$$

Equivalently,

$$\bar{x} = \mathrm{Proj}_{\mathcal{X}}\left(\bar{x} - \gamma \nabla f(\bar{x})\right),$$

and hence

$$\langle \nabla f(\bar{x}), z - \bar{x} \rangle \ge 0, \qquad \forall z \in \mathcal{X},$$

or, equivalently,

$$0 \in \nabla f(\bar{x}) + N_{\mathcal{X}}(\bar{x}).$$

Thus, every accumulation point is a first order stationary point of the constrained problem.

D.4.2 Projected-PL objective convergence

We next analyze the non-convex case under the Polyak–Łojasiewicz assumption, which is standard in nonconvex optimization.

The problem with standard PŁ. The usual Polyak–Łojasiewicz (PŁ) inequality is

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left(f(x) - f(x_\star)\right), \quad \forall x \in \mathcal{X}.$$

For unconstrained problems, this is natural, because if $x_\star$ is a minimizer, then necessarily $\nabla f(x_\star) = 0$. Therefore, the norm of the full gradient is a valid measure of stationarity. However, for constrained problems, a constrained minimizer $x_\star \in \mathcal{X}_\star$ does not in general satisfy $\nabla f(x_\star) = 0$. Instead, what holds is the first-order optimality condition

$$\langle \nabla f(x_\star), z - x_\star \rangle \geq 0, \qquad \forall z \in \mathcal{X}.$$

A natural first attempt to extend the Polyak–Łojasiewicz inequality to the constrained setting is to impose

$$\frac{1}{2}\|\nabla f(x) - \nabla f(x_\star)\|^2 \geq \mu\left(f(x) - f(x_\star)\right), \qquad \forall x \in \mathcal{X}.$$

However, this condition is not the most intrinsic constrained analogue of the PŁ inequality. The quantity $\|\nabla f(x) - \nabla f(x_\star)\|$ measures whether the ambient gradient at x matches the ambient gradient at one fixed minimizer $x_\star$. But constrained stationarity is not characterized by matching $\nabla f(x_\star)$. Rather, it is characterized by the condition $0 \in \nabla f(x) + N_{\mathcal{X}}(x)$, where the normal cone depends on the current point x. This suggests that a constrained PŁ condition should be formulated in terms of a residual whose zero set coincides with the constrained first-order condition, rather than in terms of agreement with the ambient gradient at a fixed minimizer.

Projected PŁ. A more principled way to extend the Polyak–Łojasiewicz inequality to the constrained setting is to view the problem through the Projected Gradient Descent framework. For a stepsize $\gamma > 0$, the ProjGD update is $x^+ = \mathrm{Proj}_{\mathcal{X}}\big(x - \gamma \nabla f(x)\big)$. This naturally leads to the projected gradient mapping

$$G_\gamma(x) \coloneqq \frac{1}{\gamma}\left(x - \mathrm{Proj}_{\mathcal{X}}(x - \gamma \nabla f(x))\right).$$

This object is the constrained analogue of the gradient. When $\mathcal{X} = \mathbb{R}^d$, we have

$$G_\gamma(x) = \frac{1}{\gamma}\left(x - \mathrm{Proj}_{\mathbb{R}^d}(x - \gamma \nabla f(x))\right) = \nabla f(x).$$

Moreover, for a closed convex set $\mathcal{X}$, the condition $G_\gamma(x) = 0$ is equivalent to

$$x = \mathrm{Proj}_{\mathcal{X}}\left(x - \gamma \nabla f(x)\right).$$

The above identity, from the perspective of projection onto a closed convex set, is equivalent to $\langle -\gamma \nabla f(x), z - x \rangle \leq 0$ for all $z \in \mathcal{X}$, i.e.,

$$\langle \nabla f(x), z - x \rangle \geq 0, \quad \forall z \in \mathcal{X},$$

which is precisely the variational inequality form of first-order stationarity for minimizing f over $\mathcal{X}$. More explicitly, the above condition is equivalent to $-\nabla f(x) \in N_{\mathcal{X}}(x)$, or $0 \in \nabla f(x) + N_{\mathcal{X}}(x)$.

Assumption D.3 (Projected PŁ)**.** Let Assumption 3.1 hold without f being convex. Furthermore, assume that there exist constants $\mu > 0$ and $\gamma > 0$ such that

$$\frac{1}{2}\|G_\gamma(x)\|^2 \geq \mu\left(f(x) - f(x_\star)\right), \qquad \forall x \in \mathcal{X}, \tag{155}$$

where

$$G_\gamma(x) = \frac{1}{\gamma}\left(x - \mathrm{Proj}_{\mathcal{X}}\left(x - \gamma \nabla f(x)\right)\right).$$

Theorem D.19 (Linear convergence under Projected PŁ)**.** Let Assumption D.3 hold for some $\gamma \in (0, 1/L]$, and let f be L–smooth on an open set containing $\mathcal{X}$. Assume that the constrained minimum $f_\star = \min_{x \in \mathcal{X}} f(x)$ is attained. Let $L > 0$ and $x_0 \in \mathcal{X}$. Let $\{x_k\}_{k \geq 0}$ be generated by

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle,$$

with radii

$$t_k \coloneqq \gamma \|G_\gamma(x_k)\|. \tag{156}$$

Then, for every $k \geq 0$,

$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{2}\|G_\gamma(x_k)\|^2.$$

Consequently, for any $k \geq 0$,

$$f(x_k) - f_\star \leq (1 - \gamma\mu)^k (f(x_0) - f_\star).$$

In particular, if $\gamma = 1/L$, then

$$f(x_k) - f_\star \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f_\star).$$

Proof. Fix $k \geq 0$, and for simplicity, define

$$y_k \coloneqq \mathrm{Proj}_{\mathcal{X}}(x_k - \gamma\nabla f(x_k)).$$

Then, by the definition of G_γ,

$$y_k = x_k - \gamma G_\gamma(x_k), \qquad \|y_k - x_k\| = \gamma\|G_\gamma(x_k)\| \overset{(156)}{=} t_k. \tag{157}$$

In addition, $y_k \in \mathcal{X} \cap \mathcal{B}_k$, and thus

$$\langle \nabla f(x_k), x_{k+1}\rangle \leq \langle \nabla f(x_k), y_k\rangle,$$

or equivalently,

$$\langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq \langle \nabla f(x_k), y_k - x_k\rangle. \tag{158}$$

By the optimality condition of projection onto a convex set, for $y_k = \mathrm{Proj}_{\mathcal{X}}(x_k - \gamma\nabla f(x_k))$ we have

$$\langle y_k - (x_k - \gamma\nabla f(x_k)), z - y_k\rangle \geq 0, \qquad \forall z \in \mathcal{X}.$$

Choosing $z = x_k \in \mathcal{X}$, we obtain

$$\langle y_k - x_k + \gamma\nabla f(x_k), x_k - y_k\rangle \geq 0.$$

Expanding this inequality gives

$$\langle \nabla f(x_k), y_k - x_k\rangle \leq -\frac{1}{\gamma}\|y_k - x_k\|^2 \overset{(157)}{=} -\gamma\|G_\gamma(x_k)\|^2. \tag{159}$$

By combining (158) and (159), we get

$$\langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq -\gamma\|G_\gamma(x_k)\|^2. \tag{160}$$

Since f is L–smooth, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{161}$$

Because $x_{k+1} \in \mathcal{B}_k$, we also have

$$\|x_{k+1} - x_k\| \leq t_k \overset{(156)}{=} \gamma\|G_\gamma(x_k)\|. \tag{162}$$

Substituting (160) and (162) into (161) yields

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma\|G_\gamma(x_k)\|^2 + \frac{L\gamma^2}{2}\|G_\gamma(x_k)\|^2 \\ &= f(x_k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|G_\gamma(x_k)\|^2. \end{aligned}$$

Since $\gamma \leq 1/L$, we have $1 - L\gamma/2 \geq 1/2$, and thus

$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{2}\|G_\gamma(x_k)\|^2.$$

Finally, by (155),

$$\frac{1}{2}\|G_\gamma(x_k)\|^2 \geq \mu(f(x_k) - f_\star).$$

Combining this with the descent estimate gives

$$f(x_{k+1}) \leq f(x_k) - \gamma\mu(f(x_k) - f_\star).$$

Rearranging, we obtain

$$f(x_{k+1}) - f_\star \leq (1 - \gamma\mu)(f(x_k) - f_\star).$$

Iterating this inequality yields the result. □

E Convergence theory of Projected Gradient Descent

E.1 Smooth convex case

We first consider the smooth convex regime.

Proposition E.1 (Convergence for ProjGD)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L–smooth on an open set containing $\mathcal{X}$. Let $L > 0$ and $x_0 \in \mathcal{X}$. Consider the ProjGD update

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_k - \gamma \nabla f(x_k)\right), \qquad 0 < \gamma \leq \frac{1}{L}.$$

Define the projected gradient mapping by

$$G_\gamma(x) \coloneqq \frac{1}{\gamma}\Big(x - \mathrm{Proj}_{\mathcal{X}}(x - \gamma \nabla f(x))\Big),$$

and let $x_\star \in \mathcal{X}_\star \coloneqq \operatorname{argmin}_{x \in \mathcal{X}} f(x)$. The one-step statements (i)–(ii) hold for $k \geq 0$; the subsequent rate bounds hold for integers $k \geq 1$.

(i) **One-step descent:**

$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{2}\|G_\gamma(x_k)\|^2. \tag{163}$$

(ii) **Fejér-type inequality:**

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - 2\gamma\big(f(x_{k+1}) - f(x_\star)\big).$$

(iii) **Function value rate:**

$$f(x_k) - f(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{2\gamma k}.$$

(iv) **Best-iterate projected-gradient rate:**

$$\min_{0 \leq i \leq k-1} \|G_\gamma(x_i)\|^2 \leq \frac{2\big(f(x_0) - f(x_\star)\big)}{\gamma k}.$$

(v) **Distance-based projected-gradient rate for strict stepsizes:** if $0 < \gamma < 1/L$, then

$$\min_{0 \leq i \leq k-1} \|G_\gamma(x_i)\|^2 \leq \frac{\|x_0 - x_\star\|^2}{\gamma^2(1 - \gamma L)\, k}.$$

(vi) **Ergodic function value rate:** for the average $\bar{x}_k^{\mathsf{ProjGD}} \coloneqq \frac{1}{k}\sum_{i=1}^k x_i$, one has

$$f(\bar{x}_k^{\mathsf{ProjGD}}) - f(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{2\gamma k}.$$

(vii) **Gradient-difference bound:**

$$\min_{0 \leq i \leq k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \leq \frac{\|x_0 - x_\star\|^2}{\gamma^2 k}.$$

Remark E.1. *In the special case $\gamma = 1/L$, these bounds become*

$$f(x_k) - f(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2k},$$
$$\min_{0 \leq i \leq k-1} \|G_{1/L}(x_i)\|^2 \leq \frac{2L\big(f(x_0) - f(x_\star)\big)}{k},$$
$$f(\bar{x}_k^{\mathsf{ProjGD}}) - f(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2k},$$

$$\min_{0\le i\le k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \frac{L^2\|x_0 - x_\star\|^2}{k}. \tag{164}$$

Proof. Let $z_k \coloneqq \mathrm{Proj}_{\mathcal{X}}(x_k - \gamma\nabla f(x_k))$, so that $z_k = x_{k+1}$ and $G_\gamma(x_k) = \frac{1}{\gamma}(x_k - z_k)$.

(i). By the L–smoothness of f,

$$f(z_k) \le f(x_k) + \langle \nabla f(x_k), z_k - x_k\rangle + \frac{L}{2}\|z_k - x_k\|^2.$$

Since $\gamma \le 1/L$, we have $L/2 \le 1/2\gamma$, hence

$$f(z_k) \le f(x_k) + \langle \nabla f(x_k), z_k - x_k\rangle + \frac{1}{2\gamma}\|z_k - x_k\|^2. \tag{165}$$

On the other hand, the optimality condition for projection gives

$$\langle x_k - \gamma\nabla f(x_k) - z_k, x - z_k\rangle \le 0, \qquad \forall x \in \mathcal{X}. \tag{166}$$

Since $x_k \in \mathcal{X}$, setting $x = x_k$ yields

$$\langle x_k - \gamma\nabla f(x_k) - z_k, x_k - z_k\rangle \le 0.$$

Expanding this inequality, we obtain $\langle \nabla f(x_k), z_k - x_k\rangle \le -\frac{1}{\gamma}\|z_k - x_k\|^2$. Substituting this into (165) gives

$$f(z_k) \le f(x_k) - \frac{1}{\gamma}\|z_k - x_k\|^2 + \frac{1}{2\gamma}\|z_k - x_k\|^2 = f(x_k) - \frac{1}{2\gamma}\|z_k - x_k\|^2.$$

Therefore,

$$f(x_{k+1}) = f(z_k) \le f(x_k) - \frac{1}{2\gamma}\|z_k - x_k\|^2 = f(x_k) - \frac{\gamma}{2}\|G_\gamma(x_k)\|^2.$$

(ii). Setting $x = x_\star$ in (166), we obtain

$$\langle x_k - x_{k+1}, x_\star - x_{k+1}\rangle \le \gamma\langle \nabla f(x_k), x_\star - x_{k+1}\rangle.$$

Using the identity

$$2\langle a - b, c - b\rangle = \|a - b\|^2 + \|c - b\|^2 - \|a - c\|^2$$

with $a = x_k$, $b = x_{k+1}$, and $c = x_\star$, we get

$$\frac{1}{2\gamma}\Big(\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2 - \|x_{k+1} - x_k\|^2\Big) \ge \langle \nabla f(x_k), x_{k+1} - x_\star\rangle. \tag{167}$$

By convexity, $f(x_\star) \ge f(x_k) + \langle \nabla f(x_k), x_\star - x_k\rangle$, hence

$$f(x_k) - f(x_\star) \le \langle \nabla f(x_k), x_k - x_\star\rangle. \tag{168}$$

Also, by L–smoothness,

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{169}$$

Subtracting $f(x_\star)$ from both sideds of (169) and combining if with (168),

$$f(x_{k+1}) - f(x_\star) \le \langle \nabla f(x_k), x_{k+1} - x_\star\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{170}$$

Since $\gamma \le 1/L$, we have $\frac{L}{2} \le \frac{1}{2\gamma}$, so

$$f(x_{k+1}) - f(x_\star) \le \langle \nabla f(x_k), x_{k+1} - x_\star\rangle + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2.$$

Therefore, using (167)

$$f(x_{k+1}) - f(x_\star) \le \frac{1}{2\gamma}\Big(\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2\Big),$$

which is equivalent to

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - 2\gamma\big(f(x_{k+1}) - f(x_\star)\big). \tag{171}$$

(iii). Summing (171) from $i = 0$ to $k-1$, we get

$$2\gamma \sum_{i=0}^{k-1} \big(f(x_{i+1}) - f(x_\star)\big) \le \|x_0 - x_\star\|^2.$$

Since $f(x_i)$ is nonincreasing by (163), we have

$$k\big(f(x_k) - f(x_\star)\big) \le \sum_{i=0}^{k-1} \big(f(x_{i+1}) - f(x_\star)\big).$$

Therefore, $f(x_k) - f(x_\star) \le \frac{\|x_0 - x_\star\|^2}{2\gamma k}$.

(iv). Summing the descent inequality (163) for $i = 0, \dots, k-1$, we obtain

$$\frac{\gamma}{2} \sum_{i=0}^{k-1} \|G_\gamma(x_i)\|^2 \le f(x_0) - f(x_k) \le f(x_0) - f(x_\star).$$

Hence

$$\sum_{i=0}^{k-1} \|G_\gamma(x_i)\|^2 \le \frac{2\big(f(x_0) - f(x_\star)\big)}{\gamma}.$$

Dividing by k and noticing that the minimum is upper bounded by the average yields the desired result

(v). Combining (167) with (170),

$$f(x_{k+1}) - f(x_\star) \le \frac{1}{2\gamma}\Big(\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2\Big) - \Big(\frac{1}{2\gamma} - \frac{L}{2}\Big)\|x_{k+1} - x_k\|^2.$$

Equivalently,

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - 2\gamma\big(f(x_{k+1}) - f(x_\star)\big) - (1 - \gamma L)\|x_{k+1} - x_k\|^2.$$

If $0 < \gamma < 1/L$, then $1 - \gamma L > 0$, and therefore

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - (1 - \gamma L)\|x_{k+1} - x_k\|^2.$$

Since $x_{k+1} - x_k = -\gamma G_\gamma(x_k)$, we get

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - (1 - \gamma L)\gamma^2 \|G_\gamma(x_k)\|^2.$$

Summing for $i = 0, \dots, k-1$, we obtain

$$(1 - \gamma L)\gamma^2 \sum_{i=0}^{k-1} \|G_\gamma(x_i)\|^2 \le \|x_0 - x_\star\|^2,$$

and dividing by k yields

$$\min_{0 \le i \le k-1} \|G_\gamma(x_i)\|^2 \le \frac{\|x_0 - x_\star\|^2}{\gamma^2(1 - \gamma L)k}.$$

(vi). By convexity, $f(\bar{x}_k^{\mathsf{ProjGD}}) \le \frac{1}{k}\sum_{i=1}^k f(x_i)$. Hence

$$f(\bar{x}_k^{\mathsf{ProjGD}}) - f(x_\star) \le \frac{1}{k}\sum_{i=1}^{k} \big(f(x_i) - f(x_\star)\big) = \frac{1}{k}\sum_{i=0}^{k-1} \big(f(x_{i+1}) - f(x_\star)\big) \le \frac{\|x_0 - x_\star\|^2}{2\gamma k}.$$

(vii). Since $x_\star$ minimizes f over the closed convex set $\mathcal{X}$, the first-order optimality condition implies

$$\langle \nabla f(x_\star), x - x_\star \rangle \ge 0, \qquad \forall x \in \mathcal{X}.$$

Equivalently, $x_\star = \mathrm{Proj}_{\mathcal{X}}\big(x_\star - \gamma \nabla f(x_\star)\big)$. Using the nonexpansiveness of the projection, we get

$$\|x_{i+1} - x_\star\|^2 \le \|x_i - x_\star - \gamma(\nabla f(x_i) - \nabla f(x_\star))\|^2.$$

Expanding,

$$\|x_{i+1} - x_\star\|^2 \le \|x_i - x_\star\|^2 - 2\gamma\langle \nabla f(x_i) - \nabla f(x_\star), x_i - x_\star\rangle + \gamma^2 \|\nabla f(x_i) - \nabla f(x_\star)\|^2.$$

Since f is convex and L–smooth,

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle, \qquad \forall x, y \in \mathcal{X}.$$

Because $\gamma \le 1/L$, it follows that

$$\gamma \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \langle \nabla f(x_i) - \nabla f(x_\star), x_i - x_\star\rangle.$$

Hence,

$$\|x_{i+1} - x_\star\|^2 \le \|x_i - x_\star\|^2 - \gamma^2 \|\nabla f(x_i) - \nabla f(x_\star)\|^2.$$

Summing for $i = 0, \dots, k-1$, and dividing by k,

$$\min_{0 \le i \le k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \frac{\|x_0 - x_\star\|^2}{\gamma^2 k}. \qquad \square$$

Remark E.2. *Most of the results above are known and can be found in prior literature (Rosen, 1960; Levitin & Polyak, 1966; Bertsekas, 1999; Beck, 2017). Among these bounds, the most standard headline results are the $\mathcal{O}(1/K)$ function-value estimate in part* (iii) *and the $\mathcal{O}(1/K)$ projected-gradient estimate in part* (iv)*. In constrained optimization, $\|G_\gamma(x)\|$ is the canonical stationarity measure; it vanishes if and only if x is a solution of the variational inequality $0 \in \nabla f(x) + N_{\mathcal{X}}(x)$, or equivalently, if and only if x is a minimizer of f over $\mathcal{X}$.*

E.2 A NEGATIVE RESULT FOR PROJECTED GRADIENT DESCENT

We now show that the quadratic dependence on L in the gradient-difference bound

$$\min_{0 \le i \le k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \frac{L^2 \|x_0 - x_\star\|^2}{k}$$

cannot, in general, be improved to a linear dependence on L. More precisely, we prove that there is no universal constant $C > 0$ such that, for every L–smooth convex constrained problem, the iterates of Projected Gradient Descent satisfy

$$\min_{0 \le i \le k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \frac{CL \|x_0 - x_\star\|^2}{k}. \tag{172}$$

Construction. Let the feasible set be $\mathcal{X} := \mathbb{R} \times \mathbb{R}_+$, and consider the family of convex quadratic functions

$$f_\alpha(u, v) := \frac{\alpha}{4}u^2 + \alpha u v + \alpha v^2 + v = \frac{\alpha}{4}(u + 2v)^2 + v, \qquad \alpha > 0.$$

The gradient and Hessian are

$$\nabla f_\alpha(u, v) = \begin{pmatrix} \frac{\alpha}{2}u + \alpha v \\ \alpha u + 2\alpha v + 1 \end{pmatrix}, \qquad \nabla^2 f_\alpha = \begin{pmatrix} \alpha/2 & \alpha \\ \alpha & 2\alpha \end{pmatrix}.$$

The Hessian has eigenvalues 0 and $\frac{5\alpha}{2}$. Hence f_α is convex and L–smooth with

$$L = \frac{5\alpha}{2}.$$

We claim that the constrained minimizer is $x_\star = (0, 0)$. Indeed,

$$f_\alpha(u, v) = \frac{\alpha}{4}(u + 2v)^2 + v.$$

Since $v \ge 0$ for all $(u, v) \in \mathcal{X}$, and the quadratic term is always nonnegative, we have

$$f_\alpha(u, v) \ge 0 \qquad \forall (u, v) \in \mathcal{X}.$$

Moreover, $f_\alpha(0, 0) = 0$, so $(0, 0)$ is a constrained minimizer.

ProjGD with the standard stepsize. Let $x_0 = (1, 0) \in \mathcal{X}$ and run Projected Gradient Descent with the standard stepsize

$$\gamma = \frac{1}{L} = \frac{2}{5\alpha}.$$

We first compute the gradient at x_0:

$$\nabla f_\alpha(x_0) = \nabla f_\alpha(1, 0) = \begin{pmatrix} \alpha/2 \\ \alpha + 1 \end{pmatrix}.$$

Hence the unconstrained gradient step is

$$x_0 - \gamma \nabla f_\alpha(x_0) = \left(1 - \frac{\gamma\alpha}{2}, -\gamma(\alpha + 1)\right).$$

Since the second component is negative, projection onto $\mathcal{X} = \mathbb{R} \times \mathbb{R}_+$ sets it to zero, and therefore

$$x_1 = \mathrm{Proj}_{\mathcal{X}}\big(x_0 - \gamma \nabla f_\alpha(x_0)\big) = \left(1 - \frac{\gamma\alpha}{2}, 0\right).$$

Using $\gamma = \frac{2}{5\alpha}$, we get

$$\frac{\gamma\alpha}{2} = \frac{1}{5}, \qquad \text{so} \qquad x_1 = \left(\frac{4}{5}, 0\right).$$

Gradient differences. We next compute the gradients at $x_\star$, x_0, and x_1. At the constrained minimizer,

$$\nabla f_\alpha(x_\star) = \nabla f_\alpha(0, 0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Therefore

$$\nabla f_\alpha(x_0) - \nabla f_\alpha(x_\star) = \begin{pmatrix} \alpha/2 \\ \alpha \end{pmatrix},$$

and thus

$$\|\nabla f_\alpha(x_0) - \nabla f_\alpha(x_\star)\|^2 = \left(\frac{\alpha}{2}\right)^2 + \alpha^2 = \frac{5\alpha^2}{4}.$$

Now evaluate the gradient at $x_1 = (4/5, 0)$:

$$\nabla f_\alpha(x_1) = \begin{pmatrix} \frac{\alpha}{2} \cdot \frac{4}{5} \\ \alpha \cdot \frac{4}{5} + 1 \end{pmatrix} = \begin{pmatrix} 2\alpha/5 \\ 4\alpha/5 + 1 \end{pmatrix}.$$

Hence

$$\nabla f_\alpha(x_1) - \nabla f_\alpha(x_\star) = \begin{pmatrix} 2\alpha/5 \\ 4\alpha/5 \end{pmatrix},$$

and so

$$\|\nabla f_\alpha(x_1) - \nabla f_\alpha(x_\star)\|^2 = \left(\frac{2\alpha}{5}\right)^2 + \left(\frac{4\alpha}{5}\right)^2 = \frac{4\alpha^2}{25} + \frac{16\alpha^2}{25} = \frac{4\alpha^2}{5}.$$

Therefore, for $k = 2$,

$$\min_{0 \le i \le 1} \|\nabla f_\alpha(x_i) - \nabla f_\alpha(x_\star)\|^2 = \frac{4\alpha^2}{5}.$$

Contradiction with a hypothetical linear-in-L bound. Now suppose that there exists a universal constant $C > 0$ such that (172) holds for every L–smooth convex constrained problem. Applying (172) to the present example with $k = 2$, we would obtain

$$\min_{0 \le i \le 1} \|\nabla f_\alpha(x_i) - \nabla f_\alpha(x_\star)\|^2 \le \frac{CL\|x_0 - x_\star\|^2}{2}.$$

Since $L = 5\alpha/2$ and $\|x_0 - x_\star\|^2 = \|(1, 0) - (0, 0)\|^2 = 1$, the right-hand side becomes

$$\frac{CL\|x_0 - x_\star\|^2}{2} = \frac{C}{2} \cdot \frac{5\alpha}{2} = \frac{5C\alpha}{4}.$$

But we already computed that

$$\min_{0\le i\le 1} \|\nabla f_\alpha(x_i) - \nabla f_\alpha(x_\star)\|^2 = \frac{4\alpha^2}{5}.$$

Hence (172) would force

$$\frac{4\alpha^2}{5} \le \frac{5C\alpha}{4}.$$

Equivalently,

$$\alpha \le \frac{25}{16}C.$$

Therefore, if we choose $\alpha > \frac{25}{16}C$, then (172) fails.

Since C was arbitrary, no universal constant C can make (172) valid for all L–smooth convex constrained problems. We have thus shown that there is no universal constant $C > 0$ such that

$$\min_{0\le i\le k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \frac{C\,L\|x_0 - x_\star\|^2}{k}$$

holds for all L–smooth convex constrained problems. Thus, the quadratic dependence on L in the bound

$$\min_{0\le i\le k-1} \|\nabla f(x_i) - \nabla f(x_\star)\|^2 \le \frac{L^2\|x_0 - x_\star\|^2}{k}$$

is, in general, unavoidable.

E.3 (L_0, L_1)–SMOOTH CONVEX CASE

Proposition E.2 (ProjGD under asymmetric (L_0, L_1)–smoothness)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex on $\mathbb{R}^d$ and globally asymmetrically (L_0, L_1)–smooth as in Assumption D.2. Fix $x_\star \in \operatorname{argmin}_{x\in\mathcal{X}} f(x)$, assumed to exist. Let $\{x_k\}_{k\ge 0}$ be generated by the ProjGD iteration

$$x_{k+1} = \operatorname{Proj}_{\mathcal{X}}\left(x_k - \gamma_k \nabla f(x_k)\right),$$

where $\gamma_k > 0$. If

$$\gamma_k \le \frac{1}{2}\left(\frac{1}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{1}{L_0 + L_1\|\nabla f(x_\star)\|},\right) \tag{173}$$

then for every $K \ge 1$

$$\min_{0\le k\le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \le \frac{\|x_0 - x_\star\|^2}{\sum_{k=0}^{K-1}\gamma_k^2}.$$

In particular, for

$$\gamma_k = \frac{1}{2}\left(\frac{1}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{1}{L_0 + L_1\|\nabla f(x_\star)\|}\right), \tag{174}$$

the iterates satisfy

$$\min_{0\le k\le K-1} \left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|}\right)^2 \le \frac{4\|x_0 - x_\star\|^2}{K}.$$

Moreover, if $\sqrt{K} > L_1\|x_0 - x_\star\|$, then

$$\min_{0\le k\le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{(L_0 + L_1\|\nabla f(x_\star)\|)\,\|x_0 - x_\star\|}{\sqrt{K} - L_1\|x_0 - x_\star\|},$$

and if $K \geq 4L_1^2\|x_0 - x_\star\|^2$, then

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{4\left(L_0 + L_1\|\nabla f(x_\star)\|\right)^2 \|x_0 - x_\star\|^2}{K}.$$

Proof. Denote $v_k := \nabla f(x_k) - \nabla f(x_\star)$. Since $x_\star$ is a minimizer of f over $\mathcal{X}$, for every $\gamma_k > 0$ we have $x_\star = \mathrm{Proj}_{\mathcal{X}}\left(x_\star - \gamma_k \nabla f(x_\star)\right)$ (Refer to the proof of Proposition E.1 (vii)). Therefore, by the nonexpansiveness of the projection,

$$\begin{aligned}\|x_{k+1} - x_\star\|^2 &= \|\mathrm{Proj}_{\mathcal{X}}\left(x_k - \gamma_k \nabla f(x_k)\right) - \mathrm{Proj}_{\mathcal{X}}\left(x_\star - \gamma_k \nabla f(x_\star)\right)\|^2 \\ &\leq \|x_k - x_\star - \gamma_k\left(\nabla f(x_k) - \nabla f(x_\star)\right)\|^2 \\ &= \|x_k - x_\star\|^2 - 2\gamma_k\langle v_k, x_k - x_\star\rangle + \gamma_k^2\|v_k\|^2.\end{aligned}$$

Next, by Lemma D.2

$$\frac{1}{2}\left(\frac{\|v_k\|^2}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|v_k\|^2}{L_0 + L_1\|\nabla f(x_\star)\|}\right) \leq \langle v_k, x_k - x_\star\rangle.$$

Hence, under the choice (173),

$$\gamma_k\|v_k\|^2 \leq \frac{1}{2}\left(\frac{\|v_k\|^2}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|v_k\|^2}{L_0 + L_1\|\nabla f(x_\star)\|}\right) \leq \langle v_k, x_k - x_\star\rangle.$$

Substituting this into the previous bound yields

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - \gamma_k^2\|v_k\|^2.$$

Summing for $k = 0, 1, \ldots, K-1$ and dropping the nonnegative term $\|x_K - x_\star\|^2$ gives

$$\sum_{k=0}^{K-1} \gamma_k^2\|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \|x_0 - x_\star\|^2,$$

and hence

$$\left(\min_{0\leq k\leq K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2\right)\sum_{k=0}^{K-1}\gamma_k^2 \leq \sum_{k=0}^{K-1}\gamma_k^2\|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \|x_0 - x_\star\|^2.$$

Therefore

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \leq \frac{\|x_0 - x_\star\|^2}{\sum_{k=0}^{K-1}\gamma_k^2},$$

proving the first part of the statement. Now, choosing γ_k as in (174), the bound above becomes

$$\sum_{k=0}^{K-1}\frac{1}{4}\left(\frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_k)\|} + \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L_0 + L_1\|\nabla f(x_\star)\|}\right)^2 \leq \|x_0 - x_\star\|^2.$$

The rest of the proof follows directly from the proof of Theorem D.15. □

E.4 Non-convex case

Projected gradient descent also admits convergence guarantees in the non-convex setting. The following result is obtained by specializing the proximal-gradient convergence theorem of Beck (2017, Theorem 10.15).

Proposition E.3 (ProjGD in the smooth non-convex case)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. Let $f : \mathbb{R}^d \to \mathbb{R}$ be L–smooth on an open set containing $\mathcal{X}$, and assume that $f_\star := \min_{x\in\mathcal{X}} f(x)$ is finite. Let $L > 0$ and $x_0 \in \mathcal{X}$. Let $\{x_k\}_{k\geq 0}$ be generated by the ProjGD iteration

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_k - \gamma\nabla f(x_k)\right),$$

where $0 < \gamma \leq 1/L$, and define the projected gradient mapping by

$$G_\gamma(x) = \frac{1}{\gamma}\left(x - \mathrm{Proj}_{\mathcal{X}}\left(x - \gamma\nabla f(x)\right)\right).$$

Then, for every $K \geq 1$,

$$\min_{0 \leq k \leq K-1} \|G_\gamma(x_k)\|^2 \leq \frac{f(x_0) - f_\star}{\gamma \left(1 - \frac{L\gamma}{2}\right) K}.$$

In particular, for $\gamma = \frac{1}{L}$,

$$\min_{0 \leq k \leq K-1} \left\|G_{1/L}(x_k)\right\|^2 \leq \frac{2L\left(f(x_0) - f_\star\right)}{K}.$$

Moreover, $G_\gamma(x_k) \to 0$ as $k \to \infty$. Consequently, every accumulation point of $\{x_k\}$ is a first-order stationary point.

Proof. The result is an immediate specialization of Beck (2017, Theorem 10.15) to the composite problem

$$F(x) = f(x) + \iota_{\mathcal{X}}(x),$$

where $\iota_{\mathcal{X}}$ denotes the indicator function of $\mathcal{X}$. In Beck (2017)'s notation, the non-smooth term is $g(x) \equiv \iota_{\mathcal{X}}(x)$, and the proximal-gradient update with constant parameter $\bar{L} \geq L$ is

$$x_{k+1} = \operatorname{prox}_{\iota_{\mathcal{X}}/\bar{L}}\left(x_k - \frac{1}{\bar{L}}\nabla f(x_k)\right) = \operatorname{Proj}_{\mathcal{X}}\left(x_k - \frac{1}{\bar{L}}\nabla f(x_k)\right).$$

Choosing $\bar{L} = 1/\gamma$, we recover exactly the ProjGD iteration. Notice that gradient mapping in Beck (2017) is

$$G_{\bar{L}}(x) = \bar{L}\left(x - \operatorname{Proj}_{\mathcal{X}}\left(x - \frac{1}{\bar{L}}\nabla f(x)\right)\right) = \frac{1}{\gamma}\left(x - \operatorname{Proj}_{\mathcal{X}}\left(x - \gamma\nabla f(x)\right)\right) = G_\gamma(x).$$

Therefore, (Beck, 2017, Theorem 10.15) yields $G_\gamma(x_k) \to 0$, every accumulation point of $\{x_k\}$ is stationary, and

$$\min_{0 \leq k \leq K-1} \|G_\gamma(x_k)\|^2 \leq \frac{f(x_0) - f_\star}{MK},$$

where, in the constant-stepsize case, $M = \frac{\bar{L} - L/2}{\bar{L}^2}$. Substituting $\bar{L} = 1/\gamma$ gives $M = \gamma\left(1 - L\gamma/2\right)$, and hence

$$\min_{0 \leq k \leq K-1} \|G_\gamma(x_k)\|^2 \leq \frac{f(x_0) - f_\star}{\gamma \left(1 - \frac{L\gamma}{2}\right) K}.$$

Setting $\gamma = 1/L$ yields

$$\min_{0 \leq k \leq K-1} \left\|G_{1/L}(x_k)\right\|^2 \leq \frac{2L\left(f(x_0) - f_\star\right)}{K}.$$

□

F STOCHASTIC BROXGD: COMPONENTWISE GEOMETRY AND CONVERGENCE

We now extend the algorithm to the stochastic setting. In what follows, we assume that f admits a finite-sum structure of the form

$$f(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where $n \geq 1$. Throughout this section, $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty, closed, and convex, and each f_i is convex and differentiable on a common open convex neighborhood U of $\mathcal{X}$. Assume that f and every f_i attain their minima over $\mathcal{X}$. Fix a deterministic $x_0 \in \mathcal{X}$, a minimizer $x_\star$ of f, and component minimizers $x_{\star,i}$; write $H_\star = \max_i \|x_{\star,i} - x_\star\|$.

Sampling and oracle conventions. Let $\mathcal{F}_k$ contain all randomness revealed before drawing i_k, including previous oracle selections, so that x_k is $\mathcal{F}_k$-measurable. Require $\mathbb{P}(i_k = i \mid \mathcal{F}_k) = 1/n$ for every i. Radii and exact oracle outputs are measurable selections based on the available information; any additional selection randomness is included in $\mathcal{F}_{k+1}$. If $t_k = 0$, the local feasible set is the singleton $\{x_k\}$, so $x_{k+1} = x_k$. Sampling continues at subsequent iterations: a zero step for one component does not terminate the algorithm. These are the common setting and conventions for all results below. Write $g_{k,i} = \nabla f_i(x_k)$ for the gradient of component i at iteration k. The resulting method is formalized in Algorithm 3.

Algorithm 3 Stochastic BroxGD

1: **Input:** starting point $x_0 \in \mathcal{X}$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample i_k uniformly from $\{1, \ldots, n\}$ conditionally on $\mathcal{F}_k$
4: Compute the gradient $g_{k,i_k} \coloneqq \nabla f_{i_k}(x_k)$
5: Choose a radius $t_k \geq 0$
6: Compute the next iterate as the solution of the local linear minimization problem

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle g_{k,i_k}, z \rangle$$

7: **end for**

For each radius policy analyzed below, positive-radius admissibility implies $t_k \leq \|x_k - x_{\star,i_k}\|$, and the same bound holds when $t_k = 0$. Consequently, $\|x_{k+1} - x_\star\| \leq 2\|x_k - x_\star\| + H_\star$. Starting from deterministic x_0, this gives a deterministic bound at every finite horizon. Continuity on the resulting compact feasible sets ensures that the expectations used below are finite.

F.1 CONVEX OBJECTIVE WITH BOUNDED GRADIENTS

We first present the theory for the convex case with bounded gradients. In the subsequent analysis, we assume that each component function f_i is differentiable and admits at least one constrained minimizer. We introduce the following notation. For each $i \in \{1, \ldots, n\}$, let

$$x_{\star,i} \in \operatorname*{argmin}_{x \in \mathcal{X}} f_i(x), \qquad f_i^\star \coloneqq f_i(x_{\star,i}) = \min_{x \in \mathcal{X}} f_i(x).$$

As in the proof of Theorem D.7, we begin with a stochastic version of the Type-I one-step descent property established in Theorem 3.1.

Theorem F.1 (One-step behavior w.r.t. the sampled component minimizer: Type-I case)**.** Use the common setting of Appendix F. Fix $k \geq 0$, and condition on one realization of i_k. Assume that

$$g_{k,i_k} \neq 0, \qquad 0 < t_k \leq \frac{\langle g_{k,i_k}, x_k - x_{\star,i_k} \rangle}{\|g_{k,i_k}\|}.$$

Let

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle g_{k,i_k}, z \rangle .$$

Then

$$t_k \le \|x_k - x_{\star,i_k}\|, \qquad \|x_{k+1} - x_k\| = t_k, \qquad \|x_{k+1} - x_{\star,i_k}\|^2 \le \|x_k - x_{\star,i_k}\|^2 - t_k^2.$$

Proof. Since $x_{\star,i_k} \in \mathcal{X}$ and the admissibility condition is stated with respect to $x_{\star,i_k}$, the proof is identical to the deterministic Type-I argument of Theorem 3.1, with $x_{\star,i_k}$ in place of $x_\star$. Indeed, the argument uses only the facts that $x_{\star,i_k} \in \mathcal{X}$ and

$$0 < t_k \le \frac{\langle g_{k,i_k}, x_k - x_{\star,i_k} \rangle}{\|g_{k,i_k}\|}.$$

Applying the same reasoning therefore yields

$$t_k \le \|x_k - x_{\star,i_k}\|, \qquad \|x_{k+1} - x_k\| = t_k, \qquad \|x_{k+1} - x_{\star,i_k}\|^2 \le \|x_k - x_{\star,i_k}\|^2 - t_k^2.$$

□

The following result provides a stochastic extension of Theorem D.7.

Theorem F.2 (Component Polyak radii with bounded gradients)**.** Use the common setting of Appendix F. Define $H_\star := \max_{1 \le i \le n} \|x_{\star,i} - x_\star\|$ and let $\{x_k\}_{k \ge 0}$ be generated by Algorithm 3 with

$$t_k := \begin{cases} \dfrac{f_{i_k}(x_k) - f_{i_k}^\star}{\|g_{k,i_k}\|}, & g_{k,i_k} \neq 0, \\ 0, & g_{k,i_k} = 0. \end{cases}$$

Assume $G > 0$ and $\|\nabla f_i(x)\| \le G$ for all i and $x \in \mathcal{X}$. This radius requires the component optimal values $f_i^\star$. Then, for every $\eta \in (0,1)$ and every $K \ge 1$,

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left(f(x_k) - f(x_\star) \right)^2 \right] \le \frac{G^2 \|x_0 - x_\star\|^2}{(1-\eta)\, K} + \frac{G^2 {H_\star}^2}{\eta\,(1-\eta)}.$$

Moreover, for the averaged iterate $\bar{x}_K := \frac{1}{K} \sum_{k=0}^{K-1} x_k$, we have

$$\mathbb{E}\left[f(\bar{x}_K) - f(x_\star) \right] \le \sqrt{\frac{G^2 \|x_0 - x_\star\|^2}{(1-\eta)\, K} + \frac{G^2 {H_\star}^2}{\eta\,(1-\eta)}}.$$

Proof. Fix $k \ge 0$. We first verify that the chosen radius is Type-I admissible with respect to $x_{\star,i_k}$ whenever it is positive.

If $g_{k,i_k} = 0$, convexity gives $f_{i_k}(x_k) = f_{i_k}^\star$; the radius and the step are zero, so the component distance-decrease inequality below holds with equality. For a nonzero sampled gradient, proceed as follows.

Assume now that $g_{k,i_k} \neq 0$. Since $x_{\star,i_k} \in \operatorname{argmin}_{x \in \mathcal{X}} f_{i_k}(x)$, we have $f_{i_k}(x_{\star,i_k}) = f_{i_k}^\star$. Hence

$$t_k = \frac{f_{i_k}(x_k) - f_{i_k}(x_{\star,i_k})}{\|g_{k,i_k}\|}.$$

By convexity of f_{i_k},

$$f_{i_k}(x_k) - f_{i_k}(x_{\star,i_k}) \le \langle \nabla f_{i_k}(x_k), x_k - x_{\star,i_k} \rangle = \langle g_{k,i_k}, x_k - x_{\star,i_k} \rangle .$$

Therefore

$$0 \le t_k \le \frac{\langle g_{k,i_k}, x_k - x_{\star,i_k} \rangle}{\|g_{k,i_k}\|}.$$

If $t_k > 0$, then Theorem F.1 applies and yields

$$\|x_{k+1} - x_{\star,i_k}\|^2 \leq \|x_k - x_{\star,i_k}\|^2 - t_k^2.$$

If $t_k = 0$, then $x_{k+1} = x_k$, and the same inequality still holds. Thus, in all cases,

$$\|x_{k+1} - x_{\star,i_k}\|^2 \leq \|x_k - x_{\star,i_k}\|^2 - t_k^2.$$

We now rewrite this inequality with respect to the global minimizer $x_\star$. Expanding both sides gives

$$\begin{aligned}\|x_{k+1} - x_\star\|^2 - 2\langle x_{k+1} - x_\star, x_{\star,i_k} - x_\star\rangle + \|x_{\star,i_k} - x_\star\|^2 \\ \leq \|x_k - x_\star\|^2 - 2\langle x_k - x_\star, x_{\star,i_k} - x_\star\rangle + \|x_{\star,i_k} - x_\star\|^2 - t_k^2.\end{aligned}$$

After cancellation, we obtain

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2 + 2\langle x_{k+1} - x_k, x_{\star,i_k} - x_\star\rangle.$$

Since $\|x_{k+1} - x_k\| = t_k$ when $t_k > 0$, and trivially also when $t_k = 0$, we have $\|x_{k+1} - x_k\| = t_k$ in all cases. Therefore, by Cauchy-Schwarz and the definition of $H_\star$,

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2 + 2H_\star t_k.$$

Fix any $\eta \in (0, 1)$. By Young's inequality,

$$2H_\star t_k \leq \eta t_k^2 + \frac{{H_\star}^2}{\eta}.$$

Hence

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - (1-\eta)\, t_k^2 + \frac{{H_\star}^2}{\eta}.$$

Taking conditional expectation with respect to the filtration $\mathcal{F}_k$ specified in Appendix F, we get

$$\mathbb{E}\left[\|x_{k+1} - x_\star\|^2 \mid \mathcal{F}_k\right] \leq \|x_k - x_\star\|^2 - (1-\eta)\,\mathbb{E}\left[t_k^2 \mid \mathcal{F}_k\right] + \frac{{H_\star}^2}{\eta}.$$

We next lower bound $\mathbb{E}\left[t_k^2 \mid \mathcal{F}_k\right]$. If $g_{k,i_k} \neq 0$, then by the definition of t_k and the gradient bound,

$$t_k^2 = \frac{\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2}{\|g_{k,i_k}\|^2} \geq \frac{\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2}{G^2}.$$

If $g_{k,i_k} = 0$, then $\nabla f_{i_k}(x_k) = 0$. Since f_{i_k} is convex and differentiable, x_k is a global minimizer of f_{i_k} over $\mathcal{X}$, and hence

$$f_{i_k}(x_k) = f_{i_k}^\star.$$

Therefore, in this case as well,

$$t_k^2 = 0 = \frac{\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2}{G^2}.$$

Thus, in all cases,

$$t_k^2 \geq \frac{\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2}{G^2}.$$

Taking conditional expectation yields

$$\mathbb{E}\left[t_k^2 \mid \mathcal{F}_k\right] \geq \frac{\mathbb{E}\left[\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2 \mid \mathcal{F}_k\right]}{G^2}.$$

Since i_k is sampled uniformly and independently of the past,

$$\mathbb{E}\left[f_{i_k}(x_k) - f_{i_k}^\star \mid \mathcal{F}_k\right] = \frac{1}{n}\sum_{i=1}^n \left(f_i(x_k) - f_i^\star\right).$$

By Jensen's inequality for the convex map $u \mapsto u^2$, we obtain

$$\mathbb{E}\left[\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2 \mid \mathcal{F}_k\right] \geq \left(\frac{1}{n}\sum_{i=1}^n \left(f_i(x_k) - f_i^\star\right)\right)^2.$$

Since $f_i^\star \leq f_i(x_\star)$ for every i, we have

$$\frac{1}{n}\sum_{i=1}^n \left(f_i(x_k) - f_i^\star\right) \geq \frac{1}{n}\sum_{i=1}^n \left(f_i(x_k) - f_i(x_\star)\right) = f(x_k) - f(x_\star).$$

Moreover, since $x_\star$ minimizes f over $\mathcal{X}$, we have $f(x_k) - f(x_\star) \geq 0$. Therefore

$$\mathbb{E}\left[\left(f_{i_k}(x_k) - f_{i_k}^\star\right)^2 \mid \mathcal{F}_k\right] \geq \left(f(x_k) - f(x_\star)\right)^2.$$

Consequently,

$$\mathbb{E}\left[t_k^2 \mid \mathcal{F}_k\right] \geq \frac{\left(f(x_k) - f(x_\star)\right)^2}{G^2}.$$

Substituting this into the previous conditional inequality yields

$$\mathbb{E}\left[\|x_{k+1} - x_\star\|^2 \mid \mathcal{F}_k\right] \leq \|x_k - x_\star\|^2 - \frac{1-\eta}{G^2}\left(f(x_k) - f(x_\star)\right)^2 + \frac{{H_\star}^2}{\eta}.$$

Taking expectation again and using the tower property, we get

$$\mathbb{E}\left[\|x_{k+1} - x_\star\|^2\right] \leq \mathbb{E}\left[\|x_k - x_\star\|^2\right] - \frac{1-\eta}{G^2}\mathbb{E}\left[\left(f(x_k) - f(x_\star)\right)^2\right] + \frac{{H_\star}^2}{\eta}.$$

Summing from $k = 0$ to $K - 1$, we obtain

$$\mathbb{E}\left[\|x_K - x_\star\|^2\right] \leq \|x_0 - x_\star\|^2 - \frac{1-\eta}{G^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(f(x_k) - f(x_\star)\right)^2\right] + \frac{K{H_\star}^2}{\eta}.$$

Since the left-hand side is nonnegative,

$$\frac{1-\eta}{G^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(f(x_k) - f(x_\star)\right)^2\right] \leq \|x_0 - x_\star\|^2 + \frac{K{H_\star}^2}{\eta}.$$

Dividing by K yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left(f(x_k) - f(x_\star)\right)^2\right] \leq \frac{G^2\|x_0 - x_\star\|^2}{(1-\eta)\,K} + \frac{G^2{H_\star}^2}{\eta\,(1-\eta)}. \tag{175}$$

Now define $\bar{x}_K := \frac{1}{K}\sum_{k=0}^{K-1} x_k$. By convexity of f,

$$f(\bar{x}_K) - f(x_\star) \leq \frac{1}{K}\sum_{k=0}^{K-1}\left(f(x_k) - f(x_\star)\right).$$

Applying Jensen's inequality to the convex map $u \mapsto u^2$, we obtain

$$\left(f(\bar{x}_K) - f(x_\star)\right)^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\left(f(x_k) - f(x_\star)\right)^2.$$

Taking expectation and using (175), we conclude that

$$\mathbb{E}\left[\left(f(\bar{x}_K) - f(x_\star)\right)^2\right] \overset{(175)}{\leq} \frac{G^2\|x_0 - x_\star\|^2}{(1-\eta)\,K} + \frac{G^2{H_\star}^2}{\eta\,(1-\eta)}.$$

Finally, by Cauchy-Schwarz,

$$\mathbb{E}\left[f(\bar{x}_K) - f(x_\star)\right] \leq \sqrt{\mathbb{E}\left[\left(f(\bar{x}_K) - f(x_\star)\right)^2\right]} \leq \sqrt{\frac{G^2\|x_0 - x_\star\|^2}{(1-\eta)\,K} + \frac{G^2{H_\star}^2}{\eta\,(1-\eta)}}.$$

This completes the proof. □

Remark F.3 (Sanity check)**.** *In the single-client setting, we have $x_{\star,i} = x_\star$ and hence $H_\star = 0$. Therefore, the additional neighborhood term vanishes, and we recover the result of Theorem D.7 up to a constant factor, depending on the choice of η when applying Young's inequality.*

The same conclusion holds in the interpolation regime, where there exists a common constrained minimizer $x_\star$ of all functions f_i. In this case, we may take $x_{\star,i} = x_\star$ for all i, and hence $H_\star = 0$ again. Thus, the additional neighborhood term disappears, and the stochastic result reduces to the deterministic bounded-gradient result up to the same constant-factor difference.

F.2 Smooth objectives and Type-II admissibility

For smooth objectives, we will rely on the stochastic extension of Type II admissibility condition in Theorem 3.1.

Theorem F.4 (One-step behavior of stochastic BroxGD: Type-II case)**.** Use the common setting of Appendix F. Let $\{x_k\}_{k\geq 0} \subseteq \mathcal{X}$ be generated by

$$x_{k+1} \in \operatorname*{argmin}_{z\in\mathcal{X}\cap\mathcal{B}(x_k,t_k)} \langle \nabla f_{i_k}(x_k), z\rangle\,,$$

where $i_k \in \{1,\dots,n\}$. Assume that

$$\nabla f_{i_k}(x_k) \neq \nabla f_{i_k}(x_{\star,i_k}), \quad 0 < t_k \leq \frac{\langle \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k}), x_k - x_{\star,i_k}\rangle}{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|}.$$

Then[a]

$$\|x_{k+1} - x_{\star,i_k}\|^2 \leq \|x_k - x_{\star,i_k}\|^2 - t_k^2.$$

[a] Consequently,

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2 + 2H_\star t_k,$$

where $H_\star = \max_{1\leq i\leq n} \|x_{\star,i} - x_\star\|$

Proof. Fix $k \geq 0$. Since $x_{\star,i_k} \in \operatorname{argmin}_{x\in\mathcal{X}} f_{i_k}(x)$, Theorem 3.1(iii) applied to the component objective f_{i_k} yields

$$\|x_{k+1} - x_{\star,i_k}\|^2 \leq \|x_k - x_{\star,i_k}\|^2 - t_k^2.$$

This proves the main claim. For the consequence,

$$\begin{aligned}
\|x_{k+1} - x_\star\|^2 &= \|x_{k+1} - x_{\star,i_k} + x_{\star,i_k} - x_\star\|^2 \\
&= \|x_{k+1} - x_{\star,i_k}\|^2 + 2\langle x_{k+1} - x_{\star,i_k}, x_{\star,i_k} - x_\star\rangle + \|x_{\star,i_k} - x_\star\|^2 \\
&\leq \|x_k - x_{\star,i_k}\|^2 - t_k^2 + 2\langle x_{k+1} - x_{\star,i_k}, x_{\star,i_k} - x_\star\rangle + \|x_{\star,i_k} - x_\star\|^2 \\
&= \|x_k - x_\star\|^2 - t_k^2 + 2\langle x_{k+1} - x_k, x_{\star,i_k} - x_\star\rangle\,.
\end{aligned}$$

Since $x_{k+1} \in \mathcal{B}(x_k, t_k)$, we have $\|x_{k+1} - x_k\| \leq t_k$. Therefore,

$$2\langle x_{k+1} - x_k, x_{\star,i_k} - x_\star\rangle \leq 2\|x_{k+1} - x_k\|\|x_{\star,i_k} - x_\star\| \leq 2H_\star t_k.$$

Substituting this into the previous inequality gives

$$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2 + 2H_\star t_k.$$

□

Strongly convex case. Under the additional assumption of strong convexity, Theorem F.4 yields the following result.

Theorem F.5 (Linear convergence to a neighborhood)**.** Use the common setting of Appendix F, and let the iterates be generated by Algorithm 3. Assume for each $i \in \{1,\dots,n\}$, the function f_i is μ_i-strongly convex and L_i-smooth on an open convex set containing $\mathcal{X}$, with $0 < \mu_i \leq L_i$.

Define

$$\theta_i := \frac{2\sqrt{\mu_i L_i}}{L_i + \mu_i}, \qquad \rho_i := \frac{L_i - \mu_i}{L_i + \mu_i} = \sqrt{1 - \theta_i^2}, \qquad \bar{\rho} := \frac{1}{n}\sum_{i=1}^{n} \rho_i.$$

Choose $t_k := \theta_{i_k}\|x_k - x_{\star,i_k}\|$. At $k = 0$, interpret $\bar{\rho}^0 = 1$, including when $\bar{\rho} = 0$. Then, for every $k \geq 0$,

$$\|x_{k+1} - x_{\star,i_k}\|^2 \leq \rho_{i_k}^2 \|x_k - x_{\star,i_k}\|^2. \tag{176}$$

Consequently, for every $k \geq 0$,

$$\mathbb{E}\left[\|x_k - x_\star\|\right] \leq \bar{\rho}^k \|x_0 - x_\star\| + \frac{1+\bar{\rho}}{1-\bar{\rho}} H_\star.$$

Proof. Fix $k \geq 0$. If $x_k = x_{\star,i_k}$, then $t_k = 0$, and (176) is trivial. So assume $x_k \neq x_{\star,i_k}$. Strong convexity ensures that the gradient difference is nonzero. Applying Lemma C.3 on the open convex domain of f_{i_k} yields

$$\theta_{i_k}\|x_k - x_{\star,i_k}\| \overset{(56)}{\leq} \frac{\langle \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k}), x_k - x_{\star,i_k}\rangle}{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|}.$$

Since the assumptions of Theorem F.4 are satisfied, we obtain

$$\begin{aligned}
\|x_{k+1} - x_{\star,i_k}\|^2 &\leq \|x_k - x_{\star,i_k}\|^2 - \theta_{i_k}^2 \|x_k - x_{\star,i_k}\|^2 \\
&= \left(1 - \theta_{i_k}^2\right) \|x_k - x_{\star,i_k}\|^2 \\
&= \rho_{i_k}^2 \|x_k - x_{\star,i_k}\|^2.
\end{aligned}$$

This proves (176). Taking square roots, we get $\|x_{k+1} - x_{\star,i_k}\| \leq \rho_{i_k}\|x_k - x_{\star,i_k}\|$. Using the triangle inequality twice,

$$\begin{aligned}
\|x_{k+1} - x_\star\| &\leq \|x_{k+1} - x_{\star,i_k}\| + \|x_{\star,i_k} - x_\star\| \\
&\overset{(176)}{\leq} \rho_{i_k}\|x_k - x_{\star,i_k}\| + \|x_{\star,i_k} - x_\star\| \\
&\leq \rho_{i_k}\left(\|x_k - x_\star\| + \|x_\star - x_{\star,i_k}\|\right) + \|x_{\star,i_k} - x_\star\| \\
&= \rho_{i_k}\|x_k - x_\star\| + \left(1 + \rho_{i_k}\right)\|x_{\star,i_k} - x_\star\|.
\end{aligned}$$

Since i_k is sampled uniformly from $\{1, \ldots, n\}$, taking conditional expectation with respect to $\mathcal{F}_k$ yields

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x_\star\| \mid \mathcal{F}_k\right] &\leq \mathbb{E}\left[\rho_{i_k}\|x_k - x_\star\| + \left(1 + \rho_{i_k}\right)\|x_{\star,i_k} - x_\star\| \mid \mathcal{F}_k\right] \\
&= \bar{\rho}\|x_k - x_\star\| + \frac{1}{n}\sum_{i=1}^{n}\left(1 + \rho_i\right)\|x_{\star,i} - x_\star\|.
\end{aligned}$$

Taking expectation again and using tower property, we obtain

$$\mathbb{E}\left[\|x_{k+1} - x_\star\|\right] \leq \bar{\rho}\,\mathbb{E}\left[\|x_k - x_\star\|\right] + \frac{1}{n}\sum_{i=1}^{n}\left(1 + \rho_i\right)\|x_{\star,i} - x_\star\|.$$

Iterating this recursion yields

$$\mathbb{E}\left[\|x_k - x_\star\|\right] \leq \bar{\rho}^k\|x_0 - x_\star\| + \frac{1 - \bar{\rho}^k}{1 - \bar{\rho}}\frac{1}{n}\sum_{i=1}^{n}\left(1 + \rho_i\right)\|x_{\star,i} - x_\star\|.$$

Finally, since $\|x_{\star,i} - x_\star\| \leq H_\star$ for every i,

$$\frac{1}{n}\sum_{i=1}^{n}\left(1 + \rho_i\right)\|x_{\star,i} - x_\star\| \leq \frac{1}{n}\sum_{i=1}^{n}\left(1 + \rho_i\right)H_\star = \left(1 + \bar{\rho}\right)H_\star,$$

and therefore

$$\mathbb{E}\left[\|x_k - x_\star\|\right] \leq \bar{\rho}^k\|x_0 - x_\star\| + \frac{1+\bar{\rho}}{1-\bar{\rho}} H_\star.$$

□

Convex case. In the absence of strong convexity, the following result applies.

Theorem F.6. Use the common setting of Appendix F, and let the iterates be generated by Algorithm 3. Moreover, assume that for each $i \in \{1, \dots, n\}$, the function f_i is convex and L_i-smooth on an open convex set containing $\mathcal{X}$, with $L_i > 0$. Define

$$L_{\max} := \max_{1 \le i \le n} L_i, \qquad \bar{L}_2^2 := \frac{1}{n} \sum_{i=1}^{n} L_i^2.$$

Suppose that i_k is sampled uniformly from $\{1, \dots, n\}$ and choose

$$t_k := \begin{cases} \dfrac{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|}{L_{i_k}}, & \nabla f_{i_k}(x_k) \neq \nabla f_{i_k}(x_{\star,i_k}), \\ 0, & \nabla f_{i_k}(x_k) = \nabla f_{i_k}(x_{\star,i_k}). \end{cases}$$

Then, for every $K \ge 1$,[a]

$$\min_{0 \le k \le K-1} \mathbb{E}\|\nabla f(x_k) - \nabla f(x_\star)\| \le \sqrt{\frac{4L_{\max}^2 \|x_0 - x_\star\|^2}{K} + \left(8L_{\max}^2 + 2\bar{L}_2^2\right) {H_\star}^2}.$$

[a] In the common-minimizer regime $x_{\star,i} = x_\star$ for all i, we have $H_\star = 0$, and therefore

$$\min_{0 \le k \le K-1} \mathbb{E}\|\nabla f(x_k) - \nabla f(x_\star)\| \le \frac{2L_{\max}\|x_0 - x_\star\|}{\sqrt{K}}.$$

Proof. Fix $k \ge 0$. If $\nabla f_{i_k}(x_k) = \nabla f_{i_k}(x_{\star,i_k})$, then $t_k = 0$ and $x_{k+1} = x_k$. The first distance recurrence below therefore holds with equality on this event. It remains to establish that recurrence for a nonzero sampled residual. Assume now that $\nabla f_{i_k}(x_k) \neq \nabla f_{i_k}(x_{\star,i_k})$. Since f_{i_k} is convex and L_{i_k}-smooth, cocoercivity gives

$$\frac{1}{L_{i_k}} \|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|^2 \le \langle \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k}), x_k - x_{\star,i_k} \rangle .$$

Hence

$$t_k = \frac{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|}{L_{i_k}} \le \frac{\langle \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k}), x_k - x_{\star,i_k} \rangle}{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|}.$$

Therefore the assumptions of Theorem F.4 are satisfied, and we obtain

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - \frac{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|^2}{L_{i_k}^2} + \frac{2H_\star}{L_{i_k}} \|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|.$$

The preceding recurrence now holds for every sampled component, including those with zero residual. By Young's inequality,

$$\frac{2H_\star}{L_{i_k}} \|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\| \le \frac{1}{2} \frac{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|^2}{L_{i_k}^2} + 2{H_\star}^2.$$

Hence

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - \frac{1}{2} \frac{\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_{\star,i_k})\|^2}{L_{i_k}^2} + 2{H_\star}^2.$$

Taking conditional expectation with respect to $\mathcal{F}_k$ yields

$$\mathbb{E}\left[\|x_{k+1} - x_\star\|^2 \mid \mathcal{F}_k\right] \le \|x_k - x_\star\|^2 - \frac{1}{2n} \sum_{i=1}^{n} \frac{\|\nabla f_i(x_k) - \nabla f_i(x_{\star,i})\|^2}{L_i^2} + 2{H_\star}^2.$$

Taking expectation again, using the tower property and summing from $k = 0$ to $K - 1$, we get

$$\mathbb{E}\left[\|x_K - x_\star\|^2\right] \le \|x_0 - x_\star\|^2 - \frac{1}{2} \sum_{k=0}^{K-1} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \frac{\|\nabla f_i(x_k) - \nabla f_i(x_{\star,i})\|^2}{L_i^2}\right] + 2K{H_\star}^2.$$

Since $\mathbb{E}\left[\|x_K - x_\star\|^2\right] \geq 0$, it follows that

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\|\nabla f_i(x_k) - \nabla f_i(x_{\star,i})\|^2}{L_i^2}\right] \leq \frac{2\|x_0 - x_\star\|^2}{K} + 4{H_\star}^2. \tag{177}$$

For each component, L_i-smoothness and the definition of $H_\star$ give

$$\|\nabla f_i(x_{\star,i}) - \nabla f_i(x_\star)\| \leq L_i\|x_{\star,i} - x_\star\| \leq L_i D. \tag{178}$$

Splitting the gradient difference at $x_{\star,i}$ and using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ yields

$$\begin{aligned}\|\nabla f_i(x_k) - \nabla f_i(x_\star)\|^2 &\overset{(178)}{\leq} 2\|\nabla f_i(x_k) - \nabla f_i(x_{\star,i})\|^2 + 2L_i^2{H_\star}^2 \\ &\leq 2L_{\max}^2\frac{\|\nabla f_i(x_k) - \nabla f_i(x_{\star,i})\|^2}{L_i^2} + 2L_i^2{H_\star}^2.\end{aligned} \tag{179}$$

Crucially, retain the individual L_i^2 in the second term until averaging. Since ∇f is the average component gradient, Jensen's inequality gives

$$\begin{aligned}\|\nabla f(x_k) - \nabla f(x_\star)\|^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_k) - \nabla f_i(x_\star)\|^2 \\ &\overset{(179)}{\leq} \frac{2L_{\max}^2}{n}\sum_{i=1}^{n}\frac{\|\nabla f_i(x_k) - \nabla f_i(x_{\star,i})\|^2}{L_i^2} + 2\bar{L}_2^2{H_\star}^2.\end{aligned} \tag{180}$$

Taking expectations and averaging over $k = 0, \ldots, K-1$ now gives

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_k) - \nabla f(x_\star)\|^2 &\overset{(180),(177)}{\leq} 2L_{\max}^2\left(\frac{2\|x_0 - x_\star\|^2}{K} + 4{H_\star}^2\right) + 2\bar{L}_2^2{H_\star}^2 \\ &= \frac{4L_{\max}^2\|x_0 - x_\star\|^2}{K} + (8L_{\max}^2 + 2\bar{L}_2^2){H_\star}^2.\end{aligned} \tag{181}$$

Finally, choose an index minimizing the expected squared residual and apply Cauchy–Schwarz:

$$\begin{aligned}\min_{0\leq k<K}\mathbb{E}\|\nabla f(x_k) - \nabla f(x_\star)\| &\leq \sqrt{\min_{0\leq k<K}\mathbb{E}\|\nabla f(x_k) - \nabla f(x_\star)\|^2} \\ &\overset{(181)}{\leq} \sqrt{\frac{4L_{\max}^2\|x_0 - x_\star\|^2}{K} + (8L_{\max}^2 + 2\bar{L}_2^2){H_\star}^2}.\end{aligned}$$

This is the stated bound. □

Remark F.7 (Sanity check). *When all components share a constrained minimizer, choose $x_{\star,i} = x_\star$, so $H_\star = 0$ and both neighborhood terms vanish. For a single component, Theorem F.5 gives $\|x_k - x_\star\| \leq \rho^k\|x_0 - x_\star\|$, the square-root form of the squared-distance guarantee in Theorem 3.4. In the same single-component setting, Theorem F.6 gives a best gradient-residual norm bounded by $2L\|x_0 - x_\star\|/\sqrt{K}$, whereas taking square roots in Theorem D.11 gives $L\|x_0 - x_\star\|/\sqrt{K}$. Thus these norm bounds have the same order, with a factor of two in the smooth-convex comparison. For multiple components with $H_\star = 0$, the stated guarantees retain their stochastic expectation and component-dependent constants.*

G EXPERIMENTS AND NUMERICAL ILLUSTRATIONS

We first study an explicit structured-label problem, then report a controlled study across all five application families in Appendix H. A selected fair-ranking example comes from that broader study; Appendix G.3 gives the full correlation-matrix study summarized in the main text. The remaining two-dimensional quadratic examples illustrate the geometry of the method, compare convergence with theoretical predictions, and examine sensitivity to the radius schedule. They are not intended to establish large-scale or real-world performance. After the application study, we specify the shared box-constrained instance, then study its trajectories and convergence, compare methods and radius schedules, and finally illustrate alternative constraint geometries.

G.1 FINITE-LABEL STRUCTURED QUADRATIC: PROTOCOL AND ORACLE COSTS

We instantiate the regularized quadratic over a product of simplices in (213)–(214). This is a synthetic, explicitly enumerated instance with $N = 8$ blocks and 24 labels per block (192 variables), not a task with exponentially many implicit labels. We choose it because the local oracle can retain active nonnegativity constraints and supply a numerical certificate throughout the run.

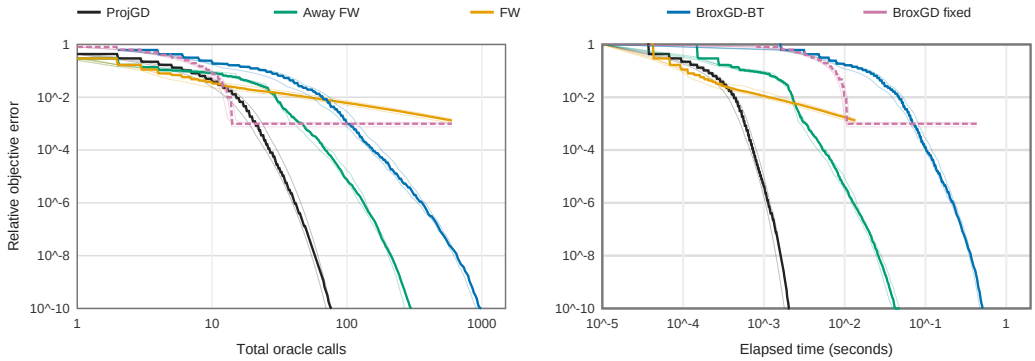


Figure 2: Relative objective error versus total oracle calls and elapsed time. Thick curves are medians over three seeds; thin curves show individual seeds. Times use three repeats per seed. Local-oracle counts include rejected backtracking trials; runtime includes their inner solves. BroxGD fixed reports best-so-far error. Oracle types differ across methods.

Instances and reference optimum. For seeds $0, 1, 2$, a QR factorization of a standard Gaussian matrix supplies an orthogonal matrix V. Set $Q = V\,\mathrm{diag}(q)V^\top$, with 192 geometrically spaced eigenvalues between 0.011 and 1.01. In each block, choose three support coordinates uniformly without replacement and assign Dirichlet$(1, 1, 1)$ weights to form $\alpha_\star$. Independently draw c_j uniformly from $[0.005, 0.03]$ off this support and set $c_j = 0$ on it. The tested objective, up to an irrelevant additive constant, is

$$f(\alpha) = \tfrac{1}{2}\alpha^\top Q\alpha + (c - Q\alpha_\star)^\top \alpha, \qquad \alpha \in (\Delta_{24})^8. \tag{182}$$

Since $\nabla f(\alpha_\star) = c$ and $c^\top(\alpha - \alpha_\star) \geq 0$ on the domain, $\alpha_\star$ is the unique constrained minimizer. Thus the reference error has no optimization-solver uncertainty. The model is a special case of (214): take $\mu = 0.01$, $B = (Q - \mu I)^{1/2}$, and $b = B^{-\top}(Q\alpha_\star - c)$. Both these matrices are nonsingular. The optimum is used for constructing the instances and evaluating error, not for choosing steps or radii.

Methods and budgets. All methods start at the uniform distribution in every block. ProjGD uses stepsize $1/L$, $L = 1.01$, and sorting-based Euclidean simplex projection. FW and away-step FW use exact quadratic line search; their global oracle chooses a minimum-gradient coordinate in every block. Away-step FW starts from the uniform mixture of the 24 product vertices that choose the same label in all blocks, and maintains a convex decomposition. These are full-vector methods, not block-coordinate variants. BroxGD-BT (contraction factor $\xi = 1/2$) follows Algorithm 2, including

its prescribed restart radius at every accepted center. BroxGD fixed uses $t = R/\sqrt{K}$, $R = \sqrt{2N} = 4$, and $K = 600$, as in Theorem 3.2; R is the diameter bound, not an optimum-dependent input. Every run has at most 600 accepted updates and is truncated for evaluation when relative objective error reaches 10^{-10}. For BroxGD fixed, we report the best objective value seen so far, as required by its theorem; the other methods are monotone on these runs. The reported 10^{-6} target is a common evaluation threshold, not an implementable stopping certificate.

Local oracle and numerical certificate. For center x, gradient g, and radius t, first test whether a global minimizing vertex lies in the ball. Otherwise bracket and bisect $a > 0$ in $z(a) = \mathrm{Proj}_{\mathcal{X}}(x - ag)$, keeping the endpoint with $\|z(a) - x\| \le t$. Each projection sorts the coordinates of each block. For the returned feasible z and $\lambda = 1/a$, put $d = z - x$ and $q = g + \lambda d$. A computable upper bound on local linear suboptimality is

$$\epsilon_{\mathrm{loc}} = \left[\langle q, z \rangle - \sum_{i=1}^{N} \min_j q_{ij} \right]_+ + \frac{\lambda}{2} \left(t^2 - \|d\|^2 \right). \tag{183}$$

Indeed, for any feasible w in the local ball, expanding $\|w - x\|^2$ gives

$$\begin{aligned} \langle g, w - z \rangle &= \langle q, w - z \rangle + \tfrac{\lambda}{2} \left(\|d\|^2 - \|w - x\|^2 + \|w - z\|^2 \right) \\ &\overset{(183)}{\ge} -\epsilon_{\mathrm{loc}}. \end{aligned}$$

We stop bisection at $\epsilon_{\mathrm{loc}} \le 10^{-10} \max\{1, \|g\| t\}$. The largest recorded certificate is below 10^{-10}, and the maximum simplex-feasibility residual is below $2 \cdot 10^{-15}$. Tightening the tolerance to 10^{-12} changes the relative-error curves by less than $8 \cdot 10^{-9}$ at matching accepted indices. These are floating-point certificate checks, not interval-arithmetic guarantees; the exact-oracle theorems are not asserted for finite-tolerance solves. This implementation uses projections internally: the outer method's projection-free oracle formulation does not make this particular local solver projection-free.

Timing and interpretation. We use NumPy 2.3.5 in a single Python process on macOS 26.6.2 arm64, request one computational thread, warm up the methods, and take three timing repeats per seed. The runtime includes gradients, oracle solves, local certificates, rejected trials, away-step bookkeeping, and best-iterate objective evaluations for BroxGD fixed; it excludes common instance construction and reference-error logging. Figure 2 shows thin seed traces and their thick pointwise median; each seed's time trace is the median of three repeats. Oracle counts mean projections for ProjGD, global linear solves for FW variants, and local linear solves including rejected trials for BroxGD; these are different operations, so runtime and inner-projection counts are essential companions.

Table 2: Median work to relative error 10^{-6} across three seeds. A dash means the target was not reached within 600 accepted updates. Times are implementation-specific.

Method	Accepted	Oracle calls	Inner projections	Time (ms)
ProjGD	41	41	41	1.12
Away-step FW	133	133	0	12.64
FW	—	—	—	—
BroxGD-BT	58	357	9682	216.79
BroxGD fixed	—	—	—	—

At this target, BroxGD-BT takes 58–66 accepted updates but 346–388 local solves and 9435–10513 inner projections. The median terminal relative errors for FW and BroxGD fixed are $1.31 \cdot 10^{-3}$ and $9.87 \cdot 10^{-4}$, respectively. These results illustrate the difference between accepted-step convergence and computational efficiency. They favor ProjGD on explicit simplices and do not establish a practical advantage for BroxGD or a scaling benefit for implicit structured prediction. Faster certified local solvers remain an application-specific research question.

Reproducibility. The accompanying scripts `scripts/experiment_structured.py`, `scripts/run_structured_suite.py`, and `scripts/summarize_structured.py`

generate the instances, all runs, certificates, summaries, and the vector figure. Raw traces and timing repeats are in `output/experiments/structured-repeats.json`; the summary is generated directly from those traces.

G.2 A NUMERICAL STUDY ACROSS THE APPLICATION FAMILIES

We next study all five model families from Appendix H: ranking/transport, PSD-constrained covariance estimation, occupancy measures, robust low-rank matrix learning, and structured-label distributions. The purpose is to locate favorable regimes, not to assume that localization is faster. These are controlled synthetic optimization instances with certified reference solutions. They do not measure downstream prediction, fairness generalization, or reinforcement-learning performance, and do not cover every possible loss and constraint combination in those families. This study is separate from the dense-Hessian illustration in Appendix G.1.

Design and coverage. The primary grid contains 180 quadratic instances: five families, two sizes, two solution-activity regimes, three curvature ranges, and three seeds per configuration. It adds six ranking instances with a fairness equality and six robust Huber matrix instances. The resulting 192 instances produce 1080 method runs, since FW is unavailable on the unbounded PSD cone. The size pairs are 6×6 and 12×12 for ranking, 8 and 16 states with four actions for occupancy measures, four and eight blocks with 24 labels for structured distributions, and 8×8 and 24×24 for both matrix families. Curvature ratios are controlled by coefficients geometrically spaced between $1/\kappa$ and 1, for $\kappa \in \{3, 30, 300\}$. Matrix symmetrization can improve this nominal condition bound.

Each primary run has a cap of 400 accepted updates or three seconds, checked between updates. We record relative objective targets 10^{-2}, 10^{-4}, and 10^{-6}; reaching 10^{-6} ends a run. A capped run is censored, not evidence of nonconvergence. To confirm the promising regimes and retain negative controls, we use five new seeds, three timing repeats, and a larger cap of 2000 updates or five seconds on ten configurations, producing another 870 runs. These include ranking matrices of orders 12 and 24 with and without fairness equalities, 32- and 64-state occupancy models, nuclear-norm matrices of orders 32 and 64, and PSD and structured-label controls. The confirmation instances all use $\kappa = 30$. Exploratory runs, including simpler target constructions and a generic-LP occupancy baseline, are retained separately and are not pooled into these reported comparisons.

Reference solutions and objectives. Write $u = \mathrm{vec}(x - x_\star)$. The quadratic tests use

$$f(x) - f(x_\star) = \tfrac{1}{2} \sum_j w_j u_j^2 + \langle c, u \rangle, \qquad w_j > 0, \tag{184}$$

where c is constructed so that $\langle c, \mathrm{vec}(x - x_\star) \rangle \geq 0$ on the feasible set. This explicitly certifies the unique minimizer; it is used for evaluation, not by the algorithms. Such quadratics are weighted least-squares models with a linear term, and can be written as regularized regression objectives by completing the square. All methods start from the same feasible point.

For ranking, $x_\star$ is a random mixture of three permutation matrices; interior instances mix this plan with the uniform plan, while boundary instances retain zero entries. A fairness equality fixes the mass in a block of rows and columns to its value at the constructed optimum. For occupancy measures, transitions are random probability vectors and the discount is 0.9; the target occupancy comes from either a fully supported policy or a policy supported on two actions per state. Off-support nonnegative entries of c certify the boundary optimum in these two families and in the structured-label tests, whose boundary targets have three labels per block. Interior instances take $c = 0$.

PSD instances have a rotated full-rank or rank-deficient covariance target, with the matrix representation of c positive semidefinite and supported on the target's nullspace. The feasible set is the unbounded PSD cone; we do not silently add a trace constraint to enable FW. Nuclear-norm instances have a low-rank target with nuclear norm 0.4 or 1 in a unit nuclear-norm ball. At the boundary, $c = -0.03UV^\top$ for the target's singular subspaces, which is a supporting linear functional. The additional robust tests replace the quadratic term in (184) by weighted Huber penalties of threshold 0.03 and add a centered ridge term of coefficient $0.1/\kappa$; the supporting linear term and known optimum are preserved. They test robustness of the optimization loss, not statistical outlier recovery.

Algorithms and fair oracle implementations. We compare ProjGD with stepsize $1/L$, a safeguarded Barzilai–Borwein projected-gradient variant (ProjGD-BB), FW, away-step FW, BroxGD-BT from Algorithm 2 with $\xi = 1/2$, and the fixed-horizon BroxGD radius $R/\sqrt{K}$ from Theorem 3.2. ProjGD-BB clips its spectral stepsize to $[1/L, 1/\mu]$ and halves it until the monotone projected Armijo test with coefficient 10^{-4} holds. FW variants use exact line search for quadratics and bounded scalar minimization for Huber objectives. BroxGD fixed reports its best iterate and includes selection evaluations in its runtime. Radius bounds are domain diameters on compact sets; on the PSD cone we use a certified bound obtained from the unconstrained quadratic minimum and strong convexity, not the reference minimizer.

Ranking uses an assignment oracle without extra fairness constraints and a linear program when those equalities are present. Occupancy FW uses policy iteration, independently checked against linear programming, rather than relying on generic-LP overhead. Nuclear FW uses a leading singular pair; at order 64 we use a partial singular-value solver. Away-step methods maintain a convex decomposition, including the common feasible initial point as an initial atom. Decomposition work is timed.

Both BroxGD and ProjGD receive applicable feasibility shortcuts. On an affine slice, projection onto the affine space is accepted if it satisfies the inequalities, and the affine-ball linear minimizer is accepted if it is feasible. PSD feasibility is screened by Cholesky factorization, and the nuclear norm has a shared Frobenius-norm sufficient check. Otherwise, ranking and occupancy projection/local subproblems use a quadratic/conic solver; explicit simplices use sorting; matrix projections use eigenvalue or singular-value thresholding. The remaining local matrix and simplex problems use a bracketed projection path and a numerical optimality certificate. These implementations can use projections internally and are not evidence that a local solve is intrinsically cheaper than projection.

Numerical accuracy, timing, and reporting. We warm up each family, deterministically randomize method order, and request one computational thread. Timings include gradients, rejection tests, numerical certificates, inner solves, and decomposition bookkeeping; they exclude data generation, common feasible initialization, and reference-error logging. Affine factorizations are timed on first use, except in fairness-constrained ranking, where the common initialization already computes and retains the factorization and projection-solver setup. Thus these are optimization-loop timings, not end-to-end deployment costs. The experiments use NumPy 2.5.3, SciPy 1.18.1, Clarabel 0.11.1, and macOS arm64. Results are implementation-specific. Tables retain success counts beside timings; a median over successful cases must not be read as a paired speedup when the success subsets differ. The held-out table instead compares methods on matched seeds after taking the median of three timing repeats. Numerical certificates and feasibility residuals are recorded; exact-oracle convergence theorems are not claimed for floating-point solves.

Table 3: Primary grid at relative objective error 10^{-4}. Each cell is median execution time in milliseconds among successful runs, followed by successes out of 18 instances. Medians from different success subsets are not paired speedups. A dash denotes no success; n/a means no global FW oracle on the unbounded cone.

Family	Optimum	ProjGD	ProjGD-BB	FW	Away FW	BroxGD fixed	BroxGD-BT
Ranking	interior	0.34 (18/18)	0.26 (18/18)	0.27 (14/18)	1.01 (17/18)	— (0/18)	1.71 (15/18)
Ranking	boundary	4.02 (18/18)	1.98 (18/18)	— (0/18)	0.16 (18/18)	2.82 (1/18)	45.84 (18/18)
PSD	interior	0.40 (18/18)	0.28 (18/18)	n/a	n/a	0.23 (1/18)	0.59 (18/18)
PSD	boundary	0.86 (18/18)	0.42 (18/18)	n/a	n/a	4.87 (1/18)	94.94 (18/18)
Occupancy	interior	0.30 (17/18)	0.25 (18/18)	— (0/18)	23.45 (16/18)	— (0/18)	1.77 (14/18)
Occupancy	boundary	6.74 (18/18)	2.72 (18/18)	3.64 (13/18)	1.93 (18/18)	— (0/18)	145.44 (18/18)
Low-rank	interior	0.30 (16/18)	0.24 (18/18)	3.12 (15/18)	5.62 (15/18)	— (0/18)	0.55 (16/18)
Low-rank	boundary	0.69 (18/18)	0.40 (18/18)	— (0/18)	— (0/18)	— (0/18)	42.92 (18/18)
Labels	interior	0.42 (17/18)	0.35 (18/18)	— (0/18)	— (0/18)	— (0/18)	1.67 (12/18)
Labels	boundary	1.47 (18/18)	0.48 (18/18)	— (0/18)	0.78 (18/18)	— (0/18)	151.26 (18/18)

Table 4: Additional primary cases at relative error 10^{-4}, in the same format as Table 3, with six instances per row.

Variant	ProjGD	ProjGD-BB	FW	Away FW	BroxGD fixed	BroxGD-BT
Fair ranking	0.85 (6/6)	1.05 (6/6)	87.40 (2/6)	202.20 (5/6)	— (0/6)	2.10 (5/6)
Huber low-rank	0.97 (6/6)	0.59 (6/6)	— (0/6)	— (0/6)	— (0/6)	84.42 (6/6)

Table 5: Held-out confirmation at relative error 10^{-4}. Ratios are baseline time divided by BroxGD-BT time, using per-seed medians of three timing repeats, then the median paired ratio. Values above one favor BroxGD-BT. Parentheses count seeds where both methods reached the target out of five; missing cases remain censored; n/a denotes an unavailable oracle.

Family / size	ProjGD	ProjGD-BB	FW	Away FW
Ranking/transport 12	0.16 (5/5)	0.42 (5/5)	5.61 (5/5)	9.55 (5/5)
Ranking/transport 24	0.17 (5/5)	0.67 (5/5)	1.88 (1/5)	43.90 (5/5)
Ranking/transport 12 fair	0.14 (5/5)	0.11 (5/5)	573.47 (5/5)	161.99 (5/5)
Ranking/transport 24 fair	0.16 (5/5)	0.66 (5/5)	289.98 (1/5)	263.26 (5/5)
PSD covariance 32	0.73 (5/5)	0.51 (5/5)	n/a	n/a
Occupancy measures 32	0.15 (5/5)	1.02 (5/5)	— (0/5)	13.55 (5/5)
Occupancy measures 64	0.16 (5/5)	3.30 (5/5)	— (0/5)	17.50 (5/5)
Low-rank matrices 32	0.61 (5/5)	0.46 (5/5)	40.93 (5/5)	127.02 (5/5)
Low-rank matrices 64	0.66 (5/5)	0.50 (5/5)	282.90 (5/5)	921.61 (5/5)
Structured labels 8 boundary	< 0.01 (5/5)	< 0.01 (5/5)	— (0/5)	< 0.01 (5/5)

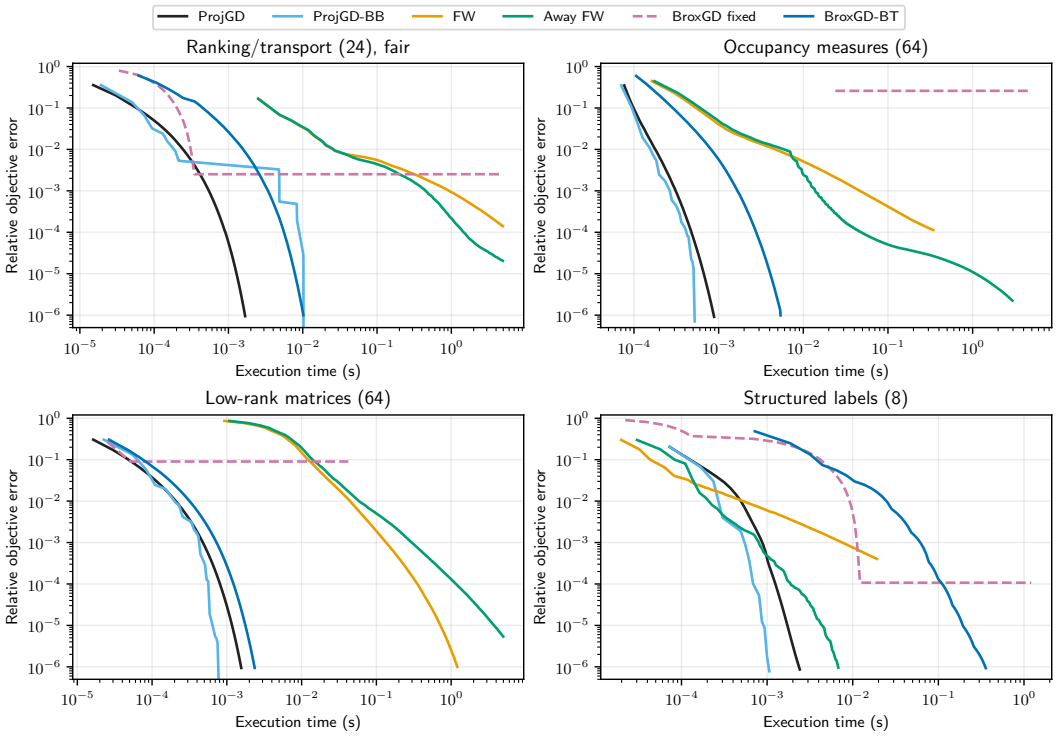


Figure 3: Held-out seed 10, selected before inspecting outcomes, in four confirmation configurations. Execution times are medians of three repeats at matching iteration indices. Curves terminate at the evaluation target or budget; BroxGD fixed reports best-so-far error. All held-out seeds enter Table 5, rather than only the plotted seed.

Where localization helps relative to FW. The clearest confirmed regime is an interior solution with inexpensive feasible local steps. On the nuclear-norm instances, BroxGD-BT reaches relative error 10^{-4} in all five held-out seeds at both sizes. Its median paired speedup over FW is $40.9\times$ at order 32 and $282.9\times$ at order 64; over away-step FW it is $127.0\times$ and $921.6\times$. The individual FW ratios range from 38.9 to 42.8 and from 278.3 to 288.0, respectively. These large ratios reflect cheap feasible gradient-like steps versus repeated spectral-oracle calls and atomic updates. They are not a scaling theorem: the sizes are modest, and stronger corrective or problem-specific matrix solvers are not tested.

Interior occupancy models also favor BroxGD-BT over the tested FW variants. Its median paired speedup over away-step FW is $13.6\times$ at 32 states and $17.5\times$ at 64 states, with all five seeds reaching the target for both methods. Vanilla FW does not reach 10^{-4} within the confirmation budgets. The global oracle here uses policy iteration, so the observed difference is not merely the overhead of a generic LP solver. On interior ranking models, BroxGD-BT beats away-step FW in all held-out seeds, with median ratios $9.6\times$ and $43.9\times$ without fairness constraints, and $162.0\times$ and $263.3\times$ with them. The fairness comparisons include the cost of solving global linear programs; specialized fairness-aware baselines could change these timings. At order 24, vanilla FW reaches the target for only one seed in each ranking variant, so its paired ratio is not representative of the whole stratum.

What the projected baselines reveal. These FW comparisons do not establish a general advantage over projection. In every confirmation stratum, the median paired runtime favors the better of ProjGD and ProjGD-BB over BroxGD-BT. In fact, on the interior nuclear and PSD tests, the first BroxGD-BT trial is feasible and equals $x_k - \nabla f(x_k)/L$, exactly the ProjGD update when projection is inactive. Their matched-index relative errors agree to below $6 \cdot 10^{-17}$. BroxGD-BT therefore inherits the favorable gradient-descent trajectory but adds overhead; ProjGD receives the same feasibility shortcuts. The interior occupancy and ranking tests likewise have cheap affine projections. ProjGD-BB occasionally takes costly boundary-crossing trials, allowing BroxGD-BT to beat that particular baseline on some instances, but fixed-step ProjGD remains faster there. A practical advantage must be assessed against both.

Active boundaries are the main obstacle in these implementations. In the primary grid, away-step FW is particularly effective on sparse ranking and occupancy optima. For explicit structured labels, ProjGD-BB is inexpensive and BroxGD-BT's repeated water-filling solves dominate runtime. At 10^{-4} in the held-out structured control, median work is 10 accepted ProjGD-BB updates and 12 projections, versus 48 accepted BroxGD-BT updates, 276 local-oracle calls, and 3782 inner projections. The PSD boundary similarly requires repeated spectral projections inside the local solve. The robust Huber matrix tests reproduce this limitation: all six BroxGD-BT runs reach 10^{-4}, but their median time is 84.4 ms, versus 0.59 ms for ProjGD-BB. Vanilla and away-step FW do not reach that target within the primary budget. Thus a favorable outer trajectory can coexist with unfavorable total cost.

Radius choice and accuracy matter. The fixed-horizon rule is much less reliable at high accuracy with the available diameter bounds. None of the 50 confirmation instances reaches 10^{-4} with this rule, although all do with BroxGD-BT. This is a finite-budget observation, consistent with a conservative radius and an accuracy floor at a fixed horizon, not a counterexample to its best-iterate theorem. At the coarser 10^{-2} target, BroxGD fixed exceeds the faster projected baseline by a factor above 1.25 on only two of the twenty held-out ranking instances; each ranking stratum's median still favors a projected method. The isolated exploratory speedups therefore do not support a reliable ProjGD advantage.

Numerical checks and interpretation. Across the 1950 reported runs there are no solver exceptions or failed final-feasibility checks. Independent checks verify the constructed first-order optimality conditions on all 242 distinct instances, and ten occupancy-oracle comparisons agree with an independent LP to $1.2 \cdot 10^{-15}$ in objective value. The largest recorded local/projection optimality certificate is $3.82 \cdot 10^{-10}$, the largest solver residual is $1.82 \cdot 10^{-10}$, and the largest local certificate divided by its initial objective gap is $4.18 \cdot 10^{-8}$. These are floating-point diagnostics, not formal error bounds for the entire trajectory. The study supports interior occupancy, ranking, and low-rank models as useful comparisons with global FW oracles; it identifies active-boundary local solvers as a bottleneck and supplies no consistent speed advantage over the strongest projected baseline tested.

Can a less aggressive radius search improve these results? We separately compare halving with contraction factors 0.8 and 0.9, and with two or four safeguarded bisection refinements. All variants reset the initial radius to $\|\nabla f(x_k)\|/L$ and use the same acceptance test (109); they change radius selection, not the gradient, local subproblem, or final full-step update. A bisection variant first halves until it obtains an accepted radius ℓ and retains the immediately preceding rejected radius h. It tests the midpoint $(\ell + h)/2$, replacing the accepted endpoint if the test passes and the rejected endpoint otherwise. After its fixed refinement budget, it takes the retained accepted point. If the initial trial passes, it performs no refinement. Every additional local solve is counted and timed, including accepted trial points later replaced by a larger accepted trial.

Why the refinement preserves the certificate. Under Assumption D.1 with exact local solves and exact acceptance tests, geometric contraction by any fixed $\xi \in (0, 1)$ still terminates: its trial radii tend to zero, so the small-radius argument of Theorem D.4 applies. The accepted-step certificate proof also extends to a retained accepted radius ℓ and rejected radius $h > \ell$. Write $\mathcal{D}(t)$ for the optimal linear decrease at radius t. If the accepted ball multiplier τ is positive, complementarity gives $\|s(\ell)\| = \ell$, and nesting of feasible balls yields

$$\tau\ell^2 \overset{(117)}{\leq} \mathcal{D}(\ell) \leq \mathcal{D}(h) \overset{(109)}{<} L\|s(h)\|^2 \overset{(108)}{\leq} Lh^2. \tag{185}$$

Thus the analysis multiplier $\lambda = \max\{L, \tau\}$ satisfies $\lambda \leq L(h/\ell)^2$; for $\tau = 0$ it equals L. The objective-decrease and endpoint-subgradient certificates in (114) are unchanged. Contraction by ξ gives $h/\ell = 1/\xi$, recovering the generalized multiplier bound L/ξ^2 in (114). After m midpoint refinements of a halving bracket, its width is divided by 2^m while its lower endpoint cannot decrease, so $h/\ell \leq 1 + 2^{-m}$. The bound is then $L(1 + 2^{-m})^2$. An accepted initial trial already has $\lambda = L$, by the original proof. This argument needs no monotonicity of the acceptance predicate and makes no claim that the retained point has the globally largest acceptable radius. It explains the potential for better accepted-step progress, not a reduction in total oracle work. The floating-point experiments retain the same 10^{-12} additive acceptance tolerance as the preceding study; the exact certificate argument is not applied literally to those solves.

Radius-search protocol. A pilot uses one seed in each of 20 configurations: the five families, interior and boundary optima, and $\kappa \in \{30, 300\}$, with the larger primary-grid size in each family. It compares the five radius variants and both projected baselines, for 140 runs capped at 300 updates or 1.5 seconds. Confirmation uses seeds 10–14 and three timing repeats in six configurations: order-24 ranking with and without fairness, 64-state occupancy, order-32 interior nuclear-norm problems, order-24 boundary PSD problems, and eight-block boundary label problems. All confirmation instances use $\kappa = 30$, with caps of 1000 updates or three seconds, producing 630 additional runs. As before, time caps are checked between updates, all methods start from the same feasible point, all methods receive feasibility shortcuts, method order is randomized, and timings include rejected trials and inner solves. These comparisons isolate radius search; the FW implementations and earlier measurements are unchanged. The archive is `output/experiments/radius-search/`, with drivers and analysis scripts in `scripts/study/`.

Table 6: Radius-search confirmation at relative error 10^{-4}. Entries are median paired speedups over halving (halving time / variant time), after taking the median of three timing repeats per seed. Above one favors the variant. Parentheses give the number of mutually successful seeds out of five. The first numerical column gives halving runtime in milliseconds, median across seeds.

Regime	Halving (ms)	Factor 0.8	Factor 0.9	Bisection 2	Bisection 4
Ranking, interior	5.63	1.52 (5/5)	1.75 (5/5)	0.90 (5/5)	0.67 (5/5)
Fair ranking, interior	5.73	1.56 (5/5)	1.79 (5/5)	0.89 (5/5)	0.65 (5/5)
Occupancy, interior	3.45	1.56 (5/5)	1.78 (5/5)	0.91 (5/5)	0.66 (5/5)
Low-rank, interior	0.68	1.02 (5/5)	1.02 (5/5)	1.01 (5/5)	1.01 (5/5)
PSD, boundary	116.41	0.51 (5/5)	0.27 (5/5)	0.97 (5/5)	0.83 (5/5)
Labels, boundary	234.53	0.50 (5/5)	0.26 (5/5)	0.98 (5/5)	0.82 (5/5)

Findings: fewer outer steps need not mean less work. At relative error 10^{-4}, factor 0.9 gives median paired speedups over halving of $1.75\times$ for ranking, $1.79\times$ for fair ranking, and $1.78\times$ for

occupancy measures. Factor 0.8 gives 1.52–$1.56\times$ on these three interior configurations. For ranking, the median accepted-step count falls from 126 to 70 with factor 0.9, and local-oracle calls fall from 252 to 140. Two bisection refinements also reduce the median step count, to 72, but require 288 local solves; four refinements need 67 steps and 402 solves. Their median speedups are therefore only 0.90 and 0.67, respectively. The analogous occupancy counts are 103 steps / 206 calls for halving, 57 / 120 for factor 0.9, and 59 / 236 for two refinements. Gentler contraction helps here because the first rejected radius is already close to an acceptable one.

The result reverses at active boundaries. On PSD problems, factor 0.9 reduces the median accepted steps from 25 to 18 but increases local solves from 134 to 502, making it about $3.7\times$ slower. On explicit labels, the corresponding counts are 48 / 276 and 35 / 1038, with about a $3.8\times$ slowdown. Two bisection refinements are approximately tied with halving on both boundary configurations (median speedups 0.97 and 0.98); four are slower (0.83 and 0.82). The encouraging PSD bisection speedup in the small pilot does not persist in the confirmation median. On the interior nuclear-norm control, every initial trial is accepted, so all five rules give identical iterates and call counts; differences of a few percent in timing are noise rather than a radius-selection improvement.

All 30 confirmation instances reach 10^{-4} with every method in all three repeats. The 770 pilot and confirmation runs have no solver exceptions or failed final-feasibility checks, and no recorded objective increases. The largest accepted-test excess is below 10^{-12}, consistent with the common numerical tolerance. These results support testing a gentler contraction on interior affine problems, but do not support replacing halving universally with either 0.9 or fixed-budget bisection. None of these refinements overturns the projected-baseline conclusion: fixed-step ProjGD remains faster in each confirmation configuration. Reusing radius information across outer iterations, instead of resetting to $\|\nabla f(x_k)\|/L$, is a separate possible improvement not tested here.

Reproducibility and limits. The complete design, pilot history, raw traces, statuses, and analysis scripts are provided under `scripts/study/` and `output/experiments/application-study/`. Tables and figures are generated from the saved traces. All tested instances, including unsuccessful and unfavorable ones, remain in the archive. The examples establish implementation-dependent regimes within these synthetic models; they neither prove a general separation between oracle models nor establish an advantage on real application datasets. In particular, no positive claim about BroxGD versus an adaptive projected method follows merely from outperforming vanilla FW.

G.2.1 A SELECTED FAIR-RANKING EXAMPLE

We illustrate a possible low-accuracy benefit on a synthetic fair-ranking problem: a 24×24 doubly stochastic plan with one exposure equality and a weighted quadratic loss of curvature ratio 30. This instance (seed 12) is selected from the broader study in Appendix G.2; it is not representative of every seed. BroxGD uses the fixed-horizon radius $t = \sqrt{48/2000}$ from Theorem 3.2, reporting its best iterate. We compare it with fixed-step ProjGD, adaptive ProjGD-BB, FW, and away-step FW. All methods use the same feasible initialization, and timings include their inner solves.

With the study's generic projection solver, at relative error 10^{-2} BroxGD takes 0.37 ms, versus 2.92 ms for ProjGD, 2.90 ms for ProjGD-BB, 6.36 ms for FW, and 6.73 ms for away-step FW. Thus it is about $8\times$ faster than the projected methods and 17–$18\times$ faster than the FW variants in this case. Local steps avoid a boundary-constrained projection solve at this target. This advantage is accuracy- and instance-dependent: the fixed-radius error plateaus near $1.16 \cdot 10^{-3}$, and the win does not persist across the tested seeds. We therefore present it as an illustration of oracle cost, not evidence of a uniformly superior method. Stronger projection implementations and new correlation-matrix tests are examined in Appendix G.3; full positive and negative comparisons remain in the appendix.

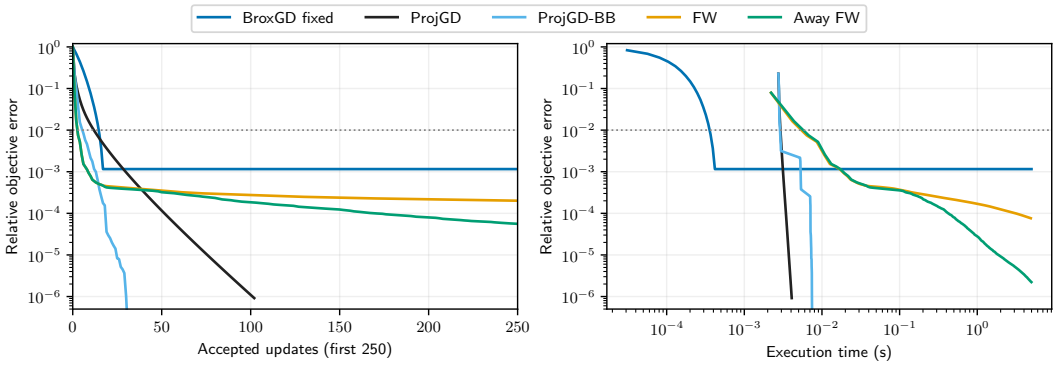


Figure 4: One selected fair-ranking instance: relative objective error against accepted updates and execution time, with a dotted 10^{-2} target. Times are medians of three repetitions. The local fixed-radius method reaches this target quickly but later plateaus; the broader seed comparisons and other experiments remain in Appendix G.

G.3 Searching for a substantial projection-cost advantage

This subsection gives the full protocol and controls for the correlation-matrix experiments in Section 4: can the local linear subproblem be much cheaper than a well-implemented projection, across new instances? We first strengthen the ranking projections, then test correlation-matrix constraints. All these experiments are synthetic; they measure optimization cost, not statistical estimation performance.

A stronger ranking control. On doubly stochastic matrices, the affine projection can be computed by row and column centering, without a dense pseudoinverse. A single block-exposure equality adds a rank-one correction. We implement these formulas for every method and compare a warm-started Dykstra projection onto the affine constraints and the nonnegative orthant with the previous generic quadratic-programming solver. The pilot uses orders 32 and 64, seeds 0 and 1, a curvature ratio of 3, and an 80% mixture of three permutation matrices with 20% uniform mass. At order 64, predetermined geometric BroxGD reaches relative error 10^{-2} in about 0.2 ms, compared with about 23 ms for generic-solver ProjGD, but only about 0.4 ms for the specialized ProjGD projection. Thus a large advantage over the generic solver does not establish a large advantage over projection itself. The selected fair-ranking measurements in Appendix G.2.1 retain their original implementation; this control limits their interpretation.

Correlation-matrix model and exact local shortcut. Consider the weighted least-squares problem

$$\min_{X\in\mathcal{E}_n} f(X)=\frac{1}{2}\sum_{i,j} w_{ij}(X_{ij}-(X_\star)_{ij})^2, \qquad \mathcal{E}_n=\{X=X^\top\succeq 0:\operatorname{diag}(X)=\mathbf{1}\}. \tag{186}$$

Here $X_\star = 0.6BB^\top + 0.4I$, where independent Gaussian rows of $B\in\mathbb{R}^{n\times r}$ are normalized to unit length. This supplies a known interior optimum. The weights are a shuffled geometric sequence from $1/3$ to 1, symmetrized across the diagonal. We initialize at I. For $G=\nabla f(X)$ and $H=G-\operatorname{Diag}(\operatorname{diag}(G))$, the affine-ball minimizer is

$$X^+ = X - t\frac{H}{\|H\|_F} \qquad (H\neq 0). \tag{187}$$

Whenever this matrix is positive semidefinite, it is also the exact BroxGD solution by Proposition C.1; a successful Cholesky factorization checks positive definiteness numerically. If $H=0$, we keep X unchanged. This computation does not project onto $\mathcal{E}_n$. In contrast, when the affine ProjGD candidate is indefinite, our Euclidean projector alternates the unit-diagonal and PSD projections with Dykstra corrections, following the correlation-projection approach of Higham (2002). Each PSD projection uses an eigendecomposition. All methods share the same inexpensive feasibility shortcut whenever their own candidate permits it.

Protocol and controls. Pilots investigate orders 64–512, two covariance-rank scalings, several fixed radii, and the predetermined geometric rule. Confirmation fixes $(n, r) \in \{(128, 5), (256, 10), (512, 20)\}$ and uses ten new seeds, 10–19, with three timing repetitions per method. BroxGD uses the fixed-horizon rule in Theorem 3.2, with the certified bound $R = \sqrt{n(n-1)}$ on $\|I - X_\star\|_F$ and horizons $K_{\mathrm{rad}} = 2000n/128$ or $10000n/128$. The former is the primary rule; both are retained. We measure best-iterate relative objective error. These horizon parameters select the radii; the recorded runs have a separate cap of 200 updates or three seconds, checked between updates. If the local affine-ball candidate becomes indefinite, the run stops with a recorded boundary event: we do not count a target beyond that event as attained, or silently substitute an approximate local oracle.

The projected baselines are ProjGD with stepsize $1/L$ and safeguarded spectral ProjGD-BB, with $L = \max_{ij} w_{ij}$. Projections are warm-started and stopped by a primal–dual gap at most $\tau \max\{1, \|V - I\|_F^2\}$ for input V. We test $\tau = 10^{-10}$ and the less accurate 10^{-6} and 10^{-4} alternatives. The latter distinguish a separation for accurately solved oracles from a practical comparison allowing approximate projections. A further control, *feasibility-backtracked gradient descent*, starts from $X - H/L$ and halves its step until a Cholesky check succeeds. It avoids global projections but can stall near the boundary; no general convergence claim is attached to this control. Timings include feasibility checks, inner iterations, projection certificates, gradient evaluations, and best-iterate selection, with one computational thread and randomized method order. Common instance generation and reference-error logging are excluded.

Confirmed low-accuracy advantage. Table 7 reports the final comparison, with the projection gap checked after every Dykstra cycle. Across all 30 size–seed pairs, the primary BroxGD radius reaches relative error 10^{-2} faster than both projected baselines, including their loose-projection variants. The paired median speedup against the faster of ProjGD and ProjGD-BB is $6.0\times$, $6.2\times$, and $7.0\times$ at orders 128, 256, and 512 with accurate projections; it is $1.9\times$, $2.0\times$, and $2.2\times$ with loose projections. Even the smallest paired loose-projection speedup is $1.72\times$. These figures use 450 confirmation runs and 540 strengthened projection runs; the final comparison retains 270 local/control runs and the 540 strengthened projection runs. An earlier 360-run tolerance check and all superseded projected timings remain archived. Checking projection certificates every cycle, instead of every fifth cycle, materially improves ProjGD and is included in these final numbers.

Table 7: Correlation-matrix confirmation: median time (ms) to relative error 10^{-2}, first taking the median of three repeats per seed, then across ten seeds. BroxGD uses $K_{\mathrm{rad}} = 2000n/128$; BroxGD (small) uses $10000n/128$. Accurate projections use $\tau = 10^{-10}$; loose projections use 10^{-4}. Every entry succeeds on all ten seeds except the marked feasibility control (nine).

n	BroxGD	BroxGD (small)	ProjGD	ProjGD-BB	ProjGD loose	ProjGD-BB loose	Feas. GD
128	1.30	2.82	7.75	7.77	2.48	2.48	0.62
256	5.82	12.96	35.97	35.55	11.57	11.27	3.57*
512	26.93	61.22	191.50	189.72	60.65	59.30	18.08

Where the time is saved. At the shared initial matrix I, the median accurately solved projection costs approximately 54, 49, and 54 times as much as the first local linear solve at the three sizes. Even the loose projection costs approximately 15, 13, and 14 times as much. These are matched initial-point comparisons, not averages over different later trajectories. The complete local method has a smaller speedup because it takes more outer updates. The smaller-radius variant also reaches the 1% target on all 30 instances, but its extra iterations erase much of the advantage over loose projections.

Figure 1 in the main text illustrates the error curves and these matched oracle costs.

The stronger projection-avoiding control. Feasibility-backtracked gradient descent is faster than BroxGD on all 29 instances where it reaches the 1% target. It stalls on the remaining order-256 instance within the recorded budget, whereas BroxGD reaches the target there. Thus the experiment establishes a consistent advantage over the tested *projected* implementations, and a substantial cost difference between their projections and these local solves; it does not establish that BroxGD is the

fastest method available. The advantage is also limited in accuracy: the primary fixed radius reaches 10^{-3} on 29 of 30 instances, but reaches 10^{-4} on none before a boundary event or the budget. Both projected baselines reach 10^{-6} on every instance at every tested projection tolerance.

FW solver check. A separate check solves the correlation-matrix global linear oracle as an SDP, using Clarabel on its diagonal dual and an exact quadratic outer line search. At order 128, seeds 10–12, even the first FW update takes 34.5–41.1 seconds and leaves relative error between 0.298 and 0.374. BroxGD reaches the 1% target in about 1.3 ms on these instances. These are one-update lower bounds on this SDP-based FW implementation's time to target, not completed FW convergence runs or a comparison with every approximate or specialized SDP oracle. Away-step FW has the same first update from the single initial atom. The expensive SDP check is kept separate from the ten-seed projected-method confirmation.

Limits and reproducibility. The regime depends on covariance rank and requested accuracy. Keeping rank five while increasing the order to 256 or 512 makes the cheap local path hit the PSD boundary before the 1% target. Increasing the rank to $n/12.8$ in the pilot instead lets ProjGD avoid the expensive projection and removes the local advantage. The successful scaling $r = n/25.6$ was chosen from pilots, then checked on fresh seeds. Fixed radii can later plateau or require a boundary local solve; these measurements do not establish a high-accuracy or uniform advantage. Nor do they prove an intrinsic oracle-complexity separation or compare every specialized correlation-matrix solver. All pilots, failed local continuations, tolerance controls, and confirmation traces are retained in `output/experiments/projection-gap/`; the drivers and independent oracle checks are in `scripts/study/`. New numerical checks compare small ranking projections against an independent quadratic solver and correlation projections against a variational certificate obtained from an SDP linear oracle.

G.4 PROBLEM SETUP

We consider the quadratic objective

$$f(x) = \frac{1}{2} x^\top \boldsymbol{Q} x, \qquad x \in \mathbb{R}^2,$$

where $\boldsymbol{Q} \in \mathbb{R}^{2\times 2}$ is symmetric positive definite with eigenvalues $\mu = 1$ and $L = 100$. Thus, f is μ–strongly convex and L–smooth. We choose $\boldsymbol{Q}$ as a rotated diagonal matrix,

$$\boldsymbol{Q} = \boldsymbol{R} \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix} \boldsymbol{R}^\top, \quad \boldsymbol{R} = \begin{pmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{pmatrix}$$

so that explicitly

$$\boldsymbol{Q} = \begin{pmatrix} 25.75 & -42.86825749 \\ -42.86825749 & 75.25 \end{pmatrix}.$$

By construction, the unconstrained minimizer of f is the origin $x_\star^{\mathrm{unc}} = (0, 0)$. The feasible set is the box $\mathcal{X} \coloneqq \{x \in \mathbb{R}^2 : \|x - c\|_\infty \le 1\}$, $c = (3, 3)$, that is, $\mathcal{X} = [2, 4] \times [2, 4]$. The starting point is $x_0 = (4, 4) \in \mathcal{X}$. In this experiment, the constrained minimizer is $x_\star \approx (3.3295734, 2)$, which lies on the lower edge of the box. At iteration k, BroxGD computes

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle,$$

where the radius is chosen according to the theoretically prescribed rule

$$t_k \overset{(20)}{=} \theta \|x_k - x_\star\|, \qquad \theta = \frac{2\sqrt{\mu L}}{L + \mu} = \frac{2\sqrt{100}}{101} \approx 0.198020.$$

Theorem 3.4 predicts

$$\|x_{k+1} - x_\star\|^2 \overset{(21)}{\le} \left(1 - \theta^2\right) \|x_k - x_\star\|^2 \approx 0.960788 \cdot \|x_k - x_\star\|^2.$$

For the convergence plots we run Algorithm 1 for 200 iterations; the method and radius comparisons below use 100 iterations. Since the problem is two-dimensional, the subproblem can be solved exactly by enumerating the relevant boundary candidates: corners of the box lying inside the ball, intersections of the circle and the box edges, and the unconstrained minimizer of the linear function over the ball whenever it is feasible.

G.5 Geometry of the iterates

Figure 5 shows the contour lines of the quadratic objective, the box constraint, the unconstrained minimizer, the constrained minimizer, the first few balls $\mathcal{B}(x_k, t_k)$, and the trajectory of the iterates.

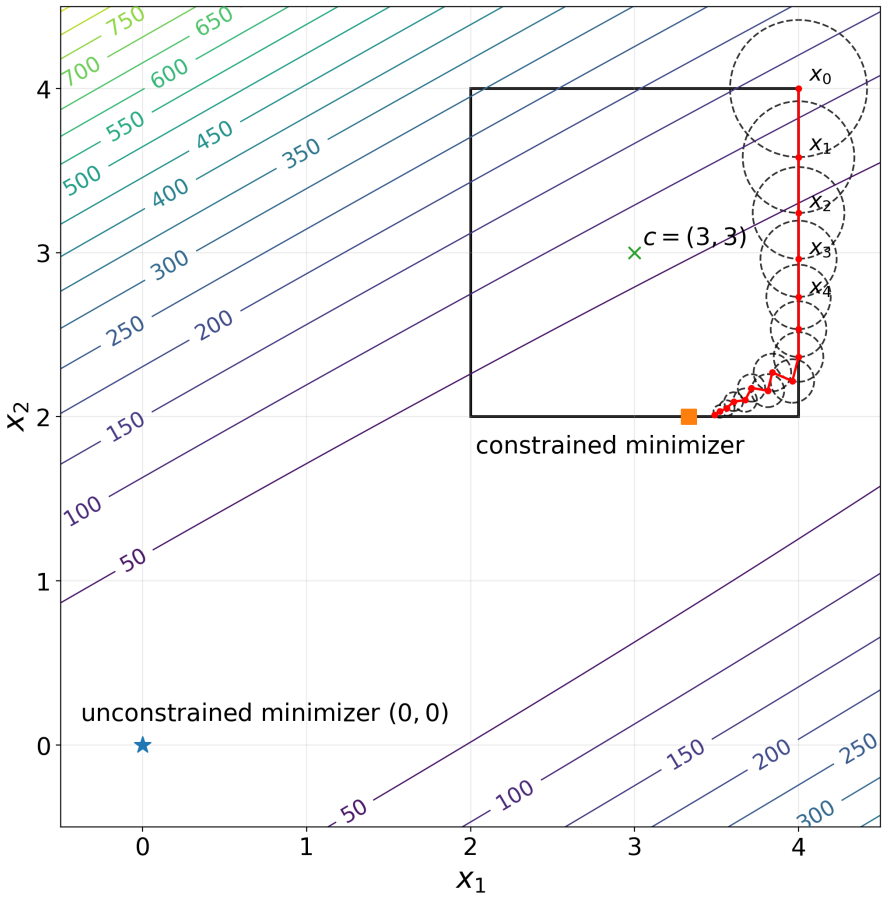


Figure 5: Two-dimensional illustration of BroxGD with the radius rule $t_k = \theta\|x_k - x_\star\|$. The plot shows the contour lines of the quadratic objective, the feasible box $\mathcal{X} = [2, 4] \times [2, 4]$ with center $c = (3, 3)$, the unconstrained minimizer $(0, 0)$, the constrained minimizer $x_\star$, the first few trust-region balls, and 15 iterates $\{x_k\}$. In this case, $L = 100$, $\mu = 1$ and $\theta \approx 0.19802$.

Several features are worth noting. First, the method starts at the upper-right corner of the box and initially moves down along the right boundary. This is natural, because the gradient points roughly toward the lower-right direction, but the iterates must remain inside the feasible set and inside the local ball. Second, once the iterates get close to the lower boundary, they quickly move toward the constrained minimizer $x_\star$. Third, the shrinking radii t_k make the local feasible regions progressively smaller as the iterates approach the solution.

G.6 Gradient-difference convergence

We next examine the quantity $\|\nabla f(x_k) - \nabla f(x_\star)\|^2$. For its gradient-difference radius rule, Theorem D.11 gives the upper bound

$$\min_{0 \le k \le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \overset{(147)}{\le} \frac{L^2 \|x_0 - x_\star\|^2}{K}.$$

The present run instead uses the strongly convex distance-based radius from Theorem 3.4. Thus, the $L^2\|x_0 - x_\star\|^2/(k+1)$ curve is a reference benchmark from a different radius rule, not a bound

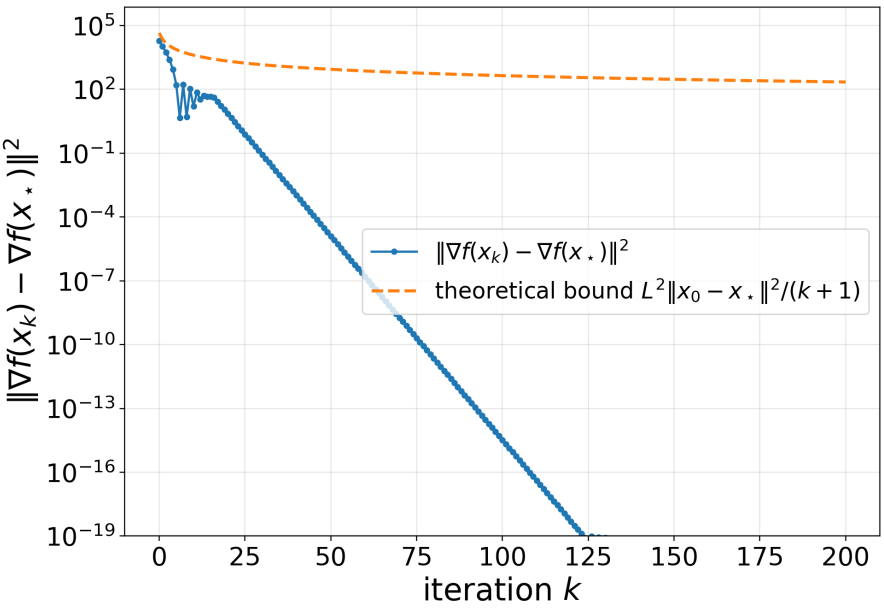


Figure 6: Semi-log plot of the gradient-difference quantity $\|\nabla f(x_k) - \nabla f(x_\star)\|^2$, together with its running minimum and the reference curve $L^2\|x_0 - x_\star\|^2/(k+1)$.

supplied by Theorem D.11 for this trajectory. Figure 6 shows the raw gradient-difference sequence, its running minimum, and this reference curve.

The behavior is strikingly better than what the $\mathcal{O}(1/k)$ bound alone would suggest. After a short transient phase, the quantity $\|\nabla f(x_k) - \nabla f(x_\star)\|^2$ decreases essentially linearly on the logarithmic scale until it reaches machine precision. In particular, the running minimum decays much faster than the generic $1/k$ benchmark. This is consistent with the fact that the present problem is strongly convex and that the radius rule was designed using the strong convexity and smoothness constants.

G.7 Distance convergence and linear theory

Finally, Figure 7 shows the squared distance to the constrained minimizer, together with the linear upper bound predicted by Theorem 3.4:

$$\psi(k) := \left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x_0 - x_\star\|^2.$$

The empirical behavior supports the theorem. The observed sequence decays much faster than the theoretical geometric upper bound and reaches numerical precision after roughly 125 iterations. This experiment also highlights an important implementation detail. The radius rule in Theorem 3.4 is $t_k = \theta\|x_k - x_\star\|$, where $x_\star$ is the *constrained* minimizer. If one mistakenly uses the unconstrained minimizer instead, then the predicted linear behavior may fail completely.

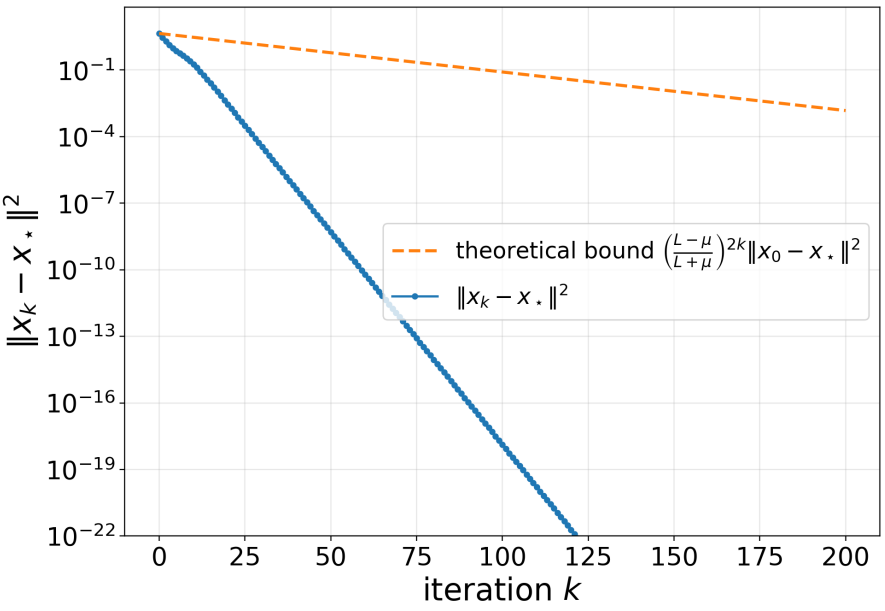


Figure 7: Semi-log plot of the squared distance to the constrained minimizer for BroxGD, together with the geometric upper bound predicted by Theorem 3.4.

G.8 Comparison with Projected Gradient Descent and Frank–Wolfe

We now compare BroxGD with two standard first-order methods for constrained smooth convex optimization: Projected Gradient Descent and Frank–Wolfe. We use exactly the same problem instance as in Appendix G.4. The three methods are implemented as follows.

BroxGD. We use the same method and the same theoretically prescribed radius rule as before:

$$x_{k+1} \in \operatorname*{argmin}_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle, \qquad t_k = \theta \|x_k - x_\star\|.$$

Projected Gradient Descent. We run Projected Gradient Descent with the standard stepsize $1/L$:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}\left(x_k - \frac{1}{L}\nabla f(x_k)\right). \tag{ProjGD}$$

Frank–Wolfe. We run the classical Frank–Wolfe method

$$s_k \in \operatorname*{argmin}_{s \in \mathcal{X}} \langle \nabla f(x_k), s \rangle, \qquad x_{k+1} = (1-\gamma_k)x_k + \gamma_k s_k, \qquad \gamma_k = \frac{2}{k+2}. \tag{FW}$$

Since $\mathcal{X} = [2,4]^2$ is a box, the linear minimization oracle is explicit. We run all three methods for 100 iterations.

Figure 8 shows the semi-log plot of the squared distance to the constrained minimizer $\|x_k - x_\star\|^2$. The figure reveals a clear separation between the three methods. First, ProjGD is the fastest method on this example. Its convergence is essentially linear on the logarithmic scale and reaches extremely high accuracy already within the 100-iteration horizon. This is not surprising: ProjGD has direct access to the Euclidean projection onto the feasible set, and the present problem is a smooth strongly convex quadratic.

Second, BroxGD also exhibits clear linear convergence, in full agreement with the theory. Moreover, the empirical decay is substantially faster than the conservative geometric upper bound. Thus, at least on this simple problem, the theorem provides a valid but somewhat pessimistic worst-case estimate. The observed behavior is significantly better.

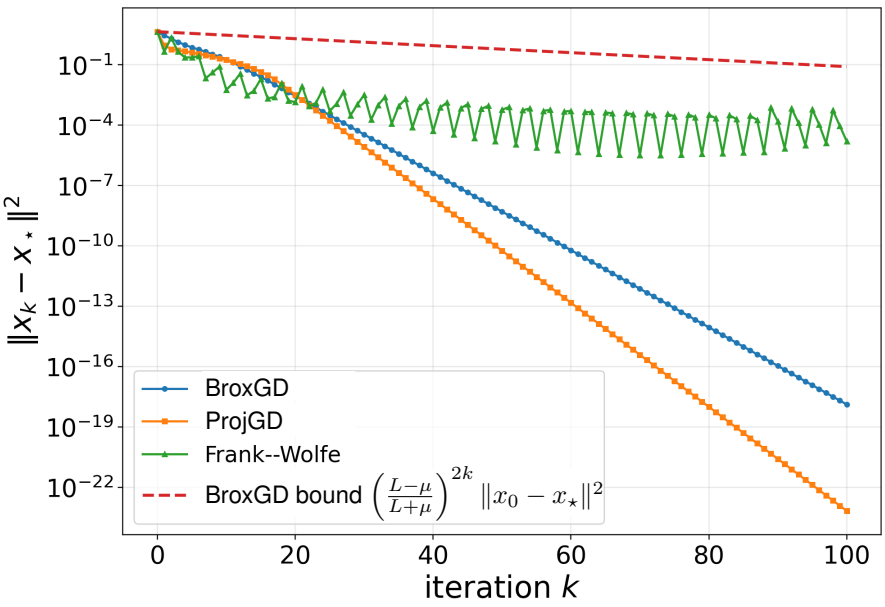


Figure 8: Semi-log plot of the squared distance to the constrained minimizer for BroxGD, Projected Gradient Descent, and Frank–Wolfe, all run on the same two-dimensional strongly convex quadratic problem for 100 iterations. For BroxGD, we also plot the geometric upper bound $\left((L-\mu)/(L+\mu)\right)^{2k} \|x_0 - x_\star\|^2$.

Third, Frank–Wolfe is much slower than the other two methods. Its trajectory displays the familiar zig-zagging behavior associated with projection-free methods on constrained problems. At the end of the 100-iteration horizon, the observed squared distances are summarized in Table 8.

Table 8: Observed squared distances after 100 iterations.

	BroxGD	ProjGD	FW
$\|x_{100} - x_\star\|^2$	1.32×10^{-18}	6.71×10^{-24}	1.58×10^{-5}

On this instance, BroxGD converges faster than FW and slower than ProjGD. The comparison counts iterations, not runtime or oracle cost: projection onto this box is cheap, and the local oracle is solved by a specialized two-dimensional enumeration. These observations do not establish the same ordering for other constraint sets.

G.9 Comparison with geometric radius schedules

We now compare the theoretically prescribed adaptive radius rule $t_k = \theta\|x_k - x_\star\|$, with a family of geometric radius schedules of the form

$$t_k = \theta\|x_0 - x_\star\| r^k, \qquad r \in (0, 1).$$

Here r is a radius-decay factor; the certified schedule in Theorem D.3 uses $r = \sqrt{q}$ for the objective contraction factor $q = 1 - \mu/(4L)$. The linear distance contraction in (21) motivates a geometric schedule $t_k = cr^k$. Here we set $c = \theta\|x_0 - x_\star\|$. The motivation for this comparison is that the adaptive rule depends on the unknown current distance to the solution, whereas the geometric rule uses only the initial distance and a fixed decay parameter r. Hence, it is natural to ask whether a well-tuned geometric schedule can mimic the behavior of the adaptive rule in practice.

We use exactly the same problem instance as before. The constant appearing in the radius rule is $\theta \approx 0.198020$. For the geometric rule, we test 10 equally spaced values of r in the interval $[0.8, 0.95]$. All methods are run for 100 iterations, and we compare the quantity $\|x_k - x_\star\|$.

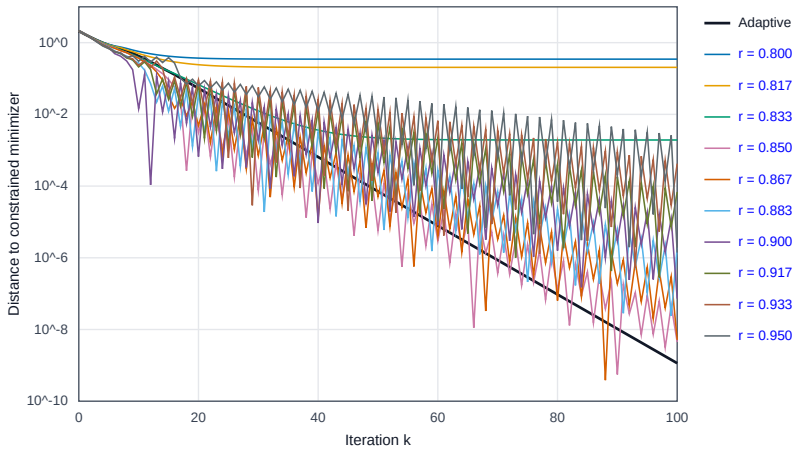


Figure 9: Semi-log plot comparing the adaptive radius rule $t_k = \theta\|x_k - x_\star\|$ with geometric radius schedules $t_k = \theta\|x_0 - x_\star\|r^k$ for 10 values of $r \in [0.8, 0.95]$.

Figure 9 reveals a clear pattern. The adaptive rule produces a stable and nearly straight line on the semi-log scale, indicating clean linear convergence. Several geometric schedules also perform well, but their behavior is much more sensitive to the choice of r.

If r is too small, the radius shrinks too quickly. In that case the method becomes overly conservative, and progress stalls long before the iterate gets close to the solution. This phenomenon is clearly visible for $r = 0.800$ and $r = 0.817$, where the method stagnates at a relatively large distance from $x_\star$. If r is too large, the radius shrinks too slowly. Then the local subproblems remain too large for too long, which leads to oscillatory behavior and poorer asymptotic performance. This can be seen for $r = 0.917$, $r = 0.933$, and $r = 0.950$, where the convergence is still present but noticeably less efficient.

The best-performing geometric schedules are clustered around $r \approx 0.85$ to 0.90. In particular, the choices $r = 0.850$ and $r = 0.867$ almost coincide with the adaptive baseline over the entire horizon, and $r = 0.900$ also performs very well. At the end of the 100-iteration run, the observed distances are summarized in Table 9.

Table 9: Observed distances after 100 iterations for different choices of r.

Setting	$\|x_{100} - x_\star\|$	Setting	$\|x_{100} - x_\star\|$	Setting	$\|x_{100} - x_\star\|$
Adaptive baseline	1.15×10^{-9}	$r = 0.800$	3.47×10^{-1}	$r = 0.817$	2.05×10^{-1}
$r = 0.833$	1.93×10^{-3}	$r = 0.850$	4.45×10^{-9}	$r = 0.867$	5.15×10^{-9}
$r = 0.883$	1.91×10^{-6}	$r = 0.900$	7.27×10^{-8}	$r = 0.917$	6.86×10^{-5}
$r = 0.933$	4.22×10^{-4}	$r = 0.950$	1.39×10^{-6}		

A well-tuned geometric schedule nearly matches the adaptive rule on this example, while the adaptive rule avoids selecting a decay parameter. Both baselines use the known reference solution: the geometric schedule still requires the initial distance, and the adaptive schedule requires the current distance at every iteration. Thus, this is a controlled heuristic comparison, not a test of a solution-independent implementation. The certified practical schedules in Appendix D.1 provide separate theoretical guarantees.

G.10 ALTERNATIVE CONSTRAINT GEOMETRIES

G.10.1 SHIFTED BOXES

The two shifted boxes in Figure 10 illustrate how the local balls interact with the boundary at different locations relative to the unconstrained minimizer. Each run uses the distance-based radius with respect to its own constrained solution.

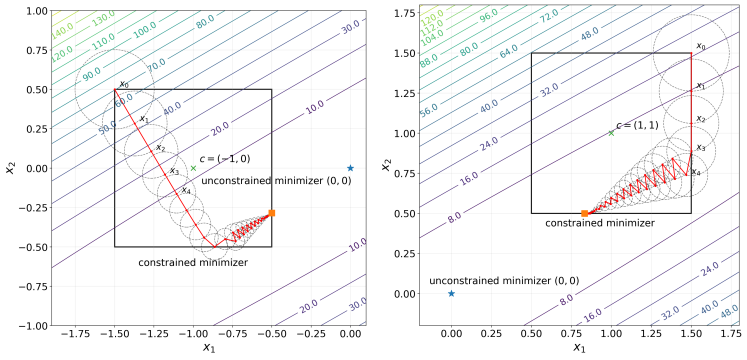


Figure 10: Illustration of BroxGD dynamics for an L–smooth ($L = 100$) and μ–strongly convex ($\mu = 1$) quadratic $f : \mathbb{R}^2 \to \mathbb{R}$, and a box constraint of radius 0.5 centered at $c = (-1, 0)$ (left) and $c = (1, 1)$ (right), with the radius rule $t_k = \theta\|x_k - x_\star\|$, where $\theta = 2\sqrt{\mu L}/(L + \mu) \approx 0.19802$ (see Theorem 3.4). Shown: 100 iterates $\{x_k\}$ of BroxGD and the corresponding balls $\mathcal{B}(x_k, t_k)$. Note that $\|x_{k+1} - x_k\| = t_k$ for all k (Theorem 3.1(ii) says that this is always the case).

G.10.2 THREE BALL GEOMETRIES

Here we perform an experiment with the same two-dimensional function as before, but with three different constraints: ℓ_1, ℓ_2 and ℓ_∞ balls of radius one centered at various points. Figure 11 shows 100 iterates of BroxGD in each case. The point of this experiment is to showcase the different ways the iterates of our methods interact with the boundary of constraint sets of different geometries.

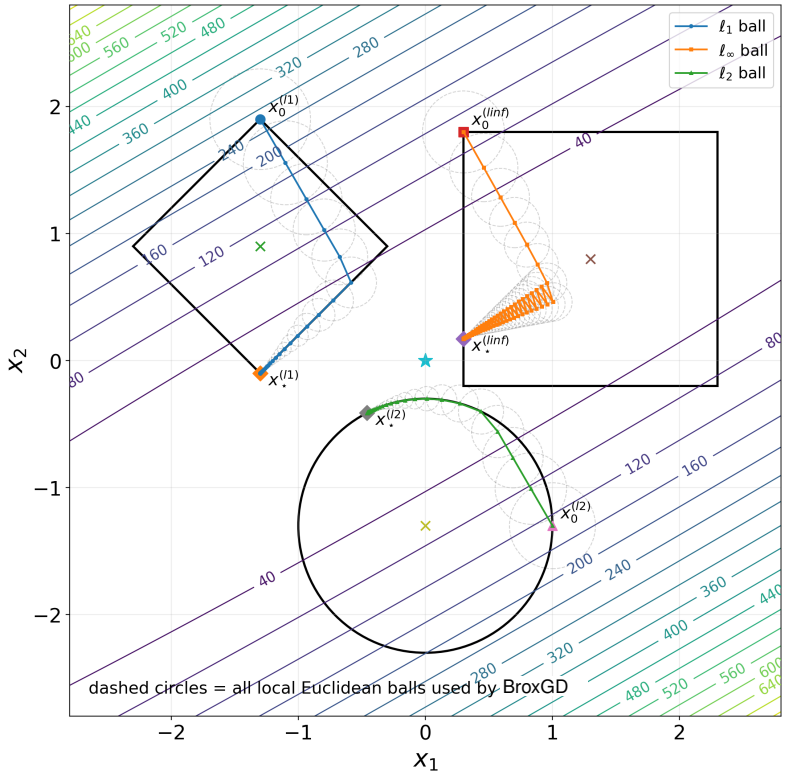


Figure 11: Illustration of BroxGD dynamics for an L–smooth ($L = 100$) and μ–strongly convex ($\mu = 1$) quadratic $f : \mathbb{R}^2 \to \mathbb{R}$, and three ball constraints of radius 1 (with ℓ_1, ℓ_2 and ℓ_∞ ball geometries), with the BroxGD radius rule $t_k = \theta\|x_k - x_\star\|$, where $\theta \approx 0.19802$ (see Theorem 3.4). Shown: 100 iterates $\{x_k\}$ of BroxGD and the corresponding balls $\mathcal{B}(x_k, t_k)$. Note that $\|x_{k+1} - x_k\| = t_k$ for all k (Theorem 3.1(ii) says that this is always the case).

G.11 DISCUSSION AND SCOPE

These experiments illustrate the predicted geometry and show that tuned geometric radii can work well on a toy strongly convex quadratic. They also expose sensitivity to the decay parameter and the distinction between iteration counts and computational cost. Broader theoretical and empirical work, including realistic and large-scale applications, remains necessary. No such application experiments are reported here.

H Potential applications and local-oracle implementations

This appendix develops five possible applications: ranking and transport, PSD-constrained learning, occupancy-measure optimization, robust low-rank learning, and structured prediction. These are mathematical models and implementation directions. Correlation-matrix experiments are summarized in Section 4, a selected fair-ranking instance in Appendix G.2.1, an explicit finite-label instance in Appendix G.1, and Appendix G.2 tests controlled synthetic instances from all five families. These experiments distinguish favorable optimization regimes from costly boundary solves; downstream application performance remains untested. We distinguish cases with a closed-form exact oracle from cases requiring an iterative inner solver. The guarantees of Appendix D.1 apply only when their objective assumptions and exact-oracle requirements hold. A small outer-iteration count alone does not bound wall-clock time.

H.1 A common template: locally inactive inequalities

Consider a feasible set with affine equalities and additional convex constraints,

$$\mathcal{X} = \{x \in \mathbb{R}^d : Ax = b,\ x \in \mathcal{K}\}, \qquad P_A = I - A^\top (AA^\top)^\dagger A, \tag{188}$$

where $\mathcal{K}$ is closed and convex and $\dagger$ denotes the Moore–Penrose pseudoinverse. The operator P_A is the orthogonal projector onto $\ker A$. A useful sufficient condition at $x_k \in \mathcal{X}$ is

$$\{x_k + D : AD = 0,\ \|D\| \le t_k\} \subseteq \mathcal{K}. \tag{189}$$

This is a local containment condition, not a claim that inequalities are inactive globally.

Proposition H.1 (Exact oracle on a locally affine feasible set)**.** If (189) holds for $t_k > 0$ and $h_k = P_A \nabla f(x_k) \neq 0$, the exact BroxGD solution is

$$x_{k+1} = x_k - t_k \frac{h_k}{\|h_k\|}. \tag{190}$$

If $h_k = 0$, every local feasible point has the same linear objective and one may choose $x_{k+1} = x_k$. For convex differentiable f, this zero projected gradient certifies constrained optimality.

Proof. By (188) and (189), the local feasible displacements are exactly $\{D : AD = 0, \|D\| \le t_k\}$. For any such displacement,

$$\langle \nabla f(x_k), D\rangle \overset{(188)}{=} \langle h_k, D\rangle \ge -\|h_k\|\|D\| \ge -t_k\|h_k\|.$$

The displacement in (190) attains this bound. When $h_k = 0$, the linear objective is constant on the affine space. Every $y \in \mathcal{X}$ has $y - x_k \in \ker A$, so convexity gives

$$f(y) \ge f(x_k) + \langle \nabla f(x_k), y - x_k\rangle = f(x_k).$$

□

Applying P_A may require a linear solve, but a factorization for fixed A can be reused. Checking (189) has its own cost. If the sufficient radius cap is zero, this formula is unavailable; the exact local oracle remains well defined for positive radii and must retain the active inequalities. Moreover, clipping a radius prescribed by Appendix D.1 to a local slack bound changes the algorithm and does not automatically preserve its stated rate.

H.2 Fair ranking, differentiable matching, and nonlinear transport

A square stochastic ranking or matching plan can be modeled by

$$\mathcal{X} = \{P \in \mathbb{R}^{n\times n} : P\mathbf{1} = \mathbf{1},\ P^\top \mathbf{1} = \mathbf{1},\ P \ge 0,\ C\,\mathrm{vec}(P) = c\}. \tag{191}$$

The last equality can encode linear exposure-fairness requirements. Nonlinear transport uses the same construction with prescribed marginal vectors in place of $\mathbf{1}$. A representative objective is

$$f(P) = \ell(P) + \frac{\lambda}{2}\|P - P_{\mathrm{ref}}\|_F^2, \qquad \lambda \ge 0, \tag{192}$$

with a convex differentiable loss ℓ; when ℓ is smooth and $\lambda > 0$, the objective is smooth and strongly convex.

For ordinary doubly stochastic constraints without the additional matrix C, set

$$J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad G_T = (I - J)\nabla f(P_k)(I - J). \tag{193}$$

This is the Frobenius-orthogonal projection onto the space of matrices with zero row and column sums. The observable condition

$$0 < t_k \le \min_{i,j}(P_k)_{ij} \tag{194}$$

ensures that every tangent displacement of Frobenius norm at most t_k preserves nonnegativity, since each entry of a displacement has absolute value at most its Frobenius norm. Hence Proposition H.1 gives

$$P_{k+1} = P_k - t_k \frac{G_T}{\|G_T\|_F} \qquad (G_T \neq 0). \tag{195}$$

For a buffered constraint $P_{ij} \ge \delta/n$, the analogous sufficient cap is $\min_{i,j}\{(P_k)_{ij} - \delta/n\}$, when positive. Extra fairness equalities require the projector onto the full equality nullspace, not just the double-centering operation in (193).

Computing G_T costs $\mathcal{O}(n^2)$ beyond gradient evaluation by subtracting row and column means and adding the grand mean. This is the cost of the closed-form oracle under (194); it is not a uniform bound for the boundary-constrained problem. A global LMO over the unmodified Birkhoff polytope is a linear assignment problem, while Euclidean projection is a constrained quadratic problem. For $n \ge 2$, the squared Frobenius diameter of that polytope is $2n$, so the classical smooth vanilla-FW bound involves Ln/K. Additional fairness constraints can change both global-oracle complexity and diameter.

At zero entries one instead solves the genuine local problem

$$\min_D\{\langle \nabla f(P_k), D\rangle : D\mathbf{1} = 0,\ D^\top\mathbf{1} = 0,\ P_k + D \ge 0,\ C\,\mathrm{vec}(D) = 0,\ \|D\|_F \le t_k\}. \tag{196}$$

A zero entry imposes a one-sided condition on its displacement; it does not require $t_k = 0$. Efficient active-set methods for (196), and whether the interior formula is useful for enough iterations, remain to be studied. Global projections may themselves admit efficient specialized solvers. No practical ranking or transport advantage follows from the per-step interior formula alone.

Numerical evidence. The ranking study in Appendix G.2 includes ordinary doubly stochastic plans and an additional fairness equality. Interior BroxGD-BT steps outperform the tested FW variants, while cheap affine ProjGD steps remain faster; at sparse boundary optima, away-step FW is particularly effective.

H.3 PSD-constrained metric, covariance, precision, and kernel learning

For symmetric matrices, consider a shifted PSD cone and a regularized metric-learning objective,

$$\mathcal{X}_\delta = \{M \in \mathbb{S}^d : M \succeq \delta I\}, \qquad \delta \ge 0, \tag{197}$$

$$f(M) = \frac{1}{N}\sum_{(i,j)\in\mathcal{E}} \ell\big(y_{ij}, \langle M, (a_i - a_j)(a_i - a_j)^\top\rangle\big) + \frac{\lambda}{2}\|M - M_{\mathrm{ref}}\|_F^2, \tag{198}$$

where $\mathcal{E}$ has N pairs, and ℓ is convex in its second argument. Related matrix losses arise in covariance and kernel estimation; precision estimation additionally needs a loss whose domain contains the chosen positive-definite feasible region, for example with $\delta > 0$.

Write $\sigma_k = \lambda_{\min}(M_k) - \delta$. If

$$0 < t_k \le \sigma_k, \tag{199}$$

then every symmetric D with $\|D\|_F \le t_k$ is feasible, because the eigenvalue perturbation bound gives

$$\lambda_{\min}(M_k + D) \ge \lambda_{\min}(M_k) - \|D\|_{\mathrm{op}} \ge \lambda_{\min}(M_k) - \|D\|_F \overset{(199)}{\ge} \delta.$$

Consequently, with the gradient taken in the symmetric matrix space, the exact oracle is

$$M_{k+1} = M_k - t_k \frac{\nabla f(M_k)}{\|\nabla f(M_k)\|_F} \qquad (\nabla f(M_k) \neq 0). \tag{200}$$

This update costs $\mathcal{O}(d^2)$ beyond the gradient and the slack certificate. Computing a fresh smallest eigenvalue can itself be expensive and is not included in that count.

One sufficient certificate avoids repeated eigenvalue computations. Start from $M_0 = \beta I$, with $\beta > \delta$, and choose positive radii satisfying

$$\sum_{k=0}^{\infty} t_k \leq \beta - \delta. \tag{201}$$

Every trajectory with $\|M_{j+1} - M_j\|_F \leq t_j$ then satisfies

$$\lambda_{\min}(M_k) \geq \beta - \|M_k - M_0\|_F \geq \beta - \sum_{j<k} t_j \overset{(201)}{\geq} \delta + t_k.$$

Thus the local ball is feasible at every step. This is only a feasibility certificate: a finite total path length can prevent reaching a distant minimizer. A prescribed convergence schedule must separately meet its theorem hypotheses and this total-radius budget; arbitrary summable radii have no general convergence guarantee.

Euclidean projection onto (197) is performed by eigenvalue thresholding. A global linear oracle on this unbounded cone can be unbounded below: if a symmetric cost G has a unit eigenvector v with eigenvalue $\gamma < 0$, then the feasible matrices $M_s = \delta I + svv^\top$ satisfy

$$\langle G, M_s \rangle = \delta \operatorname{tr}(G) + s\gamma \longrightarrow -\infty.$$

Adding a trace bound can make a global FW oracle available but changes the feasible problem. Near the PSD boundary, (200) need not be feasible and the exact local oracle must include the PSD constraint. The potential benefit is avoiding spectral projection during certified interior steps, not avoiding all spectral work throughout an arbitrary run.

Numerical evidence. The PSD covariance tests in Appendix G.2 retain the unbounded cone. On the interior confirmation instances, BroxGD-BT and ProjGD follow the same trajectory, with lower ProjGD overhead; boundary local solves are substantially more expensive. The correlation-matrix variant in Appendix G.3 adds unit-diagonal constraints and tests a regime where projection is much more expensive than a feasible affine-ball local solve. Metric and precision losses are not separately evaluated.

H.4 Occupancy measures for safe, fair, exploratory, or imitation RL

In a known finite discounted Markov decision process, a state-action occupancy vector satisfies linear flow equations. Safety and fairness can add linear inequalities:

$$\mathcal{X} = \{d : Ad = b,\ d \geq 0,\ Cd \leq c\}. \tag{202}$$

Representative convex coverage or imitation objectives include

$$f(d) = \tfrac{1}{2}\|\Phi d - \mu_{\text{target}}\|^2 + \tfrac{\lambda}{2}\|d - d_{\text{ref}}\|^2, \qquad \lambda \geq 0. \tag{203}$$

Here A, b encode discounted flow, and Φ maps occupancy measures to features. With $\lambda > 0$ this quadratic is strongly convex. Coordinates forced to zero by the flow equations can be eliminated before computing a minimum slack.

For positive free coordinates and strictly slack nonzero inequality rows, define

$$a_k = \min_i d_{k,i}, \qquad b_k = \min_{j:\|C_{j:}\|>0} \frac{c_j - (Cd_k)_j}{\|C_{j:}\|}, \qquad 0 < t_k \leq \min\{a_k, b_k\}. \tag{204}$$

An empty minimum is $+\infty$; feasible zero rows impose no restriction. For any displacement D with $\|D\| \leq t_k$, its coordinates have absolute value at most t_k and $C_{j:}D \leq \|C_{j:}\|t_k$. These bounds and (204) preserve all inequalities. Proposition H.1 therefore yields

$$d_{k+1} = d_k - t_k \frac{P_A \nabla f(d_k)}{\|P_A \nabla f(d_k)\|}, \tag{205}$$

when the denominator is nonzero. A factorization for the fixed flow matrix can be reused; its cost and sparsity depend on the MDP.

The potential benefit is replacing a globally constrained projection or planning subproblem by a flow-nullspace calculation on certified interior iterations. At boundary occupancies, the local problem retains $d_k + D \geq 0$ and $C(d_k + D) \leq c$ together with $AD = 0$ and $\|D\| \leq t_k$. A zero minimum-slack bound does not imply that this feasible set is a singleton. Such boundary problems need a different solver, and neither cheap factorization updates nor a small number of inner iterations is guaranteed here. This is a model-based proposal; estimating transition dynamics and optimizing from samples introduce additional errors outside the exact deterministic theory.

Numerical evidence. The controlled occupancy tests in Appendix G.2 use known transitions and weighted quadratic matching losses. Interior BroxGD-BT beats away-step FW with a policy-iteration oracle, but fixed-step ProjGD remains faster. The tests do not include additional safety inequalities or learning the dynamics from samples.

H.5 A COMMON CERTIFICATE FOR BOUNDARY-ACTIVE ATOMIC PROBLEMS

The remaining applications use a different mechanism. Let $\mathcal{A}$ be compact and $\mathcal{X} = \operatorname{conv}(\mathcal{A})$, let $x_k \in \mathcal{X}$, and choose a finite active set $S \subset \mathcal{A}$ whose convex hull contains x_k. The restricted local oracle solves

$$\min_{\beta \in \Delta_{|S|}} \left\{ \left\langle g_k, \sum_{j=1}^{|S|} \beta_j a_j \right\rangle : \left\| \sum_{j=1}^{|S|} \beta_j a_j - x_k \right\| \leq t_k \right\}, \tag{206}$$

where $\Delta_s = \{\beta \in \mathbb{R}^s : \beta \geq 0, \sum_j \beta_j = 1\}$. The master is a finite-dimensional convex second-order cone problem. Let $\bar{x}$ be its solution and let $\lambda \geq 0$ be a multiplier for the constraint $\frac{1}{2}(\|x - x_k\|^2 - t_k^2) \leq 0$. Define

$$q = g_k + \lambda(\bar{x} - x_k), \qquad \lambda(\|\bar{x} - x_k\|^2 - t_k^2) = 0. \tag{207}$$

The factor $1/2$ fixes the multiplier normalization. With $t_k > 0$ and $x_k \in \operatorname{conv}(S)$, a point in the relative interior of the restricted hull can be chosen sufficiently close to x_k to make the ball inequality strict, so the usual convex multiplier condition applies relative to that hull.

Proposition H.2 (Global certificate by atomic pricing)**.** Suppose $\bar{x} \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)$, $\lambda \geq 0$, and (207) holds. If

$$\min_{a \in \mathcal{A}} \langle q, a \rangle \geq \langle q, \bar{x} \rangle, \tag{208}$$

then $\bar{x}$ solves the full exact BroxGD with cost g_k.

Proof. By the convex-hull representation, (208) implies $\langle q, z - \bar{x} \rangle \geq 0$ for every $z \in \mathcal{X}$. For any locally feasible z, expand the squared norm to obtain

$$\langle g_k, z - \bar{x} \rangle \overset{(207)}{=} \langle q, z - \bar{x} \rangle + \frac{\lambda}{2}\big(\|\bar{x} - x_k\|^2 - \|z - x_k\|^2 + \|z - \bar{x}\|^2\big) \overset{(208),(207)}{\geq} 0.$$

For the last step, if $\lambda = 0$ the quadratic term vanishes; otherwise complementarity gives $\|\bar{x} - x_k\| = t_k \geq \|z - x_k\|$. Thus no locally feasible point improves the linear objective. □

If pricing finds a violating atom, it can be added and the master solved again. This gives a route to a certified oracle, not a bound on the number of pricing rounds. In particular, an infinite atomic set need not yield finite exact termination. Approximate master solutions or approximate pricing require explicit error certificates and corresponding inexact-oracle analysis before exact-oracle outer rates can be invoked. Modern away-step, pairwise, blended, lazy, and fully corrective FW methods also redistribute weights; comparison with vanilla FW alone cannot establish an advantage.

H.6 Robust low-rank matrix learning over a nuclear-norm ball

Consider matrix completion, multitask learning, or multivariate regression under

$$\min_{X\in\mathbb{R}^{m\times n},\ \|X\|_*\le\tau}\left\{f(X)=\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\rho(X_{ij}-Y_{ij})+\frac{\mu}{2}\|X\|_F^2\right\},\qquad \tau>0, \tag{209}$$

where Ω is nonempty, ρ is a convex robust loss, and $\mu\ge 0$. Huber loss gives a smooth example; absolute-value loss gives a nonsmooth one. The norm $\|\cdot\|_*$ is the nuclear norm, whereas the local ball uses the Frobenius norm.

At $\|X_k\|_*=\tau$, the nuclear-norm constraint is genuinely active: the points $(1+s)X_k$ with $s>0$ are outside the feasible set and approach X_k as s tends to zero. Thus the cheap interior-ball mechanism does not apply at such iterates. Instead use the atomic representation

$$\{X:\|X\|_*\le\tau\}=\operatorname{conv}\big(\{0\}\cup\{\tau uv^\top:\|u\|=\|v\|=1\}\big). \tag{210}$$

Maintain a finite convex decomposition of X_k in these atoms, including the zero atom when needed, and solve (206) in their coefficients, using Frobenius distance and $G_k=\nabla f(X_k)$ or a subgradient for a nonsmooth loss. After obtaining $\bar{X}$ and a multiplier, form

$$Q_k=G_k+\lambda(\bar{X}-X_k),\qquad A_{\mathrm{new}}=-\tau u_1v_1^\top, \tag{211}$$

where u_1,v_1 are a leading left/right singular-vector pair of Q_k. If $Q_k=0$, any atom, including zero, minimizes pricing. Otherwise this rank-one atom solves the global atomic linear problem. The certificate is

$$\langle Q_k,A_{\mathrm{new}}\rangle\ge\langle Q_k,\bar{X}\rangle. \tag{212}$$

By Proposition H.2, it certifies the exact local solution. Singular-vector computations can exploit structure, but finite power or Lanczos iterations do not by themselves give an exact pricing certificate.

Projection onto a nuclear-norm ball can be obtained by an SVD and projection of the singular values onto an ℓ_1 ball; specialized partial-SVD implementations may suffice in favorable cases. The local master offers a different way to reuse an existing decomposition, remove weights, and add directions through pricing. A smooth strongly convex loss permits a linear outer rate under Theorem D.3, provided the oracle is exact and the required bounds are known. A nonsmooth robust loss instead requires the paper's subgradient theory and its own radius hypotheses. Neither a small active set nor a small number of singular-vector calls is proved here. Runtime, rank growth, pricing accuracy, and comparison against corrective FW and specialized proximal solvers are necessary parts of a future evaluation.

Numerical evidence. The interior quadratic matrix tests in Appendix G.2 show large speedups over FW, but no consistent advantage over projected methods. Boundary Huber tests favor ProjGD-BB over BroxGD-BT. Their local solver uses a projection path; they do not test the restricted master-and-pricing implementation proposed above or statistical robustness to outliers.

H.7 Structured prediction over large products of simplices

For training examples $i=1,\dots,N$, let $\mathcal{Y}_i$ be a finite set of structured labels, possibly of exponential cardinality. A feasible domain appearing in structured-prediction duals is

$$\mathcal{X}=\prod_{i=1}^{N}\Delta(\mathcal{Y}_i). \tag{213}$$

A representative objective is

$$f(\alpha)=\tfrac{1}{2}\|B\alpha-b\|^2+\tfrac{\mu}{2}\|\alpha\|^2,\qquad \mu>0. \tag{214}$$

This explicitly regularized example is strongly convex in the simplex coordinates. A standard structural-SVM dual need not be strongly convex in those coordinates merely because its primal is regularized; a linear-rate conclusion requires checking the actual objective. Nonsmooth hinge-type or robust objectives are separate possibilities.

Sparse block distributions have many zero coordinates. In blocks with at least two labels, small equality-preserving displacements can make a zero coordinate negative, so the nonnegativity boundary cannot generally be discarded. For the gradient g_k, the local problem is

$$\min_{\beta_i \in \Delta(\mathcal{Y}_i)} \left\{ \sum_i \langle g_{k,i}, \beta_i \rangle : \sum_i \|\beta_i - \alpha_{k,i}\|^2 \le t_k^2 \right\}. \tag{215}$$

When a strictly positive ball multiplier λ exists, the KKT equations have the water-filling form

$$\beta_{i,y} = \left[\alpha_{k,i,y} - \frac{g_{k,i,y} + \nu_i}{\lambda} \right]_+, \qquad \sum_y \beta_{i,y} = 1, \qquad \sum_i \|\beta_i - \alpha_{k,i}\|^2 = t_k^2. \tag{216}$$

Here $[a]_+ = \max\{a, 0\}$ and ν_i enforces normalization. These equations follow by minimizing the separable Lagrangian with the squared-ball multiplier convention of (207). An active ball need not always have a positive multiplier; when $\lambda = 0$, division in (216) is invalid and one instead checks a global linear minimizer that meets the radius constraint.

A restricted master retains a finite support S_i containing that of $\alpha_{k,i}$. At a candidate solution $\bar{\beta}$, form

$$q_{i,y} = g_{k,i,y} + \lambda(\bar{\beta}_{i,y} - \alpha_{k,i,y}). \tag{217}$$

The blockwise global certificate is

$$\min_{y \in \mathcal{Y}_i} q_{i,y} \ge \sum_{y \in \mathcal{Y}_i} q_{i,y} \bar{\beta}_{i,y} \qquad \text{for each } i, \tag{218}$$

together with feasibility and complementarity. It implies the linear optimality condition over the product domain, so the proof of Proposition H.2 applies.

For an inactive label outside S_i, both $\bar{\beta}_{i,y}$ and $\alpha_{k,i,y}$ vanish, and hence

$$q_{i,y} \overset{(217)}{=} g_{k,i,y} \qquad (y \notin S_i).$$

This motivates loss-augmented MAP inference or decoding as a pricing primitive when the gradient has the requisite structured form. However, one must either minimize the shifted cost over all labels or find the best label outside the current support and compare it with the active labels. An ordinary MAP call can repeatedly return an active label; it does not automatically implement exclusion or arbitrary support-dependent shifts. Exact pricing therefore needs a suitable exclusion, enumeration, or modified-decoding procedure, whose cost is problem dependent.

The possible benefit is storing only active labels and using restricted water-filling or cone-program masters plus certified pricing, instead of explicitly materializing exponentially large vectors. Euclidean simplex projection is inexpensive for an explicit vector, but finding all positive coordinates can be difficult in an implicit representation. Conversely, the local oracle also may need many labels or many decoding calls; sparsity is not guaranteed by a small radius alone. Corrective FW methods can likewise reweight multiple structures. A convincing application study must measure master solves, support growth, exact or certified approximate inference, and total runtime under comparable stopping criteria.

Numerical evidence. Both Appendix G.1 and the larger sweep in Appendix G.2 favor projected methods on explicitly enumerated labels. Repeated inner projections dominate BroxGD-BT runtime. Whether restricted masters and structured inference can reverse this comparison for implicit, exponentially large label spaces remains open.

H.8 What these examples establish

The first three models exhibit conditions under which the exact local oracle has a normalized tangent-gradient formula, with explicit sufficient slack certificates. The last two give a master-and-pricing route and an exact optimality certificate while retaining active inequalities. These mechanisms identify concrete questions for further work: how often cheap local formulas apply, how costly the boundary solvers are, and whether their total cost offsets any reduction in outer iterations. They do not establish a universal computational advantage over projection or modern FW variants, nor do they change the exact-oracle assumptions of the convergence theorems.

I NOTATION GUIDE

This index groups the key notation by its role. The reference column links to the relevant definition or result; page numbers provide a second route through the paper.

1. Problem and error measures | **2. Parameters and geometry** | **3. Extensions**

Reading rate statements. A best iterate, an average iterate, and the final iterate are different outputs. The candidate sets $0 \le k < K$ and $0 \le k \le K$ are retained exactly as stated in each theorem. After deterministic termination, averages use the frozen continuation. A zero sampled step does not terminate the stochastic method.

I.1 CORE PROBLEM, ITERATES, AND ERROR MEASURES

Symbol	Meaning and scope	Reference
$f,\ f_i$	Objective and finite-sum components; $f = n^{-1}\sum_{i=1}^{n} f_i$ in the stochastic extension.	§ F p. 71
$\mathcal{X},\ \mathcal{X}_\star$	Feasible set and its set of minimizers: $\mathcal{X}_\star = \arg\min_{x\in\mathcal{X}} f(x)$.	Assump. 3.1 p. 6
$x_\star,\ f_\star$	A fixed constrained minimizer and its value $f_\star = f(x_\star)$.	Assump. 3.1 p. 6
$x_k,\ t_k$	Iterate and local radius at iteration k. Zero radius gives $x_{k+1} = x_k$.	Alg. 1 p. 3
$k,\ K$	Iteration index and integer horizon. Backtracking counts accepted steps; rejected trials incur extra oracle calls.	Thm. 3.3 p. 7
x_K^{best}	An iterate with smallest objective value among $x_0, \dots, x_K$ for the fixed-horizon guarantees; return a detected zero-gradient point if the method stops early.	Thm. 3.2 p. 6
$\bar{x}_K$	BroxGD average: $K^{-1}\sum_{k=0}^{K-1} x_k$. This includes the initial iterate and excludes x_K.	Thm. D.7 p. 48
$\bar{x}_K^{\mathsf{ProjGD}}$	ProjGD average: $K^{-1}\sum_{k=1}^{K} x_k$. This excludes the initial iterate and includes x_K.	Prop. E.1 p. 63
Δ_k	Objective gap $f(x_k) - f_\star$.	§ 3 p. 5
$v_k,\ Z_k$	Gradient-difference vector $v_k = \nabla f(x_k) - \nabla f(x_\star)$ and scalar residual $Z_k = \|v_k\|$. Neither is an objective gap.	§ 3 p. 5
$R_k,\ R_0$	Distance $R_k = \|x_k - x_\star\|$; R_0 is the initial distance.	Table 1 p. 5
$G_\gamma(x)$	Projected gradient mapping $\gamma^{-1}(x - \mathrm{Proj}_{\mathcal{X}}(x - \gamma\nabla f(x)))$, with $\gamma > 0$.	Prop. E.1 p. 63
$g_{\mathsf{FW}}(x)$	FW gap $\max_{s\in\mathcal{X}} \langle \nabla f(x), x - s\rangle$ on a compact domain.	Table 1 p. 5

I.2 GEOMETRY, REGULARITY, AND RADIUS PARAMETERS

Symbol	Meaning and scope	Reference
$d,\ \langle\cdot,\cdot\rangle,\ \|\cdot\|$	Ambient dimension, Euclidean inner product and norm. Matrix applications use the Frobenius inner product and norm unless another norm is explicitly specified.	§ H p. 98
$\mathcal{B}(x,t)$	Closed Euclidean ball $\{z : \|z-x\| \le t\}$.	Thm. C.1 p. 28
$\mathrm{Proj}_{\mathcal{X}},\ N_{\mathcal{X}}$	Euclidean projection and convex normal cone; $w \in N_{\mathcal{X}}(x)$ means $\langle w, z-x\rangle \le 0$ for all $z \in \mathcal{X}$.	Eq. (35) p. 20
$\iota_{\mathcal{X}},\ \partial f$	Convex indicator (zero on $\mathcal{X}$, infinity outside) and convex subdifferential.	§ D.2.2 p. 49
$L,\ \mu$	Positive smoothness bound and strong-convexity constant. In a projected-PL result, μ instead denotes its explicitly stated PL constant.	Assump. D.3 p. 61
$L_0,\ L_1$	Constants in the global asymmetric generalized-smoothness assumption.	Assump. D.2 p. 56
$c_k,\ c_\star$	$c_k = L_0 + L_1\|\nabla f(x_k)\|$ and $c_\star = L_0 + L_1\|\nabla f(x_\star)\|$.	Table 1 p. 5
$G,\ M_0$	Uniform gradient/subgradient bound G; smooth Polyak bound $M_0 = \|\nabla f(x_\star)\| + LR_0$.	Thm. D.9 p. 50
R	Certified input bound $R \ge R_0$, with $R > 0$, for fixed-horizon radii.	Thm. 3.2 p. 6
R_{lev}	Analysis bound on $\mathrm{dist}(x_k, \mathcal{X}_\star)$ for accepted backtracking centers; different from R_0.	Thm. 3.3 p. 7
$C_0,\ \varepsilon$	Certified initial-gap bound $C_0 \ge \Delta_0$, $C_0 > 0$, and target objective accuracy $\varepsilon > 0$.	Thm. D.3 p. 43
θ	Sharp radius multiplier $2\sqrt{\mu L}/(L+\mu)$.	Thm. 3.4 p. 8
ξ	Fixed radius-contraction factor in $(0,1)$ for backtracking; $\xi = 1/2$ gives halving.	Alg. 2 p. 7
$q,\ r$	$q = 1 - \mu/(4L)$ contracts the practical geometric objective bound; $r \in (0,1)$ contracts experimental radii. The certified schedule uses $r = \sqrt{q}$.	§ G.9 p. 94
$\gamma,\ \gamma_k$	ProjGD/GD stepsize, distinct from radius t_k. The projected-PL table row assumes its inequality at $\gamma = 1/L$.	Prop. E.1 p. 63
$\mathcal{D}_{\mathcal{X}},\ C_{\mathsf{FW}}$	Feasible-set diameter and FW curvature constant; distinct from initial-distance and initial-gap bounds.	Eq. (38) p. 23
G_K^{FW}	Trajectory gradient bound $\max_{0\le k\le K}\|\nabla f(x_k^{\mathsf{FW}})\|$ in the generalized-smooth FW comparison.	Table 1 p. 5

I.3 BACKTRACKING, STOCHASTIC QUANTITIES, AND APPLICATION GEOMETRY

Symbol	Meaning and scope	Reference
$j,\ y_{k,j},\ s_{k,j}$	Trial index, trial oracle point, and displacement $s_{k,j} = y_{k,j} - x_k$.	Eq. (108) p. 44
$a_k,\ \mathcal{D}_{k,j}$	Center gradient $a_k = \nabla f(x_k)$ and predicted decrease $\mathcal{D}_{k,j} = -\langle a_k, s_{k,j}\rangle$. Subscripts are omitted for a single trial.	Eq. (108) p. 44
$F,\ g_{k+1},\ \lambda_k$	Composite objective $F = f + \iota_{\mathcal{X}}$, certified subgradient $g_{k+1} \in \partial F(x_{k+1})$, and an analysis multiplier.	Thm. 3.3 p. 7
$\mathbb{E},\ \mathbb{P}$	Expectation and probability over sampling and any oracle-selection randomness; conditioning on $\mathcal{F}_k$ fixes all information before iteration k samples its component.	§ F p. 71
$n,\ i_k,\ \mathcal{F}_k$	Number of components, sampled component, and pre-sampling information. $\mathbb{P}(i_k = i \mid \mathcal{F}_k) = 1/n$.	§ F p. 71
$g_{k,i}$	Component gradient $\nabla f_i(x_k)$; the sampled gradient is g_{k,i_k}. Elsewhere g_k denotes the deterministic gradient or the explicitly chosen subgradient.	§ F p. 71
$x_{\star,i},\ f_i^\star$	Fixed minimizer of component f_i over $\mathcal{X}$, and its minimum value.	§ F p. 71
$H_\star$	Component-minimizer heterogeneity $\max_i \|x_{\star,i} - x_\star\|$; zero under the chosen common-minimizer convention.	§ F p. 71
$L_i,\ \mu_i$	Component smoothness and strong-convexity constants, as required by the relevant theorem.	Thm. F.5 p. 75
$L_{\max},\ \bar{L}_2^2$	$L_{\max} = \max_i L_i$ and $\bar{L}_2^2 = n^{-1}\sum_i L_i^2$.	Thm. F.6 p. 77
$\theta_i,\ \rho_i,\ \bar{\rho}$	$\theta_i = 2\sqrt{\mu_i L_i}/(L_i + \mu_i)$, $\rho_i = (L_i - \mu_i)/(L_i + \mu_i)$, and $\bar{\rho} = n^{-1}\sum_i \rho_i$.	Thm. F.5 p. 75
$\rho,\ \kappa_{\mathcal{P}},\ \alpha$	LLOO distance inflation, polytope geometry parameter, and damping fraction in the local-oracle comparison. The contraction factors ρ_i are separate stochastic quantities.	§ B.7 p. 25
η	Young-inequality parameter in $(0, 1)$ in the stochastic bounded-gradient analysis; local limiting arguments also define their own auxiliary parameters.	Thm. F.2 p. 72
$\mathcal{V},\ \mathrm{Proj}_{\mathcal{V}}$	Linear tangent space of an affine feasible set and its orthogonal projector.	§ C.8 p. 38
$A,\ P_A,\ h_k$	Application equality matrix, projector onto $\ker A$, and tangent gradient $h_k = P_A \nabla f(x_k)$.	Eq. (188) p. 98
$\delta,\ \Delta$	Locally defined feasibility buffer and simplex notation in applications; neither denotes the objective-gap sequence Δ_k.	§ H p. 98