ACCELERATED COMPOSITE OPTIMIZATION UNDER GRADIENT-DEPENDENT CURVATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study convex composite optimization under gradient-dependent curvature characterized by $\|\nabla^2 f(x)\| \le \ell(\|\nabla f(x)\|)$, which generalizes Lipschitz smoothness. Existing accelerated methods under this condition mainly focus on smooth minimization, where $\nabla f(x^\star) = 0$ and the iteration complexity is governed by $\ell(0)$. In composite optimization, the smooth gradient may not vanish at a solution, so the local curvature is generally not determined by $\ell(0)$. We show that all minimizers share the same smooth gradient, whose norm is denoted by $\tau_\star$. Based on a localization of the smooth-gradient difference around the solution, we develop an accelerated proximal method with iteration complexity $O(T_\ell + R\sqrt{\ell(\tau_\star)/\varepsilon})$, where T_ℓ is independent of ε. An adaptive variant achieves the same complexity without requiring prior knowledge of $\tau_\star$. For μ-strongly convex objectives, a restarted algorithm gives $O((1+\sqrt{\ell(\tau_\star)/\mu})\log(\Delta_0/\varepsilon) + T_{\ell,\mu})$ iteration complexity. We also establish lower bounds matching the leading terms in both the convex and strongly convex settings. Numerical experiments are included to support our theoretical findings.

Keywords composite optimization, gradient-dependent smoothness, accelerated proximal methods

1 INTRODUCTION

We study a convex composite optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) := f(x) + h(x),$$

where f is a twice continuously differentiable convex function, while h is proper, closed, convex, and possibly nonsmooth. Instead of assuming global Lipschitz smoothness of f, this paper focuses on gradient-dependent curvature, where the Hessian norm is controlled by the gradient norm (see Assumption 1). Such a condition does not provide a uniform smoothness constant, and the curvature varies along the optimization trajectory. First-order methods under this smoothness condition have been studied extensively, including gradient descent and accelerated methods (Li et al., 2023; Tyurin, 2025; Gorbunov et al., 2025; Vankov et al., 2025; Borodich & Kovalev, 2026; Tyurin, 2026).

Consider the smooth case where $h = 0$. For L-Lipschitz smooth problems, the accelerated iteration complexity scales as $O(R\sqrt{L/\varepsilon})$. Under gradient-dependent curvature, recent accelerated methods replace L by $\ell(0)$ in the complexity since $\nabla f(x^\star) = 0$ (Borodich & Kovalev, 2026; Tyurin, 2026). This conclusion no longer holds when $h \neq 0$. The optimality condition $0 \in \nabla f(x^\star) + \partial h(x^\star)$ does not require the smooth gradient to vanish, and hence $\ell(0)$ does not characterize the curvature around a solution. We will show later that all minimizers share the same smooth gradient, whose norm is denoted by $\tau_\star$. Accordingly, $\ell(\tau_\star)$ replaces $\ell(0)$ in the composite optimization.

Another issue is how to quantify the curvature, which determines the stepsize. In smooth minimization, the objective gap controls the gradient norm, which in turn bounds the local curvature. For composite problems, we propose the solution-centered gradient $\nabla f(x) - \nabla f(x^\star)$, whose norm can be controlled by the composite objective gap. Since $\|\nabla f(x)\| \le \tau_\star + \|\nabla f(x) - \nabla f(x^\star)\|$, the same bound also controls the curvature of f. The localization bound provides curvature control at the current iterate, but acceleration requires additional care because gradients are evaluated at extrapolated points. We establish a corresponding pathwise curvature bound and use it to develop an accelerated proximal method.

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.

Table 1: Comparison of first-order complexity bounds under gradient-dependent smoothness for reaching an objective gap at most ε. Here G denotes a trajectory gradient bound, Δ_0 is an upper bound on the initial objective gap, and ν is the auxiliary scalar-oracle cost in Vankov et al. (2025). The terms T and $\widehat{T}_\ell$ are independent of ε, $\tau_\star = \|\nabla f(x^\star)\|$.

Work	Smoothness condition	Complexity	Proximal algorithm
Li et al. (2023)	sub-linear ℓ	$O\left(\sqrt{\ell(2G)(\Delta_0 + R^2)/\varepsilon}\right)$	no
Gorbunov et al. (2025)	(L_0, L_1)	$O\left(\sqrt{L_0(1 + L_1 R e^{L_1 R})R^2/\varepsilon}\right)$	no
Vankov et al. (2025)	(L_0, L_1)	$\nu O\left(\sqrt{L_0 R^2/\varepsilon} + (L_1 R)^{2/3} \log(\Delta_0/\varepsilon)\right)$	no
Borodich & Kovalev (2026)	(L_0, L_1)	$O\left(\sqrt{L_0 R^2/\varepsilon} + T\right)$	no
Tyurin (2026)	general ℓ	$O\left(R\sqrt{\ell(0)/\varepsilon} + T\right)$	no
This paper	general ℓ	$O\left(R\sqrt{\ell(\tau_\star)/\varepsilon} + \widehat{T}_\ell\right)$	**yes**

Equipped with these geometric bounds, we develop accelerated proximal methods for both convex and strongly convex composite optimization under gradient-dependent curvature. We also establish matching lower bounds.

Contributions. The main contributions and theoretical results are summarized as follows.

(1) We establish a localization result showing that the objective gap controls the difference between the current smooth gradient and the smooth gradient at a minimizer. This gives both a local curvature bound and a region where the bound remains valid.

(2) Based on this localization result, we develop an accelerated proximal method for composite optimization under gradient-dependent curvature. The method requires $O(T_\ell + R\sqrt{\ell(\tau_\star)/\varepsilon})$ iterations to achieve a solution x such that $F(x) - F^\star \le \varepsilon$.

(3) Since $\tau_\star$ is generally unknown, we develop an adaptive method that estimates the relevant solution scale from observable quantities. We show that the adaptive method keeps the same accelerated rate, with only an accuracy-independent additional cost.

(4) For strongly convex problems, we combine the accelerated method with a restart scheme and prove that the resulting algorithm has the iteration complexity $O(T_{\ell,\mu} + (1 + \sqrt{\ell(\tau_\star)/\mu}) \log(\Delta_0/\varepsilon))$.

(5) We also derive lower bounds for deterministic proximal first-order methods under gradient-dependent curvature. These lower bounds match the leading-order complexity of our upper bounds, showing that the dependence on the solution-scale curvature is optimal up to constant factors.

(6) We provide Lean4 (Moura & Ullrich, 2021) formalizations of our definitions, algorithms, and theoretical results. Each formalized result is linked directly from its statement in the paper, and all formalized theorems and propositions are verified by the Lean 4 kernel.

Related work. Accelerated first-order methods are well understood under global Lipschitz smoothness. For smooth convex minimization, Nesterov acceleration achieves the optimal $O(R\sqrt{L/\varepsilon})$ iteration complexity. The same dependence on L can be retained for composite objectives by replacing the gradient step with a proximal step, leading to accelerated proximal gradient methods (Tseng, 2008; Beck & Teboulle, 2009; Nesterov, 2013). Several extensions consider composite optimization without the Lipschitz smoothness condition. For example, Lu & Mei (2023) study accelerated proximal-gradient methods when the smooth component has a locally Lipschitz continuous gradient, using backtracking to estimate a local smoothness constant for each iteration. Another line of work considers relative smoothness, where the smooth component is measured relative to a reference function rather than a Euclidean quadratic, leading to Bregman proximal updates instead of gradient update steps (Bauschke et al., 2017; Lu et al., 2018; Hanzely et al., 2021).

Gradient-dependent smoothness considered in this paper was introduced through the (L_0, L_1)-smooth model in Zhang et al. (2020) and later generalized to curvature functions ℓ in Li et al. (2023). Early complexity bounds under this condition depend on curvature evaluated at the initial or trajectory-dependent gradient (Li et al., 2023; Tyurin, 2025). More recent accelerated methods improve this dependence for smooth convex optimization, with rates governed by $\ell(0)$ (Gorbunov et al., 2025; Vankov et al., 2025; Borodich & Kovalev, 2026; Tyurin, 2026). Since $\nabla f(x^\star) = 0$ in smooth minimization and ℓ is nondecreasing, $\ell(0)$ is the smallest curvature bound under this model. This improvement removes the dependence on potentially much larger gradients away from the solution. To the best of our knowledge, accelerated proximal methods under this gradient-dependent curvature model have not been systematically studied. The main related accelerated results are summarized in Table 1. Notice that the accelerated result of Li et al. (2023) additionally requires sub-linear growth of ℓ, whereas the other works in the table either specialize to the (L_0, L_1)-smooth setting or do not impose this additional growth condition.

2 Composite Localization under Gradient-Dependent Curvature

Assumption 1 (Gradient-dependent curvature)**.** Lean ↗ There exists a continuous, nondecreasing function $\ell : [0, \infty) \to (0, \infty)$ such that $\left\|\nabla^2 f(x)\right\| \le \ell\left(\|\nabla f(x)\|\right)$ for all $x \in \mathbb{R}^d$.

Throughout this paper, unless stated otherwise, we assume that for some $x^\star \in X^\star$, the following initial condition holds

$$\|x_0 - x^\star\| \le R, \qquad F(x_0) - F^\star \le \Delta_0. \tag{1}$$

The curvature bound is expressed in terms of the absolute smooth gradient $\|\nabla f(x)\|$. For the case $h = 0$, a substantial body of work has established convergence rates governed by $\ell(0)$, since the smooth gradient vanishes at an optimal solution. In composite optimization, however, the optimality condition $0 \in \nabla f(x^\star) + \partial h(x^\star)$ does not imply $\nabla f(x^\star) = 0$. We therefore first identify the smooth-gradient scale associated with the solution set, which allows us to define $g_\star := \nabla f(x^\star)$ and $\tau_\star := \|g_\star\|$ independently of the choice of $x^\star \in X^\star$. In the rest of the paper, we denote $L_\star := \ell(\tau_\star)$ for the curvature scale at the solution. All proofs are deferred to Appendix A.

Proposition 2. Lean ↗ *All optimal solutions in $X^\star$ share the same smooth gradient, i.e., $\nabla f(x^\star) = \nabla f(\widetilde{x}^\star)$, for all $x^\star, \widetilde{x}^\star \in X^\star$*

For a composite problem with Lipschitz-smooth f, stationarity is commonly measured by the proximal-gradient mapping

$$G_\eta(x) := \frac{1}{\eta}\left(x - \operatorname{prox}_{\eta h}\big(x - \eta \nabla f(x)\big)\right),$$

which is zero at the optimal solution. Under Assumption 1, the curvature is controlled by $\|\nabla f(x)\|$, while $G_\eta(x)$ does not directly control the variation of this quantity. We therefore work instead with the solution-centered gradient $\nabla f(x) - g_\star$. This residual vanishes on $X^\star$ and satisfies $\|\nabla f(x)\| \le \tau_\star + \|\nabla f(x) - g_\star\|$. Thus, controlling the solution-centered gradient norm also controls the curvature. The following affine reparameterization turns this residual into the gradient of a convex smooth function while preserving both the composite objective and the proximal-gradient oracle.

Set $v^\star := -g_\star \in \partial h(x^\star)$. Define the surrogate functions

$$\phi(x) := f(x) + \langle v^\star, x - x^\star \rangle, \qquad \psi(x) := h(x) - \langle v^\star, x - x^\star \rangle.$$

Then the objective function is represented as $F = \phi + \psi$, with $\nabla\phi(x) = \nabla f(x) - g_\star$. Affine centering turns the solution-centered gradient into $\nabla\phi(x)$, which vanishes at the solution as $\nabla\phi(x^\star) = 0$.

Lemma 3. Lean ↗ *The centered functions ϕ and ψ are convex. $x^\star$ is the minimizer for both of them. For all $x \in \mathbf{dom}\, h$, the optimality gap satisfies $0 \le \phi(x) - \phi(x^\star) \le F(x) - F^\star$. Furthermore, the proximal-gradient oracle is invariant under this reparametrization, that is, for all $w, y \in \mathbf{dom}\, h$, $\operatorname{prox}_{a\psi}\big(w - a\nabla\phi(y)\big) = \operatorname{prox}_{ah}\big(w - a\nabla f(y)\big)$.*

Affine centering removes the nonzero solution gradient from the residual, but the solution scale $\tau_\star$ remains as the baseline in the curvature bound. Indeed, since $\nabla^2\phi = \nabla^2 f$ and $\nabla f(x) = g_\star + \nabla\phi(x)$, Assumption 1 together with the triangle inequality gives $\left\|\nabla^2\phi(x)\right\| \le \ell\left(\tau_\star + \|\nabla\phi(x)\|\right)$. More generally, for any $\tau \ge \tau_\star$, it holds

$$\left\|\nabla^2\phi(x)\right\| \le \ell\left(\tau + \|\nabla\phi(x)\|\right). \tag{2}$$

Hence the original gradient-dependent curvature model is reduced to a solution-centered model with baseline scale τ. This reduction therefore leaves us with the task of deriving a localization bound that controls the residual $\|\nabla\phi(x)\|$ in terms of the composite objective gap $F(x) - F^\star$.

Under L-Lipschitz smoothness, the objective gap satisfies

$$\phi(x) - \phi(x^\star) \geq \frac{\|\nabla\phi(x)\|^2}{2L}.$$

With the smoothness condition considered in this paper, there is no uniform curvature constant that can fit into this argument. The quadratic self-bounding relation above must be replaced by a quantity that accumulates the varying local curvature. To capture this relationship, we define Lean ↗

$$Q_\tau(s) := \int_0^s \frac{u}{\ell(\tau+u)}\,du, \qquad Q_{\tau,\infty} := \int_0^\infty \frac{u}{\ell(\tau+u)}\,du.$$

As the map Q_τ is continuous and strictly increasing, it admits an inverse on $[0, Q_{\tau,\infty})$. When $\ell \equiv L$, $Q_\tau(s) = \frac{s^2}{2L}$, which reduces to the self-bounding inequality for Lipschitz smooth functions.

Theorem 4. Lean ↗ *For any $\tau \geq \tau_\star$ and every $x \in \mathbf{dom}\, h$, the composite objective gap satisfies the following bound: $F(x) - F^\star \geq Q_\tau(\|\nabla f(x) - \nabla f(x^\star)\|)$. For every $\delta \in [0, Q_{\tau,\infty})$, whenever $F(x) - F^\star \leq \delta$, we have $\|\nabla f(x) - \nabla f(x^\star)\| \leq Q_\tau^{-1}(\delta)$.*

The localization theorem provides control of the solution-centered gradient norm at a single point. The next lemma extends this pointwise control along a path and provides a uniform curvature bound.

Lemma 5. Lean ↗ *For $s \geq 0$, define $\rho_\tau(s) := \int_s^{2s} du/\ell(\tau+u)$. Let $c : [0,1] \to \mathbb{R}^d$ be piecewise continuously differentiable. If $\|\nabla\phi(c(0))\| \leq s$ and $\int_0^1 \|\dot c(t)\|\,dt \leq \rho_\tau(s)$, then $\sup_{t\in[0,1]} \|\nabla\phi(c(t))\| \leq 2s$ and $\sup_{t\in[0,1]} \left\|\nabla^2\phi(c(t))\right\| \leq \ell(\tau+2s)$.*

3 Accelerated Proximal Gradient via Composite Localization

We now use the localization and pathwise bounds from Section 2 to construct an accelerated proximal method. Throughout this section, the solution scale $\tau_\star$ is assumed to be available. We first develop the accelerated method and its rate for convex objectives, then apply the same method in a restarted scheme for strongly convex objectives.

3.1 Convex objectives

Under Assumption 1, the curvature bound is local and depends on the smooth gradient. Therefore, an accelerated step must ensure that the corresponding bound remains valid along the update path. To initialize the localized accelerated scheme, choose Γ_0 satisfying

$$\Gamma_0 \geq \frac{2\Delta_0}{R^2}, \qquad \Gamma_0 R^2 < Q_{\tau_\star,\infty}. \tag{3}$$

At iteration k of Algorithm 1, the gap scale $\delta_k = R^2/A_k$ determines a solution-centered gradient norm bound M_k, which in turn specifies the local curvature $L_k = \ell(\tau_\star + 2M_k)$ and the path-length bound $\rho_k = \rho_{\tau_\star}(M_k)$. The weight increment α_k is chosen to satisfy the resulting curvature and path-length restrictions, while the remaining updates follow the standard three-sequence accelerated proximal scheme.

We next establish the one-step potential inequality, which is a key component of the convergence analysis. The following lemma identifies the local quadratic model condition sufficient for the accelerated potential to contract.

Lemma 6. Lean ↗ *Define $A^+ = A + \alpha$, $y = (Ax + \alpha z)/A^+$, $z^+ = \mathrm{prox}_{\alpha h}(z - \alpha\nabla f(y))$, and $x^+ = (Ax + \alpha z^+)/A^+$. If $L\alpha^2 \leq A^+$ and $f(x^+) \leq f(y) + \langle\nabla f(y), x^+ - y\rangle + L\|x^+ - y\|^2/2$, then*

$$A^+(F(x^+) - F^\star) + \left\|z^+ - x^\star\right\|^2/2 \leq A(F(x) - F^\star) + \|z - x^\star\|^2/2.$$

Algorithm 1 Accelerated proximal method with composite localization Lean ↗

Require: x_0, R, $\tau_\star$, and Γ_0
$A_0 \leftarrow 1/\Gamma_0$, $z_0 \leftarrow x_0$
for $k = 0, 1, \dots, N-1$ **do**
  $\delta_k \leftarrow R^2/A_k$
  $M_k \leftarrow Q_{\tau_\star}^{-1}(\delta_k)$, $L_k \leftarrow \ell(\tau_\star + 2M_k)$, and $\rho_k \leftarrow \rho_{\tau_\star}(M_k)$
  $\bar{\alpha}_k \leftarrow (1 + \sqrt{1 + 4A_kL_k})/(2L_k)$
  Choose the largest $\alpha_k \in (0, \min\{A_k, \bar{\alpha}_k\}]$ satisfying $8R\alpha_k \le \rho_k(A_k + \alpha_k)$
  $A_{k+1} \leftarrow A_k + \alpha_k$
  $y_k \leftarrow (A_kx_k + \alpha_kz_k)/A_{k+1}$
  $z_{k+1} \leftarrow \mathrm{prox}_{\alpha_k h}(z_k - \alpha_k\nabla f(y_k))$
  $x_{k+1} \leftarrow (A_kx_k + \alpha_kz_{k+1})/A_{k+1}$
end for
Return x_N

Lemma 6 shows that the accelerated potential decreases whenever its local quadratic upper model is valid. Under Assumption 1, such a model is not available globally, since the curvature depends on $\|\nabla f(\cdot)\|$. The localization result controls this scale at x_k, while Lemma 5 extends the resulting curvature bound along a sufficiently short path. Accordingly, we track the potential arising in Lemma 6, which is denoted as

$$\mathcal{E}_k := A_k\big(F(x_k) - F^\star\big) + \frac{1}{2}\|z_k - x^\star\|^2.$$

The next lemma establishes a bound on $\mathcal{E}_k$ to control the length of the accelerated path, and allows Lemma 5 to verify the required local quadratic model.

Lemma 7. Lean ↗ *Suppose that $\mathcal{E}_k \le R^2$, $\|x_k - x^\star\| \le \sqrt{2}R$, and $\|\nabla\phi(x_k)\| \le M_k$. Perform one step of Algorithm 1, then $\mathcal{E}_{k+1} \le \mathcal{E}_k$, $F(x_{k+1}) - F^\star \le R^2/A_{k+1}$, $\|z_{k+1} - x^\star\| \le \sqrt{2}R$, and $\|x_{k+1} - x^\star\| \le \sqrt{2}R$. Moreover, $\|\nabla\phi(y_k)\| \le 2M_k$. Hence, $\|\nabla f(y_k) - \nabla f(x^\star)\| \le 2M_k$.*

Remark 8. All of the conclusions in Lemma 7 are valid when $\tau_\star$ is replaced by any upper bound $\tau \ge \tau_\star$, with $L_k := \ell(\tau + 2M_k)$ and $\rho_k := \rho_\tau(M_k)$.

Lemma 7 shows the local model and the potential estimate hold in a single step. Combined with the localization theorem, this allows the one-step argument to be iterated globally. As the objective gap decreases, the localized curvature approaches the solution-scale curvature $\ell(\tau_\star)$. The following theorem shows that, after finitely many iterations, the method enters a region where the local curvature is controlled by the solution-scale curvature, and attains the standard accelerated rate governed by $\ell(\tau_\star)$.

Theorem 9. Lean ↗ *Under Assumption 1, suppose Γ_0 satisfies* (3)*. Then there exists a finite constant T_ℓ such that for every $N > T_\ell$, $F(x_N) - F^\star \le 512\,\ell(\tau_\star)R^2/(N - T_\ell)^2$. Equivalently, Algorithm 1 reaches $F(x_N) - F^\star \le \varepsilon$ with complexity $N_\varepsilon = O(T_\ell + R\sqrt{\ell(\tau_\star)/\varepsilon})$, where T_ℓ is independent of ε.*

We next specialize the result to the (L_0, L_1)-smooth setting, where $\ell(s) = L_0 + L_1 s$.

Corollary 10. Lean ↗ *Let $\ell(s) = L_0 + L_1 s$ with $L_1 \ge 0$, set $L := L_0 + L_1\tau_\star$, and suppose that $\varepsilon > 0$ satisfies $L_1^2\varepsilon < L$. Then Algorithm 1 reaches $F(x_N) - F^\star \le \varepsilon$ with complexity*

$$N_\varepsilon = O(R\sqrt{L/\varepsilon} + (1 + L_1R)\log(2 + \Gamma_0(1 + L_1R)^2/L)).$$

3.2 Strongly convex objectives

We next extend the aforementioned convex result to the case where F is μ-strongly convex through the restart scheme in Algorithm 2. At a restart stage, let Δ denote the current objective-gap upper bound, so that $F(\bar{x}) - F^\star \le \Delta$. Strong convexity gives $\|\bar{x} - x^\star\| \le R_\Delta := \sqrt{2\Delta/\mu}$. Hence Algorithm 1 can be initialized with $x_0 = z_0 = \bar{x}$, using $(\bar{x}, R_\Delta, \Delta)$ in place of (x_0, R, Δ_0) in (1), and its run until the resulting objective-gap bound is at most $\Delta/2$.

Algorithm 2 Restarted accelerated proximal method Lean ↗

Require: x_0, μ, $\tau_\star$, Δ_0, and ε, with $2\Delta_0 < Q_{\tau_\star,\infty}$
1: $\bar{x}_0 \leftarrow x_0$
2: $J \leftarrow \lceil \log_2(\Delta_0/\varepsilon) \rceil$
3: **for** $j = 0, 1, \ldots, J-1$ **do**
4: $\quad \Delta_j \leftarrow 2^{-j}\Delta_0$ and $R_j \leftarrow \sqrt{2\Delta_j/\mu}$
5: $\quad$ Starting from $x_0 = z_0 = \bar{x}_j$, run Algorithm 1 with $R = R_j$ and $\Gamma_0 = \mu$ until $A_k \geq 4/\mu$
6: $\quad \bar{x}_{j+1} \leftarrow x_k$
7: **end for**
8: **Return** $\bar{x}_J$

Although both the curvature scale and the path length vary from stage to stage under Assumption 1, the analysis of convex problem provides the contraction guarantee for each stage. The following theorem gives the resulting iteration complexity.

Theorem 11. Lean ↗ *Suppose that F is μ-strongly convex, ℓ is locally Lipschitz at $\tau_\star$, and $2\Delta_0 < Q_{\tau_\star,\infty}$. Then there exists a finite constant $T_{\ell,\mu}$ such that for any $\varepsilon < \Delta_0$, Algorithm 2 reaches $F(\bar{x}_J) - F^\star \leq \varepsilon$ with complexity $N_\varepsilon = O((1 + \sqrt{\ell(\tau_\star)/\mu}) \log(\Delta_0/\varepsilon) + T_{\ell,\mu})$.*

Based on this result, restart scheme has the usual logarithmic dependence on the target accuracy, with the asymptotic stage complexity determined by $\ell(\tau_\star)/\mu$. The non-uniform curvature contributes only through the accuracy-independent constant $T_{\ell,\mu}$. In the (L_0, L_1)-smooth setting, this constant can be bounded explicitly.

Corollary 12. Lean ↗ *Under the assumptions of Theorem 11, let $\ell(s) = L_0 + L_1 s$ with $L_1 \geq 0$. Set $L := L_0 + L_1\tau_\star$ and $R_0 := \sqrt{2\Delta_0/\mu}$. Then, for any $\varepsilon < \Delta_0$, Algorithm 2 reaches $F(\bar{x}_J) - F^\star \leq \varepsilon$ with complexity $N_\varepsilon = O((1 + \sqrt{L/\mu}) \log(\Delta_0/\varepsilon) + L_1 R_0)$.*

4 Adaptive Accelerated Proximal Gradient

The method developed in Section 3 relies on prior knowledge of the solution scale $\tau_\star$, which need not be available in practice. We remove this dependence by adapting to the unknown solution scale while maintaining the accelerated convergence rate. The key idea is to track a computable upper bound of $\tau_\star$ throughout the iteration. The construction is based on a generalized cocoercivity bound with the curvature centered at $c = \|\nabla f(x)\|$. For $c \geq 0$, define

$$P_c(s) := \int_0^s \frac{s-u}{\ell(c+u)}\, du = s^2 \int_0^1 \frac{1-v}{\ell(c+vs)}\, dv.$$

The following proposition shows P_c controls the variation of the smooth gradient through the objective gap. Moreover, P_c is globally invertible, which allows this relation to be converted into a computable upper bound of $\tau_\star$.

Proposition 13. Lean ↗ *For every $x, y \in \mathbb{R}^d$, let $c := \|\nabla f(x)\|$ and $r := \|\nabla f(x) - \nabla f(y)\|$. Then $P_c(r) \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle$. Furthermore, for every $x \in \mathbf{dom}\, h$ such that $F(x) - F^\star \leq \delta$, it holds $\tau_\star \leq \|\nabla f(x)\| + P^{-1}_{\|\nabla f(x)\|}(\delta)$.*

Proposition 13 provides a rule for determining an initial upper bound on $\tau_\star$, as specified in the following construction.

Corollary 14. Lean ↗ *Choose $\Gamma_0 \geq 2\Delta_0/R^2$, define $\delta_0 := \Gamma_0 R^2$ and $c_0 := \|\nabla f(x_0)\|$. It holds that $\widehat{\tau}_0 := c_0 + P^{-1}_{c_0}(\delta_0) \geq \tau_\star$.*

The upper bound $\widehat{\tau}_0$ does not by itself guarantee that the localized accelerated iteration can be initialized, since the initial gap scale δ_0 may lie outside the range of $Q_{\widehat{\tau}_0}$. If $\delta_0 < Q_{\widehat{\tau}_0,\infty}$, the method can start directly with the safe scale $\widehat{\tau}_0$. Otherwise, we introduce a warm-start mechanism, which maintains a valid upper bound on $\tau_\star$ while reducing the gap until $Q^{-1}_{\widehat{\tau}_k}(\delta_k)$ is well defined. The complete procedure is given in Algorithm 3 and the corresponding iteration complexity is presented in Theorem 15.

Algorithm 3 Adaptive accelerated proximal method Lean ↗

Require: x_0, R, and Γ_0

$A_0 \leftarrow 1/\Gamma_0$, $z_0 \leftarrow x_0$, $\delta_0 \leftarrow \Gamma_0 R^2$
$c_0 \leftarrow \|\nabla f(x_0)\|$, $M_0 \leftarrow P_{c_0}^{-1}(\delta_0)$, $\widehat{\tau}_0 \leftarrow c_0 + M_0$
$\delta^\circ \leftarrow Q_{\widehat{\tau}_0}(M_0)/2$, warm $\leftarrow [\delta_0 \geq Q_{\widehat{\tau}_0,\infty}]$
for $k = 0, 1, \ldots, N-1$ **do**
  if warm and $\delta_k > \delta^\circ$ **then**
    $c_k \leftarrow \|\nabla f(x_k)\|$, $M_k \leftarrow P_{c_k}^{-1}(\delta_k)$
    $\widehat{\tau}_k \leftarrow \min\{\widehat{\tau}_k, c_k + M_k\}$
  else
    warm $\leftarrow$ false, $M_k \leftarrow Q_{\widehat{\tau}_k}^{-1}(\delta_k)$
  end if
  $L_k \leftarrow \ell(\widehat{\tau}_k + 2M_k)$, $\rho_k \leftarrow \rho_{\widehat{\tau}_k}(M_k)$
  $\bar{\alpha}_k \leftarrow (1 + \sqrt{1 + 4A_k L_k})/(2L_k)$
  Choose the largest $\alpha_k \in (0, \min\{A_k, \bar{\alpha}_k\}]$ such that $8R\alpha_k \leq \rho_k(A_k + \alpha_k)$
  $A_{k+1} \leftarrow A_k + \alpha_k$, $y_k \leftarrow (A_k x_k + \alpha_k z_k)/A_{k+1}$
  $z_{k+1} \leftarrow \mathrm{prox}_{\alpha_k h}(z_k - \alpha_k \nabla f(y_k))$, $x_{k+1} \leftarrow (A_k x_k + \alpha_k z_{k+1})/A_{k+1}$
  $\delta_{k+1} \leftarrow R^2/A_{k+1}$
  $\widehat{\tau}_{k+1} \leftarrow \begin{cases} \widehat{\tau}_k, & \text{warm}, \\ \min\{\widehat{\tau}_k, \|\nabla f(y_k)\| + 2M_k\}, & \text{otherwise}. \end{cases}$
end for
Return x_N

Theorem 15. Lean ↗ *Under Assumption 1, choose $\Gamma_0 \geq 2\Delta_0/R^2$. Then Algorithm 3 is well-defined. There exists a finite constant $\widehat{T}_\ell$ such that for every $N > \widehat{T}_\ell$, $F(x_N) - F^\star \leq 512\,\ell(\tau_\star)R^2/(N - \widehat{T}_\ell)^2$. Consequently, the algorithm reaches $F(x_N) - F^\star \leq \varepsilon$ with complexity $N_\varepsilon = O(\widehat{T}_\ell + R\sqrt{\ell(\tau_\star)/\varepsilon})$.*

We also specialize the adaptive result to the (L_0, L_1)-smooth setting, where the initialization and the resulting complexity admit explicit bounds.

Corollary 16. Lean ↗ *Under the assumptions of Theorem 15, let $\ell(s) = L_0 + L_1 s$ with $L_1 \geq 0$. Then $Q_{\tau,\infty} = +\infty$ for every finite τ, so there is no warm-start steps. With $c_0 := \|\nabla f(x_0)\|$, the choice $\widehat{\tau}_0 := c_0 + L_1\delta_0 + \sqrt{L_1^2\delta_0^2 + 2(L_0 + L_1 c_0)\delta_0}$ satisfies $\widehat{\tau}_0 \geq \tau_\star$. Algorithm 3 reaches $F(x_N) - F^\star \leq \varepsilon$ with complexity $N_\varepsilon = O(R\sqrt{(L_0 + L_1\tau_\star)/\varepsilon} + \widehat{T}_\ell)$.*

5 Oracle Complexity Lower Bounds

The previous sections established upper bounds for composite optimization under gradient-dependent curvature. We now derive lower bounds for deterministic proximal first-order methods. These results show that the dependence on $\ell(\tau_\star)$ achieved by our algorithms is optimal up to some constants.

We begin by specifying the oracle model and the class of algorithms under consideration. An instance in dimension d is represented by $P = (f, h, x^{(0)})$, where $x^{(0)}$ is the initial point and $F := f + h$ is the corresponding objective with optimal value $F^\star := \inf_x F(x)$. The algorithm accesses f through the first-order oracle $\mathcal{O}_f^{(1)}(x) := (f(x), \nabla f(x))$ and h through the proximal oracle $\mathcal{P}_h(w, a) := \mathrm{prox}_{ah}(w)$, where $a > 0$. We measure complexity by the number of calls to $\mathcal{O}_f^{(1)}$. Let $A[P] = \{x^{(t)}\}_{t=0}^\infty$ denote the sequence generated by an algorithm A on instance P. This section considers deterministic proximal first-order methods in the black-box oracle model. More precisely, the algorithm may query either $\mathcal{O}_f^{(1)}$ or $\mathcal{P}_h$, and may perform any finite number of proximal-oracle calls between two consecutive first-order oracle calls. Let $\mathcal{A}_p(d)$ denote this class of algorithms in dimension d, and $\mathcal{A}_p$ denote the corresponding dimension-free class. The dimension-free class allows the method to depend on the dimension, while the worst-case below extends over all finite dimensions.

We next introduce the problem classes over which the lower bounds are stated. For $R > 0$, denote by $\mathcal{P}_{\ell,\tau}(R)$ the class of finite-dimensional instances satisfying Assumption 1 with $\tau_\star \le \tau$ and $\mathbf{dist}(x^{(0)}, X^\star) \le R$. For $\mu, \Delta_0 > 0$, let $\mathcal{P}_{\ell,\tau,\mu}(\Delta_0)$ denote the subclass of μ-strongly convex instances satisfying Assumption 1, together with $F(x^{(0)}) - F^\star \le \Delta_0$.

It remains to specify the complexity measure. For an algorithm A and an instance P, its ε-complexity is

$$T_\varepsilon(A, P) := \inf\left\{t \ge 0 : F(x^{(t)}) - F^\star \le \varepsilon\right\}.$$

Accordingly, for an algorithm class $\mathcal{A}$ and a problem class $\mathcal{P}$, we define the worst-case oracle complexity by

$$T_\varepsilon(\mathcal{A}, \mathcal{P}) = \inf_{A\in\mathcal{A}} \sup_{P\in\mathcal{P}} T_\varepsilon(A, P).$$

With these notations, we are ready to state the lower bounds for the composite problem.

Theorem 17. Lean ↗ *There exists a constant $c_0 > 0$ such that, for the convex problem class $\mathcal{P}_{\ell,\tau}(R)$, if $0 < \varepsilon < c_0\ell(\tau)R^2$, then $T_\varepsilon(\mathcal{A}_p, \mathcal{P}_{\ell,\tau}(R)) = \Omega(R\sqrt{\ell(\tau)/\varepsilon})$. For the μ-strongly convex problem class $\mathcal{P}_{\ell,\tau,\mu}(\Delta_0)$, if $\ell(\tau)/\mu \ge 4$ and $0 < \varepsilon < c_0\Delta_0$, then $T_\varepsilon(\mathcal{A}_p, \mathcal{P}_{\ell,\tau,\mu}(\Delta_0)) = \Omega(\sqrt{\ell(\tau)/\mu}\log(\Delta_0/\varepsilon))$.*

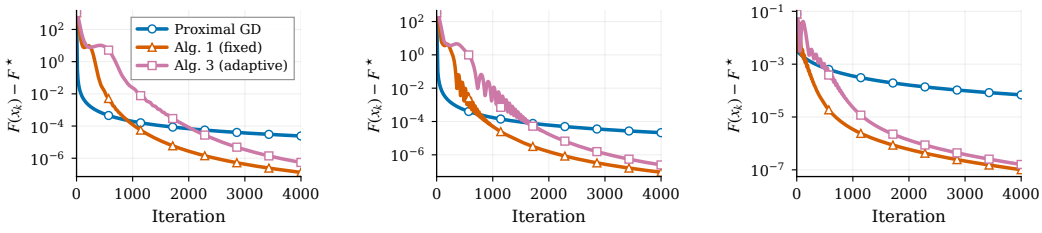
6 Numerical Experiments

We compare the proposed accelerated methods with proximal gradient descent on both convex and strongly convex composite problems. For each problem, all methods start from the same initial point, and we plot the objective gap $F(x_k) - F^\star$ against the iteration number. For the proximal-GD baseline, we use the gradient-dependent stepsize of Tyurin (2025) together with the standard proximal update. In both settings, we add regularizers so that $\tau_\star > 0$ in every problem. Detailed objective definitions and parameter settings are provided in Appendix B.

Convex composite objectives. We consider three two-dimensional examples based on exponential, quartic, and log-cosine functions, with $h(x, y) = |x|$. The smooth components and their corresponding curvature functions are given by

$$\begin{aligned} f(x,y) &= e^x + e^{1-x} + \tfrac{\lambda}{6}|y|^6, && \ell(s) = 2\sqrt{e} + s, \\ f(x,y) &= \tfrac{1}{4}(x-2)^4 + \tfrac{\lambda}{6}|y|^6, && \ell(s) = 4 + s, \\ f(x,y) &= -\log\cos(x-c) + \tfrac{\lambda}{8}|y|^8, && \ell(s) = 1 + s^2, \end{aligned}$$

where $c = 1/2 + \pi/4$, and the log-cosine example is defined on $(c - \pi/2, c + \pi/2) \times \mathbb{R}$. The coefficient $\lambda > 0$ is specified for each example in the appendix. The first two examples satisfy an (L_0, L_1)-smooth condition, whereas the log-cosine example has a nonlinear curvature function ℓ.



(a) Exponential (b) Quartic (c) Log-cosine.

Figure 1: Comparison of the proposed accelerated proximal methods and proximal gradient descent under gradient-dependent curvature.

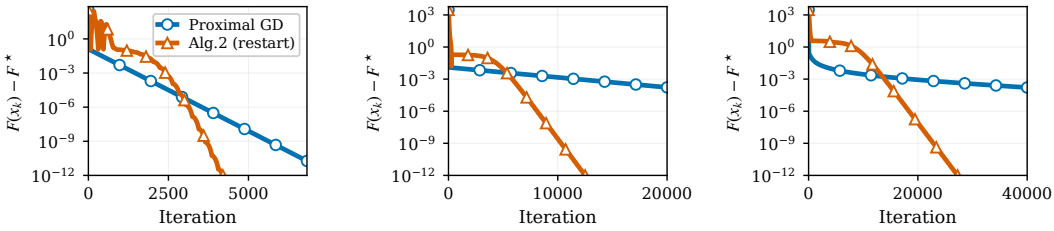
The algorithm performances are presented in Figure 1. The results show that proximal GD can decrease the objective gap faster during the initial iterations, particularly for the exponential and quartic examples. The accelerated methods catch up after this initial phase and eventually achieve

substantially smaller gaps. The fixed-scale method converges faster than the adaptive method, as it is given the exact value of $\tau_\star$. Nevertheless, the gap between the two accelerated methods gradually narrows, showing that the adaptive method can recover comparable acceleration while estimating the solution-gradient scale during the iterations.

Strongly convex composite objectives. For strongly convex problems, we conduct experiments on three examples, quartic, exp-square, and ellipsoid-exp. Their definitions and corresponding curvature functions are

$$
\begin{aligned}
f(x,y) &= \tfrac{1}{4}(x-2)^4 + \tfrac{1}{2}(x-2)^2 + \tfrac{\mu}{2}y^2, && \ell(s) = 2 + \tfrac{4\sqrt{6}}{9} + s,\\
f(x,y) &= e^{x^2/2} - 1 + \tfrac{\mu}{2}y^2, && \ell(s) = (1+W(s^2))e^{W(s^2)/2},\\
f(u,v) &= e^{u^\top Au/2} - 1 + \tfrac{1}{2}v^\top Dv, && \ell(s) = 4(1+W(s^2))e^{W(s^2)/2}.
\end{aligned}
$$

Here $A, D \succ 0$, and W is the principal Lambert function (see the details in the Appendix). The smooth component of quartic satisfies an (L_0, L_1)-smooth condition, while the other two satisfy a generalized ℓ-smoothness condition.



(a) Quartic (b) Exp-square (c) Ellipsoid-exp

Figure 2: Comparison of the proposed accelerated proximal method and proximal gradient descent on strongly convex composite problems under gradient-dependent curvature.

Figure 2 above shows a comparison between our Algorithm 2 and the proximal gradient descent method. From these plots, we can observe that after an initial phase, the restart algorithm decreases the objective gap faster than proximal GD on all examples. These verify that our proposed algorithm indeed achieves an accelerated convergence rate.

7 Conclusion

In this paper, we studied convex composite optimization under gradient-dependent curvature. The main difficulty compared with smooth minimization is that the smooth gradient does not generally vanish at a solution, and the relevant solution-scale curvature is therefore determined by $\ell(\tau_\star)$. Using a solution-centered gradient, we established localization and pathwise curvature bounds. We used these bounds to develop accelerated proximal methods for both convex and strongly convex problems. We also proposed an adaptive variant that does not require prior knowledge of $\tau_\star$. The resulting methods achieve accelerated rates with leading complexity governed by $\ell(\tau_\star)$, and the lower bounds show that this dependence is optimal up to constant factors.

AI disclosure

In this work, we used GPT-5.6 to assist with the formalization of proofs in Lean through Lean code generation. We also used the same model for language polishing and editing, as well as for polishing and editing the experimental code. The authors reviewed all AI-assisted work. Lean formalizations were checked by the Lean proof checker and inspected by the authors. Generative AI was not used to formulate the main research hypotheses, design the research methodology, or interpret the experimental results; other required-disclosure tasks were not applicable. The authors take full responsibility for the final content of this paper, including text, claims, proofs, and code produced with the aid of generative AI.

REFERENCES

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, May 2017. doi: 10.1287/moor.2016.0817.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, January 2009. doi: 10.1137/080716542.

Ekaterina Borodich and Dmitry Kovalev. Nesterov finds GRAAL: Optimal and adaptive gradient method for convex optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 95533–95558, 2026.

Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017. doi: 10.1016/j.jco.2016.11.001.

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 *International Conference on Learning Representations*, 2025.

Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, April 2021. doi: 10.1007/s10589-021-00273-8.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems 36*, pp. 40238–40271. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1749.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, January 2018. doi: 10.1137/16m1099546.

Zhaosong Lu and Sanyou Mei. Accelerated first-order methods for convex optimization with locally Lipschitz continuous gradient. *SIAM Journal on Optimization*, 33(3):2275–2310, 2023.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In André Platzer and Geoff Sutcliffe (eds.), *Automated Deduction – CADE 28*, pp. 625–635, Cham, 2021. Springer International Publishing. ISBN 978-3-030-79876-5.

Arkadi Nemirovski. Information-based complexity of convex programming. Lecture notes, Technion–Israel Institute of Technology, 1994. Fall semester 1994/95.

Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, August 2013. doi: 10.1007/s10107-012-0629-5.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, Department of Mathematics, University of Washington, May 2008.

Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60493–60514, 2025.

Alexander Tyurin. Near-optimal convergence of accelerated gradient methods under generalized and (L_0, L_1)-smoothness. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Daniil Vankov, Anton Rodomanov, Angelia Nedić, Lalitha Sankar, and Sebastian U. Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, 2025.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

A TECHNICAL PROOFS

A.1 PROOF OF PROPOSITION 2

Proof. Denote $d := \widetilde{x}^\star - x^\star$ and $x_t := (1-t)x^\star + t\widetilde{x}^\star$ for $t \in [0,1]$. Since $x^\star$ and $\widetilde{x}^\star$ are both optimal, convexity of F gives

$$F^\star \le F(x_t) \le (1-t)F(x^\star) + tF(\widetilde{x}^\star) = F^\star.$$

Hence $F(x_t) = F^\star$. Therefore,

$$\begin{aligned} 0 &= (1-t)F(x^\star) + tF(\widetilde{x}^\star) - F(x_t) \\ &= \big[(1-t)f(x^\star) + tf(\widetilde{x}^\star) - f(x_t)\big] + \big[(1-t)h(x^\star) + th(\widetilde{x}^\star) - h(x_t)\big]. \end{aligned}$$

The two terms in the brackets above are non-negative since both f and h are convex functions. As $f(x_t) = f(x^\star + td)$ is an affine function of $t \in (0,1)$, it holds that $d^\top \nabla^2 f(x^\star + td)d = 0$. By the convexity of f, we conclude that $\nabla^2 f(x_t)d = 0$, which implies

$$\nabla f(\widetilde{x}^\star) - \nabla f(x^\star) = \int_0^1 \nabla^2 f(x_t) d \, dt = 0. \qquad \square$$

A.2 PROOF OF LEMMA 3

Proof. By construction, $\nabla\phi(x^\star) = \nabla f(x^\star) + v^\star = 0$. Convexity of ϕ implies that $x^\star$ minimizes ϕ. Likewise, $0 \in \partial h(x^\star) - v^\star = \partial\psi(x^\star)$, so $x^\star$ also minimizes ψ. Since $F = \phi + \psi$, we obtain

$$0 \le \phi(x) - \phi(x^\star) \le \phi(x) - \phi(x^\star) + \psi(x) - \psi(x^\star) = F(x) - F^\star.$$

To verify the equivalence of the proximal update step, write the corresponding subproblem explicitly as

$$\begin{aligned} \operatorname{prox}_{a\psi}\big(w - a\nabla\phi(y)\big) &= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \left\{ \psi(u) + \frac{1}{2a}\left\|u - w + a\nabla\phi(y)\right\|^2 \right\} \\ &= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \left\{ h(u) + \frac{1}{2a}\left\|u - w + a\nabla f(y)\right\|^2 + C \right\} \\ &= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \left\{ h(u) + \frac{1}{2a}\left\|u - w + a\nabla f(y)\right\|^2 \right\} \\ &= \operatorname{prox}_{ah}\big(w - a\nabla f(y)\big), \end{aligned}$$

where

$$C = \langle v^\star, x^\star - w\rangle + a\,\langle v^\star, \nabla f(y)\rangle + \frac{a}{2}\left\|v^\star\right\|^2$$

is independent of u, which completes the proof. $\square$

A.3 PROOF OF THEOREM 4

Proof. Let $g := \nabla\phi$ and consider the gradient flow $\dot{z}(t) = -g(z(t))$ with $z(0) = x$. Since $\phi \in C^2$, its gradient is locally Lipschitz. Hence the vector field $V(z) := -\nabla\phi(z)$ is locally Lipschitz continuous, and the gradient-flow equation admits a unique local solution. Letting $s(t) := \|g(z(t))\|$, convexity of ϕ gives $\nabla^2\phi \succeq 0$, and hence

$$\frac{d}{dt}\frac{s(t)^2}{2} = -g(z(t))^\top \nabla^2\phi(z(t))g(z(t)) \le 0.$$

Thus $s(t) \le s(0)$ along the trajectory. In particular, $\|\dot{z}(t)\| = s(t) \le s(0)$, so the trajectory cannot escape to infinity in finite time and the solution exists for all $t \ge 0$.

Whenever $s(t) > 0$, we have

$$s'(t) = -\frac{g(z(t))^\top \nabla^2\phi(z(t))g(z(t))}{s(t)} \ge -\ell(\tau + s(t))s(t),$$

where the inequality follows from (2). At the same time,

$$\frac{d}{dt}\phi(z(t)) = -s(t)^2.$$

Therefore,

$$\frac{d}{dt}Q_\tau(s(t)) = \frac{s(t)s'(t)}{\ell(\tau + s(t))} \geq -s(t)^2 = \frac{d}{dt}\phi(z(t)).$$

Since $s(t)$ is nonincreasing, it has a limit. This limit must be zero. Otherwise, if $s(t) \to \bar{s} > 0$, then

$$\frac{d}{dt}\phi(z(t)) = -s(t)^2 \leq -\frac{\bar{s}^2}{2}$$

for all sufficiently large t, which would imply $\phi(z(t)) \to -\infty$, contradicting the fact that $x^\star$ minimizes ϕ.

Integrating the above differential inequality from 0 to T gives

$$\phi(x) - \phi(z(T)) \geq Q_\tau(s(0)) - Q_\tau(s(T)).$$

Letting $T \to \infty$, using $s(T) \to 0$ and $\phi(z(T)) \geq \phi(x^\star)$, yields

$$\phi(x) - \phi(x^\star) \geq Q_\tau\left(\|\nabla\phi(x)\|\right).$$

Upper bounding the left-hand side by $F(x) - F^\star$ proves the result. □

A.4 Proof of Lemma 5

Proof. Define $m(t) := \|\nabla\phi(c(t))\|$. On each continuously differentiable segment of c, whenever $m(t) > 0$, the chain rule and (2) give

$$m'(t) = \frac{\nabla\phi(c(t))^\top \nabla^2\phi(c(t))\dot{c}(t)}{\|\nabla\phi(c(t))\|} \leq \left\|\nabla^2\phi(c(t))\right\| \|\dot{c}(t)\| \leq \ell(\tau + m(t)) \|\dot{c}(t)\|.$$

We first prove $m(t) \leq 2s$ for all $t \in [0,1]$. Suppose this does not hold. Then there exists $t_1 \in [0,1]$ such that $m(t_1) > 2s$. Since $m(0) \leq s$ and m is continuous, the set $\{t \in [0,t_1] : m(t) = s\}$ is nonempty. Define

$$t_0 := \max\{t \in [0,t_1] : m(t) = s\}.$$

Then, for all $t \in (t_0, t_1]$, we have $m(t) > s$. Hence, on each smooth segment between t_0 and t_1, we have

$$\frac{m'(t)}{\ell(\tau + m(t))} \leq \|\dot{c}(t)\|.$$

By the chain rule and the fundamental theorem of calculus, integrating the left-hand side from t_0 to t_1 gives

$$\int_{t_0}^{t_1} \frac{m'(t)}{\ell(\tau + m(t))}\, dt = \int_s^{m(t_1)} \frac{du}{\ell(\tau + u)}.$$

For the right-hand side, we have

$$\int_{t_0}^{t_1} \|\dot{c}(t)\|\, dt \leq \int_0^1 \|\dot{c}(t)\|\, dt \leq \int_s^{2s} \frac{du}{\ell(\tau + u)}.$$

Combining these two equations gives

$$\int_s^{m(t_1)} \frac{du}{\ell(\tau + u)} \leq \int_s^{2s} \frac{du}{\ell(\tau + u)}.$$

Since $m(t_1) > 2s$, we have

$$\int_s^{m(t_1)} \frac{du}{\ell(\tau + u)} > \int_s^{2s} \frac{du}{\ell(\tau + u)},$$

which is a contradiction. Hence $m(t) \leq 2s$ for all $t \in [0,1]$.

Finally, by (2) and the monotonicity of ℓ, for all $t \in [0,1]$ we have

$$\left\|\nabla^2\phi(c(t))\right\| \leq \ell\left(\tau + \|\nabla\phi(c(t))\|\right) \leq \ell(\tau + 2s).$$

□

A.5 Proof of Lemma 6

Proof. The local quadratic upper model, the identity $x^+ - y = \alpha(z^+ - z)/A^+$, and convexity of f give the first inequality below. The second inequality follows from convexity of h applied to $x^+ = (Ax + \alpha z^+)/A^+$.

$$A^+ f(x^+) \le Af(x) + \alpha\left[f(y) + \langle \nabla f(y), z^+ - y\rangle\right] + \frac{L\alpha^2}{2A^+}\left\|z^+ - z\right\|^2,$$
$$A^+ h(x^+) \le Ah(x) + \alpha h(z^+).$$

Define $\widehat{F}_y(w) := f(y) + \langle \nabla f(y), w - y\rangle + h(w)$. Adding the two inequalities and using $L\alpha^2/A^+ \le 1$ yields

$$\begin{aligned} A^+ F(x^+) &\le AF(x) + \alpha \widehat{F}_y(z^+) + \frac{L\alpha^2}{2A^+}\left\|z^+ - z\right\|^2 \\ &\le AF(x) + \alpha \widehat{F}_y(z^+) + \frac{1}{2}\left\|z^+ - z\right\|^2. \end{aligned}$$

The optimality condition of the proximal subproblem yields the three-point inequality

$$\alpha \widehat{F}_y(z^+) + \frac{1}{2}\left\|z^+ - z\right\|^2 + \frac{1}{2}\left\|z^+ - x^\star\right\|^2 \le \alpha \widehat{F}_y(x^\star) + \frac{1}{2}\left\|z - x^\star\right\|^2.$$

Furthermore, convexity gives $\widehat{F}_y(x^\star) \le F^\star$. Combining these bounds and using $A^+ = A + \alpha$ proves the lemma. □

A.6 Proof of Lemma 7

Proof. Since $F(x_k) - F^\star \ge 0$, the potential bound $\mathcal{E}_k \le R^2$ implies $\|z_k - x^\star\| \le \sqrt{2}R$. To analyze the update with α_k, consider any auxiliary step size $\vartheta \in [0, \alpha_k]$ and perform the same update with ϑ in place of α_k. Denote the corresponding quantities by $y(\vartheta)$, $z(\vartheta)$, and $x(\vartheta)$, respectively. In particular, $z(0) = z_k$, while $\vartheta = \alpha_k$ recovers the actual update of Algorithm 1.

The map $\vartheta \mapsto z(\vartheta)$ is continuous on $[0, \alpha_k]$ by the strong convexity of the proximal subproblem and the continuity of its minimizer with respect to the step parameter. We first show that $z(\vartheta)$ cannot reach the boundary $\|z(\vartheta) - x^\star\| = 2R$. Assume that such a boundary is reached, and let $\hat{\vartheta} \in (0, \alpha_k]$ be the first value satisfying $\left\|z(\hat{\vartheta}) - x^\star\right\| = 2R$. Then, for every $\vartheta \in [0, \hat{\vartheta}]$,

$$\begin{aligned} \|y(\vartheta) - x_k\| + \|x(\vartheta) - y(\vartheta)\| &= \frac{\vartheta}{A_k + \vartheta}\big(\|z_k - x_k\| + \|z(\vartheta) - z_k\|\big) \\ &\le (2 + 3\sqrt{2})\frac{\vartheta R}{A_k + \vartheta} < \frac{8\alpha_k R}{A_{k+1}} \le \rho_k. \end{aligned}$$

Apply Lemma 5 with $\tau_\star$ to the path along $x_k \to y(\vartheta) \to x(\vartheta)$. Since $\|\nabla\phi(x_k)\| \le M_k$, the Hessian on the segment $y(\vartheta) \to x(\vartheta)$ is bounded by $L_k = \ell(\tau_\star + 2M_k)$. Let $d(\vartheta) := x(\vartheta) - y(\vartheta)$. Since the Hessian is bounded by L_k along the segment from $y(\vartheta)$ to $x(\vartheta)$, we have

$$\begin{aligned} f(x(\vartheta)) &= f(y(\vartheta)) + \langle \nabla f(y(\vartheta)), d(\vartheta)\rangle + \int_0^1 (1-t) d(\vartheta)^\top \nabla^2 f\big(y(\vartheta) + td(\vartheta)\big) d(\vartheta)\, dt \\ &\le f(y(\vartheta)) + \langle \nabla f(y(\vartheta)), d(\vartheta)\rangle + \frac{L_k}{2}\|d(\vartheta)\|^2, \end{aligned}$$

which verifies the local quadratic upper model in Lemma 6 with $L = L_k$. Furthermore, we can show

$$\frac{L_k \vartheta^2}{A_k + \vartheta} \le \frac{L_k \alpha_k^2}{A_{k+1}} \le 1.$$

The lemma therefore gives

$$(A_k + \vartheta)\big(F(x(\vartheta)) - F^\star\big) + \frac{1}{2}\|z(\vartheta) - x^\star\|^2 \le \mathcal{E}_k \le R^2.$$

At the assumed first exit, this implies $\|z(\vartheta) - x^\star\| \le \sqrt{2}R$, contradicting $\|z(\vartheta) - x^\star\| = 2R$. Hence no such first exit occurs for $\vartheta \in [0, \alpha_k]$. Taking $\vartheta = \alpha_k$ recovers the actual update and gives $\mathcal{E}_{k+1} \le \mathcal{E}_k \le R^2$. Consequently, $F(x_{k+1}) - F^\star \le R^2/A_{k+1}$ and $\|z_{k+1} - x^\star\| \le \sqrt{2}R$. Since x_{k+1} is a convex combination of x_k and z_{k+1}, we also have $\|x_{k+1} - x^\star\| \le \sqrt{2}R$.

Finally, the distance between x_k and y_k is bounded as

$$\|y_k - x_k\| = \frac{\alpha_k}{A_{k+1}} \|z_k - x_k\| \le \frac{2\sqrt{2}\alpha_k R}{A_{k+1}} < \rho_k.$$

Applying Lemma 5 once more gives $\|\nabla\phi(y_k)\| \le 2M_k$, which also implies $\|g_k - \nabla f(x^\star)\| \le 2M_k$. □

A.7 Proof of Theorem 9

Proof. We first establish the global bounds needed for the rate analysis. By (1) and (3), the initialization gives

$$\mathcal{E}_0 \le A_0\Delta_0 + \frac{R^2}{2} \le R^2.$$

Suppose that $\mathcal{E}_k \le R^2$. Since

$$F(x_k) - F^\star \le \frac{R^2}{A_k} =: \delta_k,$$

Theorem 4 gives

$$\|\nabla\phi(x_k)\| \le Q_{\tau_\star}^{-1}(\delta_k) =: M_k.$$

As $A_{k+1} = A_k + \alpha_k > A_k$, we have $\delta_{k+1} < \delta_k$, so δ_k remains in the range of $Q_{\tau_\star}$. Induction over k using Lemma 7 gives $\mathcal{E}_k \le R^2$ and $F(x_k) - F^\star \le \frac{R^2}{A_k}$. We next show that the iterates enter the solution-scale region after finitely many iterations. Denote $\delta_0 := \Gamma_0 R^2$ and $M_0 := Q_{\tau_\star}^{-1}(\delta_0)$. By continuity of ℓ at $\tau_\star$, choose $\bar{s} > 0$ such that

$$\ell(\tau_\star + 2\bar{s}) \le 2L_\star,$$

and set $\bar{\delta} := Q_{\tau_\star}(\bar{s})$. If $M_0 > \bar{s}$, define $\bar{L} := \ell(\tau_\star + 2M_0)$, $\bar{\rho} := \min_{s\in[\bar{s},M_0]} \rho_{\tau_\star}(s)$. By the choice of α_k in Algorithm 1 and the monotonicity of $s \mapsto s/(A_k + s)$, it holds

$$\frac{\alpha_k}{A_{k+1}} = \min\left\{\frac{1}{2}, \frac{2}{1 + \sqrt{1 + 4A_kL_k}}, \frac{\rho_k}{8R}\right\}. \tag{4}$$

Before M_k reaches $\bar{s}$, all three terms on the right-hand side of (4) are bounded below by $\underline{c}$, where

$$\underline{c} := \min\left\{\frac{1}{2}, \frac{2}{1 + \sqrt{1 + 4R^2\bar{L}/\bar{\delta}}}, \frac{\bar{\rho}}{8R}\right\}.$$

Since

$$\delta_{k+1} = \frac{A_k}{A_{k+1}}\delta_k = \left(1 - \frac{\alpha_k}{A_{k+1}}\right)\delta_k \le (1 - \underline{c})\delta_k,$$

there exists a finite integer

$$T_0 := \left\lceil \frac{\log(\delta_0/\bar{\delta})}{-\log(1 - \underline{c})} \right\rceil$$

such that $M_k \le \bar{s}$ for all $k \ge T_0$. If $M_0 \le \bar{s}$, set $T_0 = 0$. For $k \ge T_0$, we have

$$L_k = \ell(\tau_\star + 2M_k) \le 2L_\star.$$

The monotonicity of ℓ also gives

$$\rho_k = \rho_{\tau_\star}(M_k) \ge \frac{M_k}{2L_\star}.$$

With the fact

$$\delta_k = Q_{\tau_\star}(M_k) \le \frac{M_k^2}{2L_\star},$$

we also have

$$M_k \geq R\sqrt{\frac{2L_\star}{A_k}}.$$

The curvature and path-length constraints in Algorithm 1 therefore imply

$$\frac{\alpha_k}{A_{k+1}} \geq \min\left\{\frac{1}{2}, \frac{1}{8\sqrt{2L_\star A_k}}\right\}.$$

While $L_\star A_k < 1/32$, the lower bound above gives $\frac{\alpha_k}{A_{k+1}} \geq \frac{1}{2}$. Since the algorithm always enforces $\alpha_k/A_{k+1} \leq 1/2$, we must have $\frac{\alpha_k}{A_{k+1}} = \frac{1}{2}$, and hence $A_{k+1} = 2A_k$. Therefore, A_k doubles at every iteration until $L_\star A_k \geq 1/32$. This requires at most

$$T_1 := \left\lceil \left[\log_2\left(\frac{1}{32L_\star A_{T_0}}\right)\right]_+ \right\rceil$$

additional iterations. Set $T_\ell := T_0 + T_1$, which is finite and independent of the target accuracy. For every $k \geq T_\ell$, we can derive

$$\begin{aligned}\sqrt{A_{k+1}} - \sqrt{A_k} &= \frac{\alpha_k}{\sqrt{A_{k+1}} + \sqrt{A_k}} \\ &\geq \frac{\sqrt{A_k}\alpha_k}{2A_{k+1}} \geq \frac{1}{16\sqrt{2L_\star}}.\end{aligned}$$

Hence, after m further iterations, it must hold

$$A_{T_\ell+m} \geq \frac{m^2}{512L_\star}.$$

Taking $m = N - T_\ell$ and using the global bound $F(x_N) - F^\star \leq R^2/A_N$ yields the result. □

A.8 Proof of Corollary 10

Proof. The statement for $L_1 = 0$ follows directly from Theorem 9. It remains to consider $L_1 > 0$. Under (L_0, L_1)-smoothness, the corresponding quantities can be written explicitly as

$$\begin{aligned}Q_{\tau_\star}(s) &= \frac{s}{L_1} - \frac{L}{L_1^2}\log\left(1 + \frac{L_1 s}{L}\right), \\ \rho_{\tau_\star}(s) &= \frac{1}{L_1}\log\left(\frac{L + 2L_1 s}{L + L_1 s}\right).\end{aligned}$$

Write $\delta = \delta_k$ and $M = M_k$. The bound

$$\frac{M^2}{2(L + L_1 M)} \leq Q_{\tau_\star}(M) = \delta \leq \frac{M^2}{2L}$$

gives

$$\sqrt{2L\delta} \leq M \leq L_1\delta + \sqrt{L_1^2\delta^2 + 2L\delta}.$$

We first bound the number of iterations needed to reach $\delta_k \leq L/L_1^2$. Whenever $\delta \geq L/L_1^2$, the previous bounds give

$$\frac{\sqrt{2}L}{L_1} \leq M \leq (1 + \sqrt{3})L_1\delta.$$

Hence $L_k = O(L_1^2\delta)$ and $A_k L_k = O((L_1 R)^2)$. The curvature restriction in (4) is therefore bounded below by $c/(1 + L_1 R)$, where $c > 0$ is a constant. In addition, the explicit formula for $\rho_{\tau_\star}$ and $L_1 M/L \geq \sqrt{2}$ give $\rho_k \geq \frac{c}{L_1}$. (4) yields

$$\frac{\alpha_k}{A_{k+1}} \geq \frac{c}{1 + L_1 R}.$$

Since $\delta_{k+1} = (A_k/A_{k+1})\delta_k$, the gap bound decreases geometrically and reaches L/L_1^2 within $O\left((1+L_1R)\log\left(2+\frac{\Gamma_0 L_1^2 R^2}{L}\right)\right)$ iterations.

Once $\delta \le L/L_1^2$, the upper bound on M gives $M = O(L/L_1)$ and hence $L_k = O(L)$. The path-length bound satisfies

$$\rho_k \ge \frac{M}{L + 2L_1 M} \ge \frac{cM}{L}.$$

Together with

$$M \ge \sqrt{2L\delta} = R\sqrt{\frac{2L}{A_k}},$$

the curvature and path-length restrictions in (4) give

$$\frac{\alpha_k}{A_{k+1}} \ge c \min\left\{1, \frac{1}{\sqrt{LA_k}}\right\}.$$

As long as $LA_k \le 1$, the above bound gives $A_{k+1} \ge (1+c)A_k$. Therefore, A_k reaches $1/L$ within $O(\log(2+\Gamma_0/L))$ iterations. Once $LA_k \ge 1$, the bound implies

$$\sqrt{A_{k+1}} - \sqrt{A_k} \ge \frac{c}{\sqrt{L}}.$$

Consequently, after m additional iterations, $A_k \ge cm^2/L$. Therefore, $F(x_k) - F^\star \le \varepsilon$ once

$$m = O\left(R\sqrt{\frac{L}{\varepsilon}}\right).$$

Adding the phase bounds gives the stated complexity, since both logarithmic factors are at most $\log(2+\Gamma_0(1+L_1R)^2/L)$ and $1 \le 1 + L_1R$. □

A.9 Proof of Theorem 11

Proof. Consider stage j, with $\Delta := \Delta_j$ and $R_\Delta = R_j$. As Algorithm 2 starts each stage with $\Gamma_0 = \mu$, the initialization rule $A_0 = 1/\Gamma_0$ in Algorithm 1 then determines $A_0 = 1/\mu$, $\Gamma_0 = \frac{2\Delta}{R_\Delta^2}$, and

$$\Gamma_0 R_\Delta^2 = 2\Delta \le 2\Delta_0 < Q_{\tau_\star,\infty}.$$

Under this setting, the parameter condition (3) holds for every stage. Repeated application of Lemma 7 gives

$$F(x_k) - F^\star \le \frac{R_\Delta^2}{A_k} =: \delta_k$$

throughout the stage. Before the stopping iteration, we know $\frac{1}{\mu} \le A_k < \frac{4}{\mu}$, and therefore $\frac{\Delta}{2} < \delta_k \le 2\Delta$. We use the geometric-decrease argument in the proof of Theorem 9 to bound the total number of iterations in each stage. Since $M_k = Q_{\tau_\star}^{-1}(\delta_k)$, we have

$$Q_{\tau_\star}^{-1}\left(\frac{\Delta}{2}\right) \le M_k \le Q_{\tau_\star}^{-1}(2\Delta).$$

By monotonicity of ℓ and the previous bounds on M_k,

$$L_k = \ell(\tau_\star + 2M_k) \le \ell\big(\tau_\star + 2Q_{\tau_\star}^{-1}(2\Delta)\big),$$
$$\rho_k = \rho_{\tau_\star}(M_k) \ge \min_{s \in [Q_{\tau_\star}^{-1}(\Delta/2), Q_{\tau_\star}^{-1}(2\Delta)]} \rho_{\tau_\star}(s) =: \underline{\rho}(\Delta) > 0.$$

The minimum is positive because $\rho_{\tau_\star}$ is continuous and positive on this compact interval. Collect these bounds by setting

$$\Lambda_\ell(\Delta) := \max\left\{\ell\big(\tau_\star + 2Q_{\tau_\star}^{-1}(2\Delta)\big), \frac{\Delta}{\underline{\rho}(\Delta)^2}\right\}.$$

Then $L_k \le \Lambda_\ell(\Delta)$ and $\rho_k \ge \underline{\rho}(\Delta) \ge \sqrt{\Delta/\Lambda_\ell(\Delta)}$. Substituting these inequalities, $A_k < 4/\mu$, and $R_\Delta^2 = 2\Delta/\mu$ into (4) gives

$$\begin{aligned}\frac{\alpha_k}{A_{k+1}} &\ge \min\left\{\frac{1}{2}, \frac{2}{1+\sqrt{1+16\Lambda_\ell(\Delta)/\mu}}, \frac{1}{8\sqrt{2}}\sqrt{\frac{\mu}{\Lambda_\ell(\Delta)}}\right\} \\ &\ge c\min\left\{1, \sqrt{\frac{\mu}{\Lambda_\ell(\Delta)}}\right\},\end{aligned}$$

where $c > 0$ is a constant. As in the proof of Theorem 9, this yields the geometric decrease

$$\delta_{k+1} = \frac{A_k}{A_{k+1}}\delta_k \le \left(1 - c\min\left\{1, \sqrt{\frac{\mu}{\Lambda_\ell(\Delta)}}\right\}\right)\delta_k.$$

Thus δ_k decreases from 2Δ to at most $\Delta/2$, equivalently A_k reaches $4/\mu$, within

$$\left\lceil \frac{\log 4}{-\log\left(1 - c\min\left\{1, \sqrt{\mu/\Lambda_\ell(\Delta)}\right\}\right)} \right\rceil = O\left(1 + \sqrt{\frac{\Lambda_\ell(\Delta)}{\mu}}\right)$$

iterations.

It remains to sum the iteration bounds over all restart stages. Local Lipschitz continuity of ℓ at $\tau_\star$ means that there exist constants $C, r > 0$ such that

$$|\ell(\tau_\star + u) - L_\star| \le Cu, \qquad 0 \le u \le r.$$

Since ℓ is nondecreasing, $\ell(\tau_\star + u) \ge L_\star$. Hence,

$$\left|\frac{1}{\ell(\tau_\star + u)} - \frac{1}{L_\star}\right| = \frac{|\ell(\tau_\star + u) - L_\star|}{L_\star \ell(\tau_\star + u)} \le \frac{Cu}{L_\star^2}.$$

Thus $\ell(\tau_\star + u) = L_\star + O(u)$ and $1/\ell(\tau_\star + u) = 1/L_\star + O(u)$ as $u \downarrow 0$. For $0 < s \le r/2$, substituting the reciprocal estimate into the definitions of $Q_{\tau_\star}$ and $\rho_{\tau_\star}$ gives

$$\begin{aligned}Q_{\tau_\star}(s) &= \int_0^s \left(\frac{u}{L_\star} + O(u^2)\right) du = \frac{s^2}{2L_\star} + O(s^3), \\ \rho_{\tau_\star}(s) &= \int_s^{2s} \left(\frac{1}{L_\star} + O(u)\right) du = \frac{s}{L_\star} + O(s^2)\end{aligned}$$

as $s \downarrow 0$. Inverting the expansion for $Q_{\tau_\star}$ gives $Q_{\tau_\star}^{-1}(\Delta/2) = \sqrt{L_\star \Delta} + O(\Delta)$ and $Q_{\tau_\star}^{-1}(2\Delta) = 2\sqrt{L_\star\Delta} + O(\Delta)$ as $\Delta \downarrow 0$, from which we can show

$$\ell\big(\tau_\star + 2Q_{\tau_\star}^{-1}(2\Delta)\big) = L_\star + O(\sqrt{\Delta}).$$

Notice that

$$\rho'_{\tau_\star}(s) = \frac{2}{\ell(\tau_\star + 2s)} - \frac{1}{\ell(\tau_\star + s)} \xrightarrow{s\to 0} \frac{1}{L_\star} > 0.$$

Thus $\rho_{\tau_\star}$ is increasing on a sufficiently small neighborhood of zero. For all sufficiently small Δ, $\underline{\rho}(\Delta) = \rho_{\tau_\star}(Q_{\tau_\star}^{-1}(\Delta/2))$. Since

$$Q_{\tau_\star}\big(Q_{\tau_\star}^{-1}(\Delta/2)\big) = \frac{\Delta}{2},$$

the previous expansions give

$$\big(Q_{\tau_\star}^{-1}(\Delta/2)\big)^2 = L_\star\Delta + O(\Delta^{3/2}), \qquad \frac{\Delta}{\underline{\rho}(\Delta)^2} = L_\star + O(\sqrt{\Delta})$$

Therefore, $\Lambda_\ell(\Delta) = L_\star + O(\sqrt{\Delta})$. Consequently, as $\Delta \downarrow 0$, it holds

$$\sqrt{\Lambda_\ell(\Delta)} - \sqrt{L_\star} = O(\sqrt{\Delta}).$$

Define

$$T_{\ell,\mu} := \frac{1}{\sqrt{\mu}} \sum_{j=0}^{\infty} \left(\sqrt{\Lambda_\ell(\Delta_j)} - \sqrt{L_\star}\right).$$

Since $\Delta_j = 2^{-j}\Delta_0$, $T_{\ell,\mu}$ is finite and independent of ε.

Let $J := \left\lceil \log_2 \frac{\Delta_0}{\varepsilon} \right\rceil$. After J stages, $F(\bar{x}_J) - F^\star \le 2^{-J}\Delta_0 \le \varepsilon$. Summing the one-stage bound gives

$$N_\varepsilon = O\left(\sum_{j=0}^{J-1}\left(1+\sqrt{\frac{\Lambda_\ell(\Delta_j)}{\mu}}\right)\right) = O\left(J\left(1+\sqrt{\frac{L_\star}{\mu}}\right)+T_{\ell,\mu}\right),$$

which proves the result. □

A.10 Proof of Corollary 12

Proof. When $L_1 = 0$, the definitions give $\Lambda_\ell(\Delta) = L_0$ and $T_{\ell,\mu} = 0$, so the bound follows from Theorem 11. For $L_1 > 0$, fix $\Delta > 0$ and set $s_- := Q_{\tau_\star}^{-1}(\Delta/2)$, $s_+ := Q_{\tau_\star}^{-1}(2\Delta)$. The inverse bounds in the proof of Corollary 10 give $s_- \ge \sqrt{L\Delta}$ and $s_+ \le 4L_1\Delta + 2\sqrt{L\Delta}$. The explicit formula for $\rho_{\tau_\star}$ gives $\rho'_{\tau_\star}(s) = L/((L+L_1 s)(L+2L_1 s)) > 0$. Thus $\underline{\rho}(\Delta) = \rho_{\tau_\star}(s_-) \ge s_-/(L+2L_1 s_-)$, which yields

$$\frac{\sqrt{\Delta}}{\underline{\rho}(\Delta)} \le \frac{L\sqrt{\Delta}}{s_-} + 2L_1\sqrt{\Delta} \le \sqrt{L} + 2L_1\sqrt{\Delta},$$

$$\ell(\tau_\star + 2s_+) \le L + 4L_1\sqrt{L\Delta} + 8L_1^2\Delta \le (\sqrt{L} + 3L_1\sqrt{\Delta})^2.$$

By the definition of Λ_ℓ, these estimates imply $0 \le \sqrt{\Lambda_\ell(\Delta)} - \sqrt{L} \le 3L_1\sqrt{\Delta}$. Since $\Delta_j = 2^{-j}\Delta_0$, the transient constant in Theorem 11 satisfies

$$T_{\ell,\mu} \le \frac{3L_1}{\sqrt{\mu}}\sum_{j=0}^{\infty}\sqrt{\Delta_j} = \frac{3L_1\sqrt{\Delta_0}}{(1-2^{-1/2})\sqrt{\mu}} = O(L_1 R_0).$$

Substituting this estimate into Theorem 11 proves the result. □

A.11 Proof of Proposition 13

Proof. The claim is immediate when $r = 0$. Suppose $r > 0$. Define $e := -(\nabla f(x) - \nabla f(y))/r$. Along the line $x + te$, define $r_t := \|\nabla f(x+te) - \nabla f(x)\|$. Since $r_0 = 0$, the curvature bound and the triangle inequality give

$$r'_t \le \left\|\nabla^2 f(x+te)e\right\| \le \ell(c + r_t).$$

Taking the derivative of P_c, we can show

$$P'_c(r_t) = \int_0^{r_t} \frac{du}{\ell(c+u)} \le t.$$

Since P'_c is strictly increasing, its inverse is well-defined. For any value of r, denote

$$t_r := P'_c(r) = \int_0^{r} \frac{du}{\ell(c+u)}$$

and $w := x + t_r e$. Then $r_t \le (P'_c)^{-1}(t)$ for $t \in [0, t_r]$. Integration along the segment from x to w gives

$$\begin{aligned} f(w) &= f(x) + t_r\langle\nabla f(x), e\rangle + \int_0^{t_r}\langle\nabla f(x+te) - \nabla f(x), e\rangle\, dt \\ &\le f(x) + t_r\langle\nabla f(x), e\rangle + \int_0^{t_r}(P'_c)^{-1}(t)\, dt \\ &= f(x) + t_r\langle\nabla f(x), e\rangle + \int_0^{r}\frac{u}{\ell(c+u)}\, du. \end{aligned}$$

The convex function $u \mapsto f(u) - \langle\nabla f(y), u\rangle$ is minimized at y. Hence

$$f(y) - \langle\nabla f(y), y\rangle \le f(w) - \langle\nabla f(y), w\rangle.$$

Using the derived upper bound and $\langle \nabla f(x) - \nabla f(y), e \rangle = -r$, we obtain

$$\begin{aligned} f(y) - \langle \nabla f(y), y \rangle &\le f(x) - \langle \nabla f(y), x \rangle - r t_r + \int_0^r \frac{u}{\ell(c+u)}\, du \\ &= f(x) - \langle \nabla f(y), x \rangle - P_c(r), \end{aligned}$$

which proves the first claim.

The expression for P_c' above shows that P_c is continuous and strictly increasing. Furthermore, for any fixed $d > 0$, $P_c'(s) \ge \int_0^d \ell(c+u)^{-1}\, du > 0$ for $s \ge d$, so $P_c(s) \to \infty$ as $s \to \infty$. Hence P_c^{-1} is well defined on $[0, \infty)$.

Finally, apply the first claim with $y = x^\star$. Composite optimality gives $-\nabla f(x^\star) \in \partial h(x^\star)$. Therefore,

$$\begin{aligned} F(x) - F^\star &\ge f(x) - f(x^\star) - \langle \nabla f(x^\star), x - x^\star \rangle \\ &\ge P_{\|\nabla f(x)\|}\left(\|\nabla f(x) - \nabla f(x^\star)\|\right). \end{aligned}$$

Hence $F(x) - F^\star \le \delta$ implies

$$\|\nabla f(x) - \nabla f(x^\star)\| \le P^{-1}_{\|\nabla f(x)\|}(\delta).$$

The triangle inequality then yields

$$\tau_\star = \|\nabla f(x^\star)\| \le \|\nabla f(x)\| + P^{-1}_{\|\nabla f(x)\|}(\delta),$$

which proves the second claim. □

A.12 Proof of Theorem 15

Proof. We first show that the algorithm is well defined and that the warm-start terminates after finitely many iterations. By Corollary 14, $\widehat{\tau}_0 \ge \tau_\star$, while (1) and the choice of Γ_0 give

$$\mathcal{E}_0 \le A_0 \Delta_0 + \frac{R^2}{2} \le R^2.$$

Suppose the k-th iteration is in the warm-start phase and that $\mathcal{E}_k \le R^2$. Then $F(x_k) - F^\star \le \delta_k := R^2/A_k$. Set $c_k := \|\nabla f(x_k)\|$ and $M_k := P_{c_k}^{-1}(\delta_k)$. By Proposition 13,

$$\|\nabla f(x_k) - \nabla f(x^\star)\| \le M_k, \qquad c_k + M_k \ge \tau_\star.$$

Hence the update of $\widehat{\tau}_k$ preserves $\widehat{\tau}_k \ge \tau_\star$. Applying Lemma 7 with M_k and $\widehat{\tau}_k$ therefore yields $\mathcal{E}_{k+1} \le \mathcal{E}_k \le R^2$. In particular, the potential and distance bounds are preserved throughout the warm-start phase.

It remains to show that this phase cannot proceed indefinitely. Let $\delta^\circ := Q_{\widehat{\tau}_0}(M_0)/2$, and define $G_R := \max_{\|x - x^\star\| \le \sqrt{2}R} \|\nabla f(x)\|$, $M_- := P_0^{-1}(\delta^\circ)$, $M_+ := P_{G_R}^{-1}(\delta_0)$. Before termination, $\delta^\circ < \delta_k \le \delta_0$. Since $\|x_k - x^\star\| \le \sqrt{2}R$, we have $c_k \le G_R$. The monotonicity of P_c^{-1} in both arguments then gives $M_- \le M_k \le M_+$. Combining this with $\widehat{\tau}_k \le \widehat{\tau}_0$ gives

$$L_k = \ell(\widehat{\tau}_k + 2M_k) \le \ell(\widehat{\tau}_0 + 2M_+) =: \widehat{L}.$$

The path-length bound satisfies

$$\rho_k = \rho_{\widehat{\tau}_k}(M_k) \ge \frac{M_k}{L_k} \ge \frac{M_-}{\widehat{L}},$$

while $A_k = R^2/\delta_k \le R^2/\delta^\circ$. Using (4), there is a constant

$$\widehat{c}_0 := \min\left\{ \frac{1}{2},\ \frac{2}{1 + \sqrt{1 + 4R^2\widehat{L}/\delta^\circ}},\ \frac{M_-}{8R\widehat{L}} \right\} > 0$$

such that $\alpha_k / A_{k+1} \ge \widehat{c}_0$ throughout the warm-start. Consequently,

$$\delta_{k+1} = \frac{A_k}{A_{k+1}} \delta_k \le (1 - \widehat{c}_0)\delta_k,$$

and the condition $\delta_k \le \delta^\circ$ is reached after finitely many iterations. Denote the first such index by K. At the exit point, $\widehat{\tau}_K \le \widehat{\tau}_0$. Since $Q_\tau(s)$ is nonincreasing in τ,

$$Q_{\widehat{\tau}_K}(M_0) \ge Q_{\widehat{\tau}_0}(M_0) = 2\delta^\circ \ge 2\delta_K.$$

Thus $Q^{-1}_{\widehat{\tau}_K}(\delta_K)$ is well defined. If the initial gap scale already satisfies $\delta_0 < Q_{\widehat{\tau}_0,\infty}$, we simply take $K = 0$.

We next consider the iterations after K. At every $k \ge K$, $M_k := Q^{-1}_{\widehat{\tau}_k}(\delta_k)$. Since $\widehat{\tau}_k \ge \tau_\star$, Theorem 4 gives

$$\|\nabla f(x_k) - \nabla f(x^\star)\| \le M_k.$$

Lemma 7 again yields $\mathcal{E}_{k+1} \le \mathcal{E}_k$, together with

$$\|\nabla f(y_k) - \nabla f(x^\star)\| \le 2M_k.$$

It follows that

$$\tau_\star \le \|\nabla f(y_k)\| + 2M_k,$$

so the update of $\widehat{\tau}_{k+1}$ remains a valid upper bound on $\tau_\star$. The same gradient bound also yields

$$\widehat{\tau}_{k+1} \le \|\nabla f(y_k)\| + 2M_k \le \tau_\star + 4M_k.$$

Since $\widehat{\tau}_{k+1} \le \widehat{\tau}_k$ and $\delta_{k+1} < \delta_k$, we have

$$Q_{\widehat{\tau}_{k+1},\infty} \ge Q_{\widehat{\tau}_k,\infty} > \delta_k > \delta_{k+1},$$

so the localization remains well defined at every subsequent iteration. Furthermore,

$$Q_{\widehat{\tau}_{k+1}}(M_k) \ge Q_{\widehat{\tau}_k}(M_k) = \delta_k > \delta_{k+1},$$

which implies $M_{k+1} < M_k$.

We now show that the local curvature eventually reaches the solution scale. By continuity of ℓ, choose $s_\star > 0$ such that

$$\ell(\tau_\star + 6s_\star) \le 2L_\star.$$

If $M_K \le s_\star$, no further entrance phase is needed. Otherwise, set $\bar{\tau} := \widehat{\tau}_K$, $\bar{M} := M_K$, $\bar{L} := \ell(\bar{\tau} + 2\bar{M})$, and $\bar{\delta} := Q_{\bar{\tau}}(s_\star)$. As long as $M_k \ge s_\star$, monotonicity gives $M_k \le \bar{M}$, $\widehat{\tau}_k \le \bar{\tau}$. Hence, $L_k \le \bar{L}$. The gap scale also satisfies

$$\delta_k = Q_{\widehat{\tau}_k}(M_k) \ge Q_{\bar{\tau}}(s_\star) = \bar{\delta},$$

so $A_k \le R^2/\bar{\delta}$, while $\rho_k \ge s_\star/\bar{L}$. Equation (4) therefore gives a uniform constant $\widehat{c}_1 > 0$ such that $\alpha_k/A_{k+1} \ge \widehat{c}_1$ throughout this phase. Hence δ_k decreases geometrically, so there exists a finite index j such that $M_j \le s_\star$.

At iteration j,

$$\widehat{\tau}_{j+1} \le \tau_\star + 4M_j \le \tau_\star + 4s_\star.$$

Since both $\widehat{\tau}_k$ and M_k are nonincreasing, for every $k \ge j+1$,

$$L_k = \ell(\widehat{\tau}_k + 2M_k) \le \ell(\tau_\star + 6s_\star) \le 2L_\star.$$

It remains to derive the accelerated rate in this region. Since $L_k \le 2L_\star$,

$$\rho_k \ge \frac{M_k}{2L_\star}.$$

Also, $\widehat{\tau}_k \ge \tau_\star$ implies

$$\delta_k = Q_{\widehat{\tau}_k}(M_k) \le \frac{M_k^2}{2L_\star},$$

and therefore

$$M_k \ge R\sqrt{\frac{2L_\star}{A_k}}.$$

Combining these bounds with (4) gives

$$\frac{\alpha_k}{A_{k+1}} \ge \min\left\{\frac{1}{2},\ \frac{1}{8\sqrt{2L_\star A_k}}\right\}.$$

While $L_\star A_k < 1/32$, this lower bound equals $1/2$, and hence $A_{k+1} = 2A_k$. Thus, after finitely many additional iterations, $L_\star A_k \geq 1/32$. For all subsequent iterations,

$$\begin{aligned}\sqrt{A_{k+1}} - \sqrt{A_k} &= \frac{\alpha_k}{\sqrt{A_{k+1}} + \sqrt{A_k}} \\ &\geq \frac{\sqrt{A_k}\alpha_k}{2A_{k+1}} \geq \frac{1}{16\sqrt{2L_\star}}.\end{aligned}$$

The warm-start and the two entrance phases above require finitely many iterations. Thus, there exists a finite constant $\widehat{T}_\ell$, independent of ε, such that

$$A_N \geq \frac{(N - \widehat{T}_\ell)^2}{512 L_\star}, \qquad N > \widehat{T}_\ell.$$

Since the potential bound gives $F(x_N) - F^\star \leq R^2/A_N$, we conclude that

$$F(x_N) - F^\star \leq \frac{512\,\ell(\tau_\star) R^2}{(N - \widehat{T}_\ell)^2}.$$

The stated iteration complexity follows immediately. □

A.13 Proof of Corollary 16

Proof. For $0 \leq u \leq s$, we have $\ell(c_0 + u) \leq L_0 + L_1 c_0 + L_1 s$. Therefore,

$$P_{c_0}(s) \geq \frac{s^2}{2(L_0 + L_1 c_0 + L_1 s)}.$$

Apply this bound at $s = \|\nabla f(x_0) - \nabla f(x^\star)\|$ and use the gap bound (1) together with $\Delta_0 \leq \delta_0$. Solving the resulting quadratic inequality gives the stated bound on the solution-centered gradient norm. The complexity then follows from Theorem 15. □

A.14 Proof of Theorem 17

Proof. Denote $L_\tau := \ell(\tau)$. Fix any $A \in \mathcal{A}_p$ and an integer oracle budget $N \geq 1$. Choose the dimension to satisfy $m \geq 4N + 5$. Let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex quadratic function with $0 \preceq \nabla^2 g \preceq L_\tau I$. Initialize the lifted method at $x^{(0)} = (z^{(0)}, 0)$. For $x = (z, r) \in \mathbb{R}^{m+1}$, define the lifted functions

$$f_g(z, r) := g(z) - \tau r, \qquad h_\sigma(z, r) := \tau|r| + \frac{\sigma}{2} r^2, \qquad \sigma \geq 0. \tag{5}$$

Since g is quadratic, $f_g \in C^\infty$, so the standing C^2 assumption holds. The function h_σ is proper, closed, convex, and possibly nonsmooth. The corresponding composite objective is

$$F_\sigma(z, r) = g(z) + \tau(|r| - r) + \frac{\sigma}{2} r^2.$$

Since the last two terms are nonnegative and vanish at $r = 0$, we have $F_\sigma^\star = g^\star$, and

$$F_\sigma(z, r) - F_\sigma^\star \geq g(z) - g^\star. \tag{6}$$

The smooth function f_g satisfies

$$\nabla f_g(z, r) = (\nabla g(z), -\tau), \qquad \nabla^2 f_g(z, r) = \begin{pmatrix} \nabla^2 g(z) & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, the gradient can be bounded as $\|\nabla f_g(z, r)\| \geq \tau$ and the Hessian satisfies Assumption 1 since

$$\big\|\nabla^2 f_g(z, r)\big\| \leq L_\tau = \ell(\tau) \leq \ell\big(\|\nabla f_g(z, r)\|\big).$$

We next compare the oracle information available for the lifted problem with that of g. Suppose that A queries the first-order oracle of f_g at (z, r). A single first-order oracle call to g at z gives

$(g(z), \nabla g(z))$, from which $f_g(z,r)$ and $\nabla f_g(z,r)$ can be obtained exactly. The proximal oracle follows

$$\operatorname{prox}_{ah_\sigma}(w_z, w_r) = \left(w_z, \frac{\operatorname{sign}(w_r)(|w_r| - a\tau)_+}{1 + a\sigma} \right), \tag{7}$$

which shows that it can be evaluated explicitly without any access to g. By the preceding discussion, each first-order oracle call made by A can be simulated using one oracle call to g, while every proximal-oracle call can be simulated without querying g.

Since A is deterministic, the above simulation defines a deterministic first-order method $\widetilde{A}$ for g. If A outputs $x^{(t)} = (z^{(t)}, r^{(t)})$ after t first-order oracle calls, the corresponding output of $\widetilde{A}$ is $z^{(t)}$. Then, by (6), we have

$$g(z^{(t)}) - g^\star \le F_\sigma(x^{(t)}) - F_\sigma^\star.$$

Hence, an ε-optimal output of A gives an ε-optimal output of $\widetilde{A}$. To apply a final-output lower bound to the first-hitting complexity defined above, simulate N calls of A, evaluate g at its $N+1$ projected outputs $z^{(0)}, \dots, z^{(N)}$, then return the one with the smallest value. This deterministic method uses at most $2N+1$ oracle calls to g; its objective gap equals $\min_{0 \le t \le N}(g(z^{(t)}) - g^\star)$. The choice $m \ge 4N+5$ satisfies the dimension requirements of the quadratic lower bounds used below with this budget.

We first consider the convex case with $\sigma = 0$. The convex quadratic minimax bound recalled in Drori (2017, Eq. (1.1)) applies in dimension $m \ge 2B+3$ to any deterministic first-order method using at most B calls. Taking $B = 2N+1$, there exists a convex quadratic g with $\mathbf{dist}(z^{(0)}, \arg\min g) \le R$ such that

$$\min_{0 \le t \le N} \big(g(z^{(t)}) - g^\star \big) \ge \frac{L_\tau R^2}{4(4N+3)^2}.$$

For every $z^\star \in \arg\min g$, the point $(z^\star, 0)$ minimizes the lifted objective F_0, and $\nabla f_g(z^\star, 0) = (0, -\tau)$. Hence the lifted problem satisfies $\tau_\star = \tau$. The initial distance satisfies

$$\mathbf{dist}\big((z^{(0)}, 0), X^\star\big) \le \mathbf{dist}(z^{(0)}, \arg\min g) \le R,$$

so the lifted instance belongs to $\mathcal{P}_{\ell,\tau}(R)$. If $\varepsilon < L_\tau R^2/(4(4N+3)^2)$, the gap domination (6) rules out an ε-optimal output of A at every $t \le N$. Choosing N of order $R\sqrt{L_\tau/\varepsilon}$ with a sufficiently small universal coefficient gives the claimed bound for $0 < \varepsilon < c_0 L_\tau R^2$. Since A was arbitrary, we obtain

$$T_\varepsilon\big(\mathcal{A}_p, \mathcal{P}_{\ell,\tau}(R)\big) = \Omega\left(R\sqrt{\frac{\ell(\tau)}{\varepsilon}} \right).$$

We next consider the μ-strongly convex case with $\sigma = \mu$. Set $\kappa := L_\tau/\mu \ge 4$ and $q := (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$. The quadratic lower bound in Nemirovski (1994, Section 12.2 and Proposition 12.3.2), applied with the objective-gap accuracy measure, gives a quadratic g satisfying $\mu I \preceq \nabla^2 g \preceq L_\tau I$ and $g(z^{(0)}) - g^\star \le \Delta_0$ such that

$$\min_{0 \le t \le N} \big(g(z^{(t)}) - g^\star \big) \ge \frac{\Delta_0}{2} \left(\frac{2q^{4N+3}}{1 + q^{8N+6}} \right)^2 \ge \frac{\Delta_0}{2} q^{8N+6}.$$

Indeed, for a budget of B calls, that result reduces the lower bound to the Chebyshev approximation problem of degree $2B+1$, with dimension requirement $m > 2B+1$. The normalization follows from $\left\| (\nabla^2 g)^{1/2}(z - z^\star) \right\|^2 = 2(g(z) - g^\star)$; we take $B = 2N+1$ and again allow a factor of two in the minimax bound. The matrix-vector oracle used in that result, with the linear term supplied, can simulate the value and gradient of the quadratic. Thus the bound applies to the first-order oracle considered here. With $\sigma = \mu$, the lifted objective F_μ is μ-strongly convex, has $\tau_\star = \tau$, and satisfies

$$F_\mu(z^{(0)}, 0) - F_\mu^\star = g(z^{(0)}) - g^\star \le \Delta_0.$$

Hence the lifted instance belongs to $\mathcal{P}_{\ell,\tau,\mu}(\Delta_0)$. The lower bound therefore applies to every output of A through budget N. For $\kappa \ge 4$, we have $\log(1/q) \le 8/(3\sqrt{\kappa})$, so

$$\frac{\Delta_0}{2} q^{8N+6} \ge \frac{\Delta_0}{2} \exp\left(-8 - \frac{64N}{3\sqrt{\kappa}} \right).$$

For a sufficiently small universal $c_0 > 0$, this excludes an ε-optimal output before a universal constant multiple of $\sqrt{\kappa}\log(\Delta_0/\varepsilon)$ calls whenever $0 < \varepsilon < c_0\Delta_0$. Since the instance classes allow arbitrary finite dimensions, the dependence of m on N is permitted. Taking the worst case over these instances yields

$$T_\varepsilon\big(\mathcal{A}_p, \mathcal{P}_{\ell,\tau,\mu}(\Delta_0)\big) = \Omega\left(\sqrt{\frac{\ell(\tau)}{\mu}}\log\frac{\Delta_0}{\varepsilon}\right).$$

□

B Experimental Details

Convex examples. We consider three convex composite problems $F(x,y) = f(x,y) + |x|$, with the smooth objectives and initial points

$$\begin{aligned} f(x,y) &= e^x + e^{1-x} + \frac{\lambda}{6}|y|^6, && (x_0,y_0) = (-6,-5),\\ f(x,y) &= \frac{(x-2)^4}{4} + \frac{\lambda}{6}|y|^6, && (x_0,y_0) = (-4,-5),\\ f(x,y) &= -\log\cos(x-c) + \frac{\lambda}{8}|y|^8, && (x_0,y_0) = (1/4,-2). \end{aligned}$$

Here $\lambda = 5\times 10^{-3}$ for the exponential and quartic examples, $\lambda = 10^{-4}$ for the log-cosine example, and $c = 1/2 + \pi/4$. For each problem, all methods use the same initial point. The proximal map is $\mathrm{prox}_{ah}(u,v) = (\mathrm{sign}(u)\max\{|u|-a,0\}, v)$. The higher-order terms in y make these objectives convex but not strongly convex. Each minimizer satisfies $x^\star > 0$ and $y^\star = 0$, so optimality gives $\nabla f(x^\star,y^\star) = (-1,0)$ and $\tau_\star = 1$.

Strongly convex examples. The quartic, exp-square, and ellipsoid-exp tests use the following smooth components of $F = f + h$, with the initial points listed on the right

$$\begin{aligned} f(x,y) &= \frac{(x-2)^4}{4} + \frac{(x-2)^2}{2} + \frac{\mu}{2}y^2, && (x_0,y_0) = (-4,-5),\\ f(x,y) &= e^{x^2/2} - 1 + \frac{\mu}{2}y^2, && (x_0,y_0) = (-4,-5),\\ f(u,v) &= e^{u^\top Au/2} - 1 + \frac{1}{2}v^\top Dv, && (u_0,v_0). \end{aligned}$$

The two-dimensional problems use $\mu = 10^{-2}$ and $\mu = 10^{-3}$, respectively. For $t \geq 0$, the principal Lambert function $W(t)$ is the unique nonnegative number satisfying $W(t)e^{W(t)} = t$; set $a = \sqrt{W(1)}$. For ellipsoid-exp, $u, v \in \mathbb{R}^{500}$, $\|e\| = 1$, and $Ae = e$. The diagonal entries of D are geometrically spaced from 10^{-3} to 10^{-1}. We set $A = Q\Lambda Q^\top$, with orthogonal Q and eigenvalues geometrically spaced from 1 to 4. The matrix Q is a Householder reflection generated from a normalized Gaussian vector, and $e = Qe_1$. The regularizers are $h(x,y) = |x|$ for quartic, $h(x,y) = 2|x-a|$ for exp-square, and $h(u,v) = 2\|u-ae\|$ for ellipsoid-exp and all $\tau_\star = 1$