Near-Optimal Acceleration for Smooth ℓ_p/ℓ_q Nondual Convex First-Order Oracle Optimization

David Martínez-Rubio DAVID.MARTINEZRUBIO@IMDEA.ORG
IMDEA Software Institute, Madrid, Spain

Brian Bullins* BBULLINS@PURDUE.EDU
Purdue University, West Lafayette, IN, USA

Cristóbal Guzmán* CRGUZMANP@UC.CL
Institute for Mathematical and Computational Engineering, Faculty of Mathematics and School of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

Mathieu Molina* MATHIEU.MOLINA.RESEARCH@GMAIL.COM
Tel Aviv University, Israel

Abstract

We study the optimization of convex objectives with $(L, \kappa-1)$-Hölder-continuous gradients in ℓ_q over RB_p^d, $1 < \kappa \le 2$. Martínez-Rubio and Guzmán [MG26] provides selectors with a movement bound for the problem of chasing high-dimensional convex nested sets for every $p < q$ and generally reduces Lipschitz convex optimization to bounds on the movement of selectors. We couple that movement with Hölder descent yielding a polynomial-runtime first-order method whose feasible output, in the high-dimensional regime $T \le d$ and for $p < \min\{q, 2\}$, has error

$$\widetilde{O}_{\kappa,p,q}\left(\frac{LR^\kappa}{T^{\kappa(1+1/p-(1/q-1/2)_+)-1}}\right),$$

after T queries to a first-order oracle, solving the COLT 2015 open problem of Guzmán [Guz15], up to logarithmic factors. At $(p, q) = (1, 2)$, the rate is $\widetilde{O}(LR^\kappa/T^{2\kappa-1})$, including $\widetilde{O}(LR^2/T^3)$ cubic decay in the smooth case.

1. Introduction and result

The open problem of Guz15 asks for the minimax risk of black-box convex optimization when the domain is an ℓ_p ball but smoothness is measured in a different ℓ_q norm. The mismatch matters most for $p < \min\{q, 2\}$: the feasible set is smaller than the natural ℓ_q ball, yet classical accelerated methods use only ℓ_p or ℓ_q geometry and do not match known lower bounds. For instance at $(p, q) = (1, 2)$ and Lipschitz gradient smoothness, the classical rate is $O(T^{-2})$ while the lower bound predicts $O(T^{-3})$.

The nonsmooth endpoint of this question was solved in (MG26) by reducing a level bundle method to movement of nested convex bodies in high dimensions. This paper shows that the same movement primitive also supplies the missing acceleration for every $1 < \kappa \le 2$.

. This draft is not yet in the form in which we would have liked to share it. In light of recent developments, we have nevertheless decided to make it available now. We plan to polish and revise this work shortly.

*. Authors marked with an asterisk are listed alphabetically.

The mechanism is not an ordinary acceleration wrapper. A descent step contributes to a reduction in function value in the form of a positive power of the gradient norm, while a not-small-enough evaluation produces a deep cut in the bundle sublevel set and forces movement inversely proportional to that norm. The means inequalities balances the two effects, yielding acceleration, after a coupling of the movement selector and the descent point.

More generally, our reduction takes as input a feasible selector with the required movement bound and inherits polynomial running time when that selector is computable in polynomial time on the generated bodies. Martínez-Rubio and Guzmán [MG26] supplies the selectors and their implementation. The contribution here is the conversion of movement into Hölder acceleration, including the coupling that preserves feasibility for constrained optimization.

If the problem is unconstrained and a global minimizer x^* is known to satisfy $\|x^*\|_p \leq R$, we can provide a simple analysis of an algorithm that runs the movement selector on the auxiliary body RB_p^d, while descent steps and the final output are allowed to fall outside and returns a point outside it. For the full solution, we want to return a feasible point, and we would not assume there is a zero gradient in the set. After presenting our solution in the simplified setting, the second half of the paper constructs a solution based on approximating a Moreau envelope which, up to handling errors, reduces the problem to the previous case.

1.1. Setting

We use $1/\infty = 0$. For $1 \leq r \leq \infty$, the conjugate exponent r^* satisfies $1/r + 1/r^* = 1$, and $B_r^d = \{x \in \mathbb{R}^d : \|x\|_r \leq 1\}$. We write $\langle u, v\rangle = \sum_{i=1}^d u_i v_i$ for the standard inner product.

Assume

$$\|\nabla f(v) - \nabla f(u)\|_{q^*} \leq L\|v-u\|_q^{\kappa-1}, \qquad 1 < \kappa \leq 2. \tag{1}$$

Integrating the gradient on a segment gives

$$f(v) \leq f(u) + \langle \nabla f(u), v-u\rangle + \frac{L}{\kappa}\|v-u\|_q^{\kappa}. \tag{2}$$

For $p < q$, define the movement exponent

$$\rho_{p,q} \coloneqq \frac{1}{p} - \left(\frac{1}{q} - \frac{1}{2}\right)_+. \tag{3}$$

Throughout, $\widetilde{O}_{\kappa,p,q}$ omits constants depending on fixed regularity and norm parameters and factors polynomial in logarithms of d, T, the confidence level, and the requested numerical accuracy; it never omits a polynomial in d.

Theorem 1 (Nondual Hölder accelerated optimization) [↓] *Let $R, L > 0$, let $\mathcal{Q} = RB_p^d$, let $1 \leq p, q \leq \infty$, and let f be convex and differentiable on a neighborhood of $\mathcal{Q}$, satisfying (1) for some $1 < \kappa \leq 2$. Fix $\delta \in (0,1)$. For $T \leq d$, and either $p < \min\{q, 2\}$ or $p = q \leq 2$, there is a polynomial-time algorithm in the real-arithmetic model, making at most T first-order oracle queries and returning $\widehat{x}_T \in \mathcal{Q}$ with probability at least $1-\delta$ such that*

$$f(\widehat{x}_T) - \min_{u \in \mathcal{Q}} f(u) \leq \widetilde{O}_{\kappa,p,q}\left(\frac{LR^\kappa}{T^{\kappa(1+1/p-(1/q-1/2)_+)-1}}\right), \tag{4}$$

Polynomial running time is understood in the real-arithmetic model of Section 9, with the selector implementation verified in Proposition 14, with fixed regularity and norm parameters. We note that there is also a deterministic algorithm if we use the deterministic stable selectors in MG26. We do not claim efficiency for that deterministic algorithm but we note that MG26 showed lower bounds of essentially the same order against randomized algorithms in comparison to the ones in the open question (Guz15) for the Lipschitz problem.

The rate above, along with the observation that for $2 \le p < q$ one can run the classical modified Nesterov method in the p-geometry (GN15, Section 4.1.B) to match the lower bounds in Guzmán and Nemirovski [GN15] up to log factors, solves the COLT 2015 open problem (Guz15). The randomized implementation attains the same upper bound; we do not infer a randomized lower bound from the deterministic one.

Related optimization methods. The diagonal $p = q$ smooth rates go back to classical work (NN85), while universal methods cover unknown Hölder regularity (Nes15). Our algorithm draws on level and bundle methods (LNN95) as realized by (MG26). The new point is the coupling to the nonstandard movement geometry and the feasible envelope construction.

An independent work (Ouy26) appeared on arXiv as we prepared the writing of our results. The work studies the specific case of quadratics when $(p, q) = (1, 2)$ for the smooth case ($\kappa = 2$).

2. Warm-up: power descent versus harmonic movement

In order to understand the solution, it is convenient to understand how the non-smooth rate of (MG26) is achieved. For intuition, we include Figure 1, borrowed from the paper.

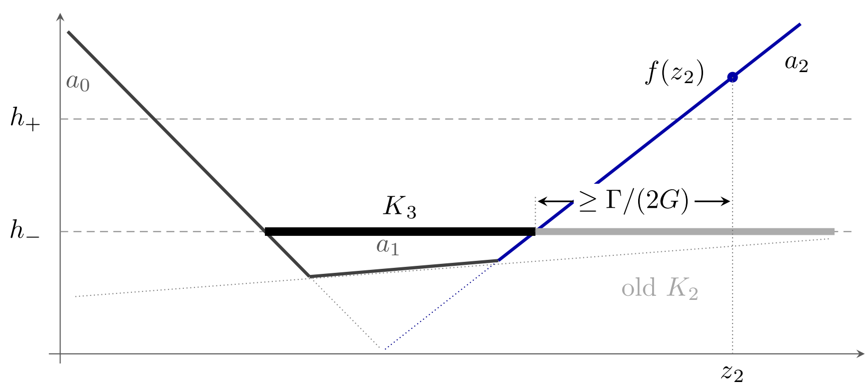

Figure 1: A one-dimensional section of a step where the function value is no less than h_+ (null step). The old model $\max\{a_0, a_1\}$ lies below h_- at z_2 (that is, $z_2 \in K_2$), whereas the new tangent satisfies $a_2(z_2) = f(z_2) > h_+$. Adding a_2 reduces the (h_-)-sublevel to the thick black segment ($K_3 \subset K_2$), separated from z_2 by at least $\Gamma/(2G)$ as the subgradient norm is no larger than G. Recall that Γ is the duality gap at the beginning of a phase.

Non-smooth solution from (MG26) Before query $t \geq 1$, let

$$m_t(x) = \max_{0 \leq i < t} \{f(x_i) + \langle g_i, x - x_i \rangle\}, \qquad \ell_t = \min_{x \in \mathcal{Q}} m_t(x), \qquad U_t = \min_{0 \leq i < t} f(x_i).$$

Convexity gives $\ell_t \leq \min_{x \in \mathcal{Q}} f(x) \leq U_t$. We run the algorithm in phases. Freeze the phase scale $\Gamma = (U - \ell)/2$ and two levels $h_- = \ell + \Gamma/2 < h_+ = \ell + 3\Gamma/2$, and query a stable selector $x_t = z_t \in K_t := \{x \in \mathcal{Q} : m_t(x) \leq h_-\}$. If $f(z_t) \leq h_+$, the upper bound improves by a constant fraction. Otherwise the new tangent $a_t(x) = f(z_t) + \langle g_t, x - z_t \rangle$ gives $K_{t+1} = K_t \cap \{a_t \leq h_-\}$. This retained halfspace is at ℓ_q-distance at least $(h_+ - h_-)/\|g_t\|_{q^*} = \Gamma/\|g_t\|_{q^*}$ from z_t. Thus every such *deep cut* forces movement of the selector, unless the new body is empty; see Figure 1. An upper bound on the total movement $\sum_{t=1}^{T} \|z_t - z_{t+1}\|_q$ after T steps coming from the chasing convex bodies problem along with the fact that the movement is lower bounded by $T \cdot \Gamma/G$ gives a maximum number of steps until we see an empty sublevel set (so the lower bound increases to at least h_-) or we have $f(z_t) \leq h_+$. Either way, we decrease the gap by a constant factor and continue to the next phase.

Hölder-smooth warm-up In this warm-up, suppose the function is convex and Hölder smooth globally but its global minimum is in RB_p^d while we initialized at 0, and we allow descent points y_t and the points x_t that we use to query the first-order oracle to be outside of RB_p^d while we have a stable selector $z_t \in RB_p^d$. We assume being able to perform exact segment function value minimization in this warm-up, for simplicity. The algorithm has some similarities to the nonsmooth solution: we run the bundle method, except that within a phase, we query at $x_t \in \operatorname{argmin}_{x \in [y_t, z_t]} f(x)$ receiving $f(x_t)$ and $g_t = \nabla f(x_t)$, then descend to y_{t+1} from x_t optimizing the unconstrained upper bound (2) on the function that Hölder smoothness gives while z_t is the proposed stable selector. By optimality, the first-order oracle query at x_t also either has function value at most h_+ or it produces a deep cut separating z_t from the next sublevel set.

Line search gives $f(x_t) \leq f(y_t)$ and $\langle g_t, z_t - x_t \rangle \geq 0$: descent improves the old value, and the tangent $a_t(u) = f(x_t) + \langle g_t, u - x_t \rangle$ satisfies $a_t(z_t) \geq f(x_t)$. For the smooth method, replace the fixed levels by the moving target below. If descent guarantees decrease D_t, set $U_{t+1} = f(x_t) - D_t$, add a_t to the bundle, and lower all cuts' target to $U_{t+1} - \Gamma$, with phase scale Γ. Convexity preserves all points of RB_p^d below this target level; the bodies stay nested. Thus the same gradient gives decrease D_t and forces selector movement at least $(\Gamma + D_t)/\|g_t\|_{q^*}$, unless the next body is empty. In particular, the deep cut at x_t is also a deep cut for z_t and we force movement similarly as if we had queried at z_t instead.

For $p < q$, Martínez-Rubio and Guzmán [MG26] provides, for every nonempty nested sequence $RB_p^d \supseteq K_0 \supseteq \cdots \supseteq K_N$, a selector $z_t \in K_t$ satisfying

$$\sum_{t<N} \|z_{t+1} - z_t\|_q = \widetilde{O}_{p,q}\left(RN^{1-\rho_{p,q}}\right),$$

where $\rho_{p,q}$ is defined in (3).

Let g be a gradient and $G = \|g\|_{q^*}$. Choose a support vector $v_q(g)$, i.e. such that $\|v_q(g)\|_q = 1$ and $\langle g, v_q(g) \rangle = G$. Minimizing the Hölder upper model along $-v_q(g)$ gives the step length and guaranteed decrease in function value $\mathcal{D}_\kappa(G)$, where

$$a = (G/L)^{1/(\kappa-1)}, \qquad \mathcal{D}_\kappa(G) := \frac{\kappa - 1}{\kappa} L^{-1/(\kappa-1)} G^{m_\kappa}, \text{ for } m_\kappa = \frac{\kappa}{\kappa - 1}.$$

Thus gradient descent controls the m_κ-power mean of the gradient norms.

If $G_t = 0$, the queried point is a global minimizer and we stop. Otherwise, a tangent cut with level violation Γ and subgradient norm G_t has depth Γ/G_t. Hence a purely movement-based estimate controls

$$\Gamma \sum_{t<N} \frac{1}{G_t},$$

or equivalently the harmonic mean of the subgradient norms. The inequality between the harmonic mean and the m_κ-power mean is

$$\frac{N}{\sum_{t<N} 1/G_t} \leq \left(\frac{1}{N} \sum_{t<N} G_t^{m_\kappa}\right)^{1/m_\kappa}. \tag{5}$$

If one phase has scale Γ, the concatenated guaranteed descent is at most $U_0 - f^\star \leq U_0 - \ell_0 = 2\Gamma$. For N nonterminal transitions in this phase, descent gives

$$\sum_{t=0}^{N-1} G_t^{m_\kappa} \lesssim_\kappa L^{1/(\kappa-1)}\Gamma,$$

whereas movement gives

$$\Gamma \sum_{t=0}^{N-1} \frac{1}{G_t} = \widetilde{O}\big(RN^{1-\rho}\big).$$

Substituting these estimates into (5) and using $m_\kappa = \kappa/(\kappa-1)$ yields

$$N^{\kappa-1+\kappa\rho} = \widetilde{O}_\kappa\left(\frac{LR^\kappa}{\Gamma}\right).$$

This is the desired exponent that yields optimality up to log factors. The line search makes both estimates simultaneously valid; the later envelope construction preserves this tradeoff using quadratic descent and phase tuning, with controlled numerical errors and feasible queries.

To sum up, we know from previous work that we have certain movement bound $\widetilde{O}(RN^{1-\rho})$ with using some stable selectors z_t. We keep descent points y_t from doing classical in the unconstrained problem, and we select the point x_t where we will do the first-order oracle query as the minimizer of f on $[y_t, z_t]$ (or an approximation with controlled errors). Optimality of the line search creates a deep cut over z_t in the nested convex bodies subproblem, while ensuring $f(x_t) \leq f(y_t)$. This is reminiscent of linear coupling and other accelerated regimes combining mirror descent and descent steps. Smaller gradient norms G_t strengthen the guaranteed cut depth Γ/G_t, while larger norms strengthen the guaranteed descent. The means inequality balances those two phenomena. The use of this inequality is reminiscent of the intuitive explanation in Allen-Zhu and Orecchia [AO14, Thought Experiment, p. 4] (that inspired the idea of this paper) and, in hindsight, of the use of the reverse Hölder inequality in higher-order acceleration (CHJ+22; ABJ+24; CGM24).

We note that convergence rates depending on the harmonic mean of the gradient norms is an identified phenomenon, see (Lev17; Ora23). Our work provides an intuitive geometric explanation to this phenomenon in the form of our realized moment bound.

3. Algorithm at a glance

The method runs in phases. Each inner bundle call builds a tangent model, and the outer level method retains its certified aggregate cut. The maximum of the retained aggregate cuts forms a model $m \le f$ on $\mathcal{Q}$. Minimizing m gives a certified lower bound, while a retained feasible point gives an upper bound. Each phase fixes a fraction of this certified gap, forms a corresponding model sublevel set, and queries the movement selector inside that set. The coupled-step module then returns a feasible upper witness and a new valid affine cut.

Algorithm 1 Feasible Hölder movement-coupled level method

Input: Feasible set $\mathcal{Q} = RB_p^d$; first-order oracle; Hölder parameters (κ, L); accuracy ε; movement selector C; query budget J from (26).

Input: Feasible envelope step Step; initial feasible upper witness y, upper certificate $U \ge f(y)$, and max-affine model $m \le f$ on $\mathcal{Q}$.

Output: A feasible witness; its certified gap is at most ε on simultaneous selector success.

```
Count every first-order call, including initialization and inner line-search calls. Before call J + 1, return the retained y.
ℓ ← min_{u∈Q} m(u).
while U − ℓ > ε do
    Γ ← (U − ℓ)/2 and U^start ← U.
    Choose the planned length N_Γ by (25); set the envelope coefficient A_Γ, bundle tolerance τ_Γ, and line-search tolerance η_Γ by (17).
    Restart the horizon-dependent selector C for N_Γ transitions (N_Γ + 1 selector points); set n ← 0.
    while U > U^start − Γ/4 do
        K ← {u ∈ Q : m(u) ≤ U − Γ}.
        if K = ∅ then
            break.
        end if
        If n = N_Γ + 1, return y; otherwise set n ← n + 1.
        z ← C(K); if it signals failure, return y.
        (y, U, a, G, η, ξ) ← Step(y, U, z, Γ; A_Γ, τ_Γ, η_Γ).   ⋄ valid cut a ≤ f, cut-normal norm G, errors η, ξ
        m(u) ← max{m(u), a(u)} for u ∈ Q.
    end while
    ℓ ← min_{u∈Q} m(u).
end while
return ŷ ← y.
```

The step $\mathsf{Step}(y, U, z, \Gamma; A, \tau, \eta)$ uses the bracketed envelope line search of Lemma 10: at each trial center it calls the certified proximal bundle solver of Proposition 9, then returns the feasible upper witness and aggregate cut described in Lemma 11. The step retains the old witness whenever it retains the old upper certificate. The displayed pseudocode uses exact model solves; Proposition 13 specifies the certified numerical replacements for its solves and emptiness tests. The query guard applies inside every subroutine; an interrupted call retains the previous witness. The selector uses the bounded-work implementation of Martínez-Rubio and Guzmán [MG26], as specified in Section 9. The certified gap guarantee holds with probability at least $1-\delta$. Thus the upper certificate U, phase scale Γ, cut-normal norm G, descent inexactness η, and cut inexactness ξ are the only quantities passed from the analytic module to the movement argument.

4. The abstract coupled phase

Maintain a max-affine bundle model $m_t \le f$ on $\mathcal{Q}$, a lower certificate $\ell_t = \min_{\mathcal{Q}} m_t$, a feasible upper witness y_t, and an upper certificate $U_t \ge f(y_t)$. A phase begins with scale $\Gamma = (U_0 - \ell_0)/2$. Its body is

$$K_t = \{u \in \mathcal{Q} : m_t(u) \le U_t - \Gamma\}\,.$$

The bodies are nested because m_t increases and U_t decreases.

The coupled step allows two geometrically different errors. The descent error η_t is vertical slack in the upper-bound decrease: it is the amount lost from the ideal decrease $\mathcal{D}(G_t)$ for a specified nonnegative descent profile $\mathcal{D}$. The cut error ξ_t is vertical slack in the separation at z_t: for cut normal g_t, it makes the retained halfspace shallower by $\xi_t/\|g_t\|_{s^*}$. Thus η_t spends the phase's progress budget, whereas ξ_t spends its cut depth. An implementation obtains descent and cuts with some errors ξ_t η_t, and we show that the rate of convergence is robust to it.

Definition 2 (Coupled step with qualified errors) *Fix a descent profile $\mathcal{D} : [0,\infty) \to [0,\infty)$. At center $z_t \in K_t$, a coupled step returns an upper certificate $U_{t+1} \le U_t$, a valid affine cut $a_t(u) = b_t + \langle g_t, u\rangle$, cut-normal norm $G_t = \|g_t\|_{s^*}$, descent inexactness $\eta_t \ge 0$, and cut inexactness $\xi_t \ge 0$, such that*

$$U_{t+1} \le U_t - \mathcal{D}(G_t) + \eta_t, \tag{6}$$

$$a_t(z_t) - (U_{t+1} - \Gamma) \ge \Gamma + \mathcal{D}(G_t) - \xi_t. \tag{7}$$

The movement norm is ℓ_s and the cut normal is measured in ℓ_{s^*}.

Lemma 3 (Cut depth and inverse normal norm) [↓] *Let G_t be the cut-normal norm and ξ_t the cut inexactness from Definition 2. If $\xi_t \le \Gamma/2$ and the next body is nonempty, then*

$$\|z_{t+1} - z_t\|_s \ge \frac{\Gamma}{2G_t}.$$

Note that if $G_t = 0$, we can stop the algorithm, so zero never occurs in an inverse-norm sum.

Theorem 4 (Movement-to-optimization reduction) [↓] *Let $1 < \kappa \le 2$ and $\rho > 0$. Suppose a selector for nested convex sets in RB_p^d satisfies $\sum_{t=0}^{N-1} \|z_{t+1} - z_t\|_s = \widetilde{O}(RN^{1-\rho})$. Consider the level updates of Algorithm 1, using coupled steps as in Definition 2 with profile $\mathcal{D} = \mathcal{D}_\kappa$, with a budget of $T \ge 1$ coupled steps in place of the accuracy and first-order-call stopping rules. Stop earlier only if the certified gap vanishes. Assume that, in each phase of* *scale Γ, the errors satisfy $\sum_{t=0}^{N-1} \eta_t \le \Gamma/16$ and $0 \le \xi_t \le \Gamma/2$ for every nonterminal prefix of length N. From an initial certified gap $O(LR^\kappa)$, the retained upper witness after at most T coupled steps has optimization error at most*

$$\widetilde{O}_\kappa\left(\frac{LR^\kappa}{T^{\kappa-1+\kappa\rho}}\right).$$

5. Module I: a global minimizer lies in the auxiliary ball

Assume in this section that $f : \mathbb{R}^d \to \mathbb{R}$ and that a global minimizer x^* satisfies $x^* \in \mathcal{Q}$. Thus $\nabla f(x^*) = 0$, and a value at an infeasible point is still a valid upper bound on $f(x^*)$.

Given a current upper witness y and movement center z, minimize f on the segment $[y, z]$:

$$x \in \operatorname*{argmin}_{w \in [y,z]} f(w), \qquad g = \nabla f(x), \qquad G = \|g\|_{q^*}.$$

Set

$$y^+ = x - (G/L)^{1/(\kappa-1)} v_q(g), \qquad U^+ = f(x) - \mathcal{D}_\kappa(G), \tag{8}$$

and use the tangent

$$a(u) = f(x) + \langle g, u - x \rangle.$$

Lemma 5 (Exact ambient coupling) [↓] *Let U be the current upper certificate, G the gradient norm in (8), and Γ the phase scale. Then*

$$f(y^+) \le U^+ \le U - \mathcal{D}_\kappa(G), \qquad a(z) - (U^+ - \Gamma) \ge \Gamma + \mathcal{D}_\kappa(G).$$

Thus the ambient module satisfies Definition 2 with $\mathcal{D} = \mathcal{D}_\kappa$ and zero descent and cut errors.

Theorem 6 (Ambient rate with exact line search) [↓] *Let $p < \min\{q, 2\}$, let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable and satisfy (1), and suppose a global minimizer obeys $\|x^*\|_p \le R$. Use RB_p^d only for the bundle lower model and the movement selector, and grant the segment-minimization oracle used in (8). Then after T coupled outer steps the ambient module returns a point $y_T \in \mathbb{R}^d$, not necessarily in RB_p^d, with*

$$f(y_T) - f(x^*) \le \widetilde{O}_{\kappa,p,q}\left(\frac{LR^\kappa}{T^{\kappa(1+\rho_{p,q})-1}}\right).$$

The theorem explains the cleanest use of an externally supplied radius bound. The auxiliary ball restricts the comparator and makes movement finite; it does not restrict the descent path or output. This argument is valid only when the optimization problem itself is unconstrained. An approximate one-dimensional search can replace the ideal oracle by assigning its value and sign errors to the budgets in Definition 2; the feasible construction below gives the fully specified polynomial algorithm used in the main theorem.

6. Why constrained feasibility requires a new module

For $\min_{\mathcal{Q}} f$, the point y^+ in (8) can be infeasible, and the constrained minimizer can have nonzero gradient. Projecting y^+ back to $\mathcal{Q}$ does not preserve either the Hölder decrease or the tangent-cut alignment. A direct κ-power proximal envelope suggests the same descent profile, but we do not establish the required linear-time proximal-solve bound for that construction.

We instead use the "inexact gradient trick" at the current phase accuracy, a classical technique from Devolder, Glineur, and Nesterov [DGN14]. For a tolerance $\omega > 0$, define

$$H_\omega := \begin{cases} \left(\dfrac{2-\kappa}{2\kappa}\right)^{(2-\kappa)/\kappa} L^{2/\kappa}\omega^{-(2-\kappa)/\kappa}, & 1 < \kappa < 2, \\ L, & \kappa = 2. \end{cases} \tag{9}$$

Lemma 7 (Inexact gradient trick) [↓] *For every $u, v \in \mathcal{Q}$,*

$$f(v) \le f(u) + \langle \nabla f(u), v-u\rangle + \frac{H_\omega}{2}\|v-u\|_q^2 + \omega.$$

For the feasible construction, use $s = \min\{q, 2\}$. For a center $c \in \mathcal{Q}$ and coefficient $A > 0$, define the quadratic envelope

$$M_A(c) := \min_{u\in\mathcal{Q}} \left\{ f(u) + \frac{A}{2}\|u-c\|_s^2 \right\}.$$

Let $J_s(w) = \nabla(\|w\|_s^2/2)$ denote the duality map.

Definition 8 (Certified envelope evaluation) *At center c, a bundle call with tolerance τ returns a convex max-affine model $m \le f$ and points $x, y \in \mathcal{Q}$ such that*

$$x \in \operatorname*{argmin}_{u\in\mathcal{Q}} \left\{ m(u) + \frac{A}{2}\|u-c\|_s^2 \right\}, \tag{10}$$

$$f(y) + \frac{A}{2}\|y-c\|_s^2 \le m(x) + \frac{A}{2}\|x-c\|_s^2 + \tau. \tag{11}$$

The aggregate residual and cut are

$$g = AJ_s(c-x), \qquad a(v) = m(x) + \langle g, v-x\rangle. \tag{12}$$

Proposition 9 (Certified bundle solve) [↓] *Let $1 < s \le 2$, let $\mathcal{Q} \subseteq RB_s^d$ be compact and convex, and let f be convex and differentiable on a neighborhood of $\mathcal{Q}$. Fix $c \in \mathcal{Q}$, $A, \tau > 0$, and $0 < \omega \le \tau/2$. Suppose that*

$$f(v) \le f(u) + \langle \nabla f(u), v-u\rangle + \frac{H_\omega}{2}\|v-u\|_s^2 + \omega \qquad (u, v \in \mathcal{Q}).$$

Set $\mu = A(s-1)$. Repeating (10) yields (11) using

$$O\left(\left(1 + \frac{H_\omega}{\mu}\right)\log\left(2 + \frac{H_\omega R^2}{\tau}\right)\right) \tag{13}$$

first-order calls.

Exact model solves are used above. We note that Proposition 13 gives the same order of calls with certified approximate solves when $H_\omega/\mu = O_s(1)$. The proof contracts the global model gap directly. At the coefficient used below, $H_\omega/\mu = O_s(1)$, so this nonaccelerated inner solver already has logarithmic oracle cost.

7. The feasible envelope coupling

Let y^{old} be the current feasible upper witness and z the movement center. We evaluate the envelope along

$$c_\lambda = y^{\mathrm{old}} + \lambda(z - y^{\mathrm{old}}), \qquad 0 \le \lambda \le 1.$$

At each queried c_λ, compute the certified residual g_λ and the scalar

$$\psi_\lambda = \left\langle g_\lambda, z - y^{\mathrm{old}} \right\rangle.$$

Lemma 10 (Certified envelope line search) [↓] *Using envelope evaluations with bundle tolerance τ, sign bisection returns a center $c \in [y^{\mathrm{old}}, z]$ and certified residual g such that*

$$M_A(c) \le M_A(y^{\mathrm{old}}) + \eta + \tau, \qquad \langle g, z - c\rangle \ge 0.$$

The number of envelope evaluations is $O(\log(2 + AR^2/\eta))$.

By Lemma 11, the envelope-value bound preserves telescoping descent up to budgeted errors, while $\langle g, z - c\rangle \ge 0$ keeps the cut deep at z. Only selector centers enter the movement sum, not line-search trials. The nested-body movement bound therefore still applies.

Lemma 11 (Feasible descent and cut) [↓] *Let U be the old feasible upper certificate, let Γ be the phase scale, and let (x, y, m) be the bundle certificate at the line-search center c. Put $G = \|g\|_{s^*}$. The aggregate affine function in (12) satisfies $a \le f$ on $\mathcal{Q}$, and the new feasible upper certificate $U^+ = \min\{U, f(y)\}$ obeys*

$$U^+ \le U - \frac{G^2}{4A} + C_s(\eta + \tau), \tag{14}$$

$$a(z) - (U^+ - \Gamma) \ge \Gamma + \frac{G^2}{2A} - C_s(\eta + \tau). \tag{15}$$

Its witness is y^{old} if $U \le f(y)$, and y otherwise. The cut is retained in both cases. Here g is an aggregate residual, not necessarily a subgradient of f. It is a τ-subgradient of M_A by (22).

This is Definition 2 with $\mathcal{D}(G) = G^2/(4A)$ and errors $C_s(\eta + \tau)$. Its numerical version in Proposition 13 uses $\mathcal{D}(G) = G^2/(8A)$. Before upper progress terminates the phase, the descent inequality gives $\sum G_t^2 \lesssim A\Gamma$. The cut inequality and movement give $\Gamma \sum 1/G_t = \widetilde{O}(RN^{1-\rho})$. Hölder with exponents 3 and 3/2 yields the quadratic phase bound

$$N^{1+2\rho} = \widetilde{O}_s\left(\frac{AR^2}{\Gamma}\right). \tag{16}$$

Choose planned phase length N_Γ and parameters

$$\tau_\Gamma \asymp \eta_\Gamma \asymp \frac{\Gamma}{N_\Gamma}, \qquad A_\Gamma \asymp_s L^{2/\kappa}\left(\frac{N_\Gamma}{\Gamma}\right)^{(2-\kappa)/\kappa}. \tag{17}$$

Substitution in (16) gives

$$N_\Gamma^{\kappa-1+\kappa\rho} = \widetilde{O}_{\kappa,s}\left(\frac{LR^\kappa}{\Gamma}\right). \tag{18}$$

Thus the feasible quadratic module has exactly the Hölder exponent found in the ambient warm-up.

Theorem 12 (Feasible movement-coupled rate for $p < \min\{q,2\}$) [↓] *Let $1 \le p < \min\{q,2\}$ and $1 < q \le \infty$, let $\mathcal{Q} = RB_p^d$, and let f be convex and differentiable on a neighborhood of $\mathcal{Q}$, satisfying (1). Use $s = q$ when $q \le 2$ and $s = 2$ when $q \ge 2$. Let $\rho_{p,q}$ be the movement exponent in (3), let U_t denote the feasible upper certificates, and let τ_Γ, η_Γ be the bundle and line-search tolerances in (17). For a first-order budget $T \ge 1$ and confidence parameter $\delta \in (0,1)$, run Algorithm 1 with initialization from Lemma 16 and the target accuracy and query budget from (27), or use its small-budget fallback. This returns $\widehat{x}_T \in RB_p^d$, with $f^\star = \min_{u\in\mathcal{Q}} f(u)$, satisfying*

$$f(\widehat{x}_T) - f^\star \le \widetilde{O}_{\kappa,p,q}\left(\frac{LR^\kappa}{T^{\kappa(1+\rho_{p,q})-1}}\right).$$

The bound holds with probability at least $1-\delta$, using at most T first-order calls, including initialization and all inner bundle and line-search calls.

8. Regime assembly and comparison with lower bounds

The feasible theorem is assembled as follows.

1. If $p < q \le 2$, use $s = q$ in the envelope and the p/q movement selector from the companion paper.

2. If $p < 2 \le q$, norm comparison gives

$$\|\nabla f(v) - \nabla f(u)\|_2 \le \|\nabla f(v) - \nabla f(u)\|_{q^*} \le L\|v-u\|_2^{\kappa-1}.$$

   Use the Euclidean envelope and the $p/2$ movement selector. Its exponent is $\rho_{p,2} = 1/p$, which is the required $q \ge 2$ exponent.

For $p < \min\{q,2\}$, the powers reported by Guzmán [Guz15], based on the lower-bound framework of Guzmán and Nemirovski [GN15], are

$$\kappa\left(\frac{3}{2} + \frac{1}{p} - \frac{1}{q}\right) - 1 \quad (q < 2), \qquad \kappa\left(1 + \frac{1}{p}\right) - 1 \quad (q \ge 2).$$

These agree with Theorem 1.

9. Implementation and scope

The reduction assumes a selector that returns feasible centers with movement $\widetilde{O}(RN^{1-\rho})$ on the generated nested bodies. A bounded-work implementation on these bodies, with the stated movement bound conditionally on the preceding history, is supplied by Proposition 14, making use of the selector's implementation in Martínez-Rubio and Guzmán [MG26, Theorem 9, Proposition 11, and Appendix E].

We work in the real-arithmetic model: arithmetic, comparisons, logarithms, rational powers, powers with the fixed norm and regularity exponents, and independent scalar standard Gaussian draws have unit cost. Exact first-order calls are counted separately. For fixed κ, p, q, the additional work is polynomial in d, $1 + LR^\kappa/\varepsilon$, $\log(1/\delta)$, and $\log(2 + \|\nabla f(0)\|_{s^*}/(LR^{\kappa-1}))$ in the new branch. The last dependence comes only from solving the inner tangent models; the first-order bound is independent of affine terms in f. The convex-solve and sampling bounds in Proposition 14, together with the hard budgets in Algorithm 1, bound the work on every run.

Acknowledgments

Cristóbal Guzmán was partially funded by ANID FONDECYT 1251029 grant, and ANID Basal FB210017 National Center for Artificial Intelligence CENIA. David Martínez-Rubio was funded by grant La Caixa Junior Leader Fellowship 2025. He thanks OpenAI for free access to their models. Mathieu Molina received funding from the European Research Council (ERC) under the European Union's Horizon Europe program (grant agreement No. 101170373), as a postdoctoral fellow at Tel Aviv University.

This work was elaborated in combination with ChatGPT/Codex. After providing the initial ideas (the contents of the warm-up Section 2 along with pointing to using the Moreau envelope analysis, which can absorb errors, plus treating Hölder smooth functions as inexact smooth functions), ChatGPT 5.6 (the newest model when we elaborated the results) was able to craft the details of the proof. This draft is not in the form we would have liked to share it and the authors will shortly improve its writing quality, organization of ideas and overall presentation.

Appendix A. Proof of the abstract phase theorem

Proof of Lemma 3. The new body lies in the halfspace

$$H_t = \{u : a_t(u) \le U_{t+1} - \Gamma\}\,.$$

If $G_t = 0$, the coupled cut is constant and its retained halfspace is empty. Suppose $G_t > 0$. A displacement w from z_t into H_t must satisfy $\langle g_t, w\rangle \le -[a_t(z_t) - (U_{t+1} - \Gamma)]$. Hölder's inequality gives $\|w\|_s \ge [a_t(z_t) - (U_{t+1} - \Gamma)]/G_t$. Equality is attained by a suitably scaled negative unit ℓ_s direction on which g_t attains its dual norm. Therefore the distance from the current center is

$$\operatorname{dist}_s(z_t, H_t) = \frac{a_t(z_t) - (U_{t+1} - \Gamma)}{G_t} \overset{①}{\ge} \frac{\Gamma - \xi_t}{G_t} \overset{②}{\ge} \frac{\Gamma}{2G_t}.$$

Here ① uses the coupled cut (7) and $\mathcal{D} \ge 0$, and ② uses $\xi_t \le \Gamma/2$. Since z_{t+1} belongs to the new body, its movement is at least this distance. ∎

Proof of Theorem 4. Fix a phase of scale Γ and a nonterminal prefix of length N. We first bound N and then sum the phase costs. Recall that U_t is the upper certificate and G_t the cut-normal norm. Here the descent profile is $\mathcal{D}_\kappa$. As long as upper progress has not terminated the phase, $U_0 - U_N < \Gamma/4$. Summing (6) and using the descent error budget gives

$$\sum_{t<N} G_t^{m_\kappa} \le C_\kappa L^{1/(\kappa-1)}\Gamma. \tag{19}$$

Unpack the assumed soft movement bound as $\sum_{t<N} \|z_{t+1} - z_t\|_s \le CR\Lambda(d,N)N^{1-\rho}$, where Λ is nondecreasing and polylogarithmic. The inverse-normal-norm lemma and this movement bound give

$$\frac{\Gamma}{2}\sum_{t<N}\frac{1}{G_t} \le R\Lambda(d,N)N^{1-\rho}. \tag{20}$$

Apply Hölder in the form

$$\begin{aligned} N &= \sum_{t<N} (G_t^{m_\kappa})^{1/(m_\kappa+1)} (G_t^{-1})^{m_\kappa/(m_\kappa+1)} \\ &\le \left(\sum_{t<N} G_t^{m_\kappa}\right)^{1/(m_\kappa+1)} \left(\sum_{t<N}\frac{1}{G_t}\right)^{m_\kappa/(m_\kappa+1)}. \end{aligned}$$

Substituting (19) and (20), raising to $m_\kappa + 1$, and then raising to $\kappa - 1$ yields

$$N^{\kappa-1+\kappa\rho} \le C_\kappa \frac{LR^\kappa \Lambda(d,N)^\kappa}{\Gamma}.$$

It remains to verify gap contraction. If upper progress occurs, the upper certificate drops by $\Gamma/4 = (U_0 - \ell_0)/8$, while the lower certificate does not decrease; the new gap is at most $7(U_0 - \ell_0)/8$. If the body $\{m \le U - \Gamma\}$ is empty, then $\ell = \min_{\mathcal{Q}} m \ge U - \Gamma$, so the new gap is at most $\Gamma = (U_0 - \ell_0)/2$. The bounds on phase costs grow geometrically as the gap shrinks and are therefore dominated by the final phase, including at most one terminal step per phase. Starting from a certified gap $O(LR^\kappa)$ and inverting the resulting bound on the total number T of coupled steps gives optimization error $\widetilde{O}_\kappa(LR^\kappa/T^{\kappa-1+\kappa\rho})$ for the retained upper witness. ■

Appendix B. The ambient module

Proof of Lemma 5. Exact minimization on $[y, z]$ gives

$$f(x) \le f(y), \qquad \langle g, z - x\rangle \ge 0. \tag{21}$$

Let $a = (G/L)^{1/(\kappa-1)}$. By (2) and the norming-vector identity,

$$f(x - av_q(g)) \le f(x) - aG + \frac{L}{\kappa}a^\kappa = f(x) - \frac{\kappa-1}{\kappa}L^{-1/(\kappa-1)}G^{\kappa/(\kappa-1)} = U^+.$$

Thus $f(y^+) \le U^+$. The first sign in (21) gives $U^+ \le U - \mathcal{D}_\kappa(G)$. Because x^* is a global minimizer, $U^+ \ge f(y^+) \ge f(x^*)$, so this is a valid upper certificate.

For the tangent cut,

$$a(z) - (U^+ - \Gamma) = \langle g, z - x\rangle + \mathcal{D}_\kappa(G) + \Gamma \overset{①}{\ge} \Gamma + \mathcal{D}_\kappa(G).$$

Here ① uses the second sign in (21). ■

Proof of Theorem 6. Apply Lemma 5 and Theorem 4 with the imported movement exponent $\rho_{p,q}$. To initialize, query at 0. Since $\nabla f(x^*) = 0$, (1) gives

$$\|\nabla f(0)\|_{q^*} \le L\|x^*\|_q^{\kappa-1} \le LR^{\kappa-1}.$$

The tangent range over RB_p^d is at most a constant times $R\|\nabla f(0)\|_{p^*} \le LR^\kappa$, where $p < q$. Hence the initial certified gap is $O(LR^\kappa)$. Geometric phases give the claimed rate. Nothing in the proof requires the upper witness to lie in the auxiliary ball. ■

Appendix C. Inexact gradient trick and the certified bundle solve

Proof of Lemma 7. It suffices to prove, for every $t \ge 0$,

$$\frac{L}{\kappa}t^\kappa \le \frac{H_\omega}{2}t^2 + \omega.$$

For $1 < \kappa < 2$, the maximum of the left side minus $Ht^2/2$ occurs at $t = (L/H)^{1/(2-\kappa)}$ and equals

$$\frac{2-\kappa}{2\kappa}L\left(\frac{L}{H}\right)^{\kappa/(2-\kappa)}.$$

The definition of H_ω makes this value equal to ω. Combining the scalar inequality with (2) proves the claim. At $\kappa = 2$, it is the usual descent lemma. ■

Proof of Proposition 9. We bound the gap between the best evaluated proximal objective and the global model minimum. Put $h(u) = A\|u - c\|_s^2/2$ and $F = f + h$. Query at c and let m_0 be its affine tangent. For $j \geq 0$, solve

$$x_j \in \operatorname*{argmin}_{u \in \mathcal{Q}} \{m_j(u) + h(u)\}, \qquad \ell_j = m_j(x_j) + h(x_j),$$

query at x_j, and retain $V_j = \min_{0 \leq i \leq j} F(x_i)$ with a corresponding witness. Stop when $E_j := V_j - \ell_j \leq \tau$; otherwise add the tangent at x_j to form m_{j+1}.

Strong convexity and $m_{j+1} \geq m_j$ give

$$\frac{\mu}{2}\|x_{j+1} - x_j\|_s^2 \leq \ell_{j+1} - \ell_j =: b_j.$$

The new model contains the tangent at x_j, so

$$\begin{aligned} E_{j+1} &\leq F(x_{j+1}) - \ell_{j+1} \\ &\overset{①}{\leq} \frac{H_\omega}{2}\|x_{j+1} - x_j\|_s^2 + \omega \\ &\leq \frac{H_\omega}{\mu} b_j + \omega. \end{aligned}$$

Here ① uses the assumed quadratic upper model. Since $V_{j+1} \leq V_j$, we have $b_j \leq E_j - E_{j+1}$. With $a = H_\omega/\mu$, this proves

$$E_{j+1} \leq \frac{a}{1+a} E_j + \frac{\omega}{1+a}, \qquad E_j \leq \omega + \left(\frac{a}{1+a}\right)^j E_0.$$

The initial tangent at c gives $E_0 = f(x_0) - m_0(x_0) \leq 2H_\omega R^2 + \omega$. Thus $E_j \leq \tau$ within (13). The retained model and witness satisfy (10) and (11). All minima are over the original $\mathcal{Q}$, so the stopping certificate is global. ■

For $1 < s \leq 2$, $\|\cdot\|_s^2/2$ is σ_s-strongly convex with $\sigma_s = s - 1$. Taking $\omega = \tau/8$ and $A \geq C_s H_\omega$ makes $H_\omega/[A(s-1)] = O_s(1)$, as required for logarithmic oracle cost.

Appendix D. Envelope line search and feasible coupling

Proof of Lemma 10. Fix one certificate (x, y, m) at a queried center c and define its lower envelope

$$\underline{M}(v) = \min_{u \in \mathcal{Q}} \left\{ m(u) + \frac{A}{2}\|u - v\|_s^2 \right\}.$$

Because $m \leq f$, $\underline{M}(v) \leq M_A(v)$ for all v. At c, the bundle gap gives

$$M_A(c) \leq f(y) + \frac{A}{2}\|y - c\|_s^2 \leq \underline{M}(c) + \tau.$$

Danskin's theorem (Dan67) gives $g = AJ_s(c - x) \in \partial \underline{M}(c)$. Combining the three facts yields the certified τ-subgradient inequality

$$M_A(v) \geq M_A(c) + \langle g, v - c \rangle - \tau, \qquad v \in \mathcal{Q}. \tag{22}$$

Write $d = z - y^{\mathrm{old}}$, $c_\lambda = y^{\mathrm{old}} + \lambda d$, and $\psi_\lambda = \langle g_\lambda, d\rangle$. If $\psi_0 \geq 0$, return c_0. If $\psi_0 < 0$, query 1; when $\psi_1 \leq 0$, apply (22) at c_1 with $v = c_0$ to get $M_A(c_1) \leq M_A(c_0) + \tau$, and return c_1.

It remains to treat $\psi_0 < 0 < \psi_1$. Bisect while maintaining endpoints λ_-, λ_+ with $\psi_{\lambda_-} \leq 0 \leq \psi_{\lambda_+}$, and stop when $\lambda_+ - \lambda_- \leq \eta/G_{\mathrm{ls}}$, where $G_{\mathrm{ls}} = 4AR^2$. The envelope is G_{ls}-Lipschitz on this segment: comparison with the prox minimizer at the other center gives

$$\left|M_A(c') - M_A(c)\right| \leq 2AR\left\|c' - c\right\|_s.$$

The τ-subgradient inequality at the left endpoint, evaluated at c_0, gives $M_A(c_{\lambda_-}) \leq M_A(c_0) + \tau$. Therefore

$$M_A(c_{\lambda_+}) \leq M_A(c_0) + \tau + \eta.$$

Finally, $\left\langle g_{\lambda_+}, z - c_{\lambda_+}\right\rangle = (1 - \lambda_+)\psi_{\lambda_+} \geq 0$. The initial bracket has length one, so the number of evaluations is $O(\log(2 + AR^2/\eta))$. ■

Proof of Lemma 11. Let

$$d_x = \|c - x\|_s, \qquad d_y = \|c - y\|_s, \qquad G = \|g\|_{s^*} = Ad_x.$$

Optimality in (10) supplies $g_m \in \partial m(x)$ with $g - g_m \in N_{\mathcal{Q}}(x)$. Hence, for every $v \in \mathcal{Q}$,

$$m(v) \geq m(x) + \langle g_m, v - x\rangle \geq m(x) + \langle g, v - x\rangle = a(v).$$

Thus $a \leq m \leq f$.

The model prox objective is $A\sigma_s$-strongly convex. Since $m(y) \leq f(y)$, the bundle gap implies

$$\frac{A\sigma_s}{2}\|y - x\|_s^2 \leq \tau.$$

The triangle inequality therefore gives

$$d_y^2 \geq \frac{1}{2}d_x^2 - \frac{2\tau}{A\sigma_s}. \tag{23}$$

The line search gives $M_A(c) \leq M_A(y^{\mathrm{old}}) + \eta + \tau$ and $\langle g, z - c\rangle \geq 0$. Since $M_A(y^{\mathrm{old}}) \leq f(y^{\mathrm{old}}) \leq U$, the upper certificate obeys

$$\begin{aligned} f(y) &\leq M_A(c) + \tau - \frac{A}{2}d_y^2 \\ &\leq U + \eta + 2\tau - \frac{A}{2}d_y^2 \\ &\overset{①}{\leq} U - \frac{G^2}{4A} + C_s(\eta + \tau). \end{aligned}$$

Here ① uses (23). Taking the minimum with the old upper value proves (14).

For the cut, the bundle gap also gives

$$m(x) - f(y) \geq \frac{A}{2}(d_y^2 - d_x^2) - \tau.$$

Consequently,

$$\begin{aligned} a(z) - (f(y) - \Gamma) &\geq \langle g, z - x \rangle + \frac{A}{2}(d_y^2 - d_x^2) + \Gamma - \tau \\ &= \langle g, z - c \rangle + \frac{A}{2}(d_x^2 + d_y^2) + \Gamma - \tau \\ &\geq \Gamma + \frac{G^2}{2A} - C_s(\eta + \tau). \end{aligned}$$

Replacing $f(y)$ by the no-larger U^+ only deepens the cut. This proves (15). ■

Appendix E. Certified numerical solves and cuts

Proposition 13 (Certified numerical coupling) [↓] *Let* $1 < s \leq 2$, $\mathcal{Q} = RB_p^d \subseteq RB_s^d$, *and* $A \geq C_s H_{\tau/8}$. *Fix* $\eta, \tau > 0$ *and* $0 < \nu \leq \min\{\tau, AR^2\}$. *Certified approximate model solves and the sign search of Lemma 10 produce a feasible witness* y^+, *certificate* $U^+ \geq f(y^+)$, *and a valid affine cut* $\widehat{a} \leq f$ *such that*

$$\begin{aligned} U^+ &\leq U - \frac{G^2}{8A} + C_s(\eta + \tau + \nu), \\ \widehat{a}(z) - (U^+ - \Gamma) &\geq \Gamma + \frac{G^2}{4A} - C_s(\eta + \tau + \nu), \end{aligned}$$

where G *is the cut-normal norm in* ℓ_{s^*}. *The number of first-order calls is* $O_s(\log(2 + AR^2/\tau)\log(2 + AR^2/\eta))$. *For outer tests, compute a certified interval* $[L_m, V_m]$ *of width* $\Gamma/16$ *for* $\min_{\mathcal{Q}} m$, *and terminate the phase if* $L_m \geq U - 5\Gamma/4$. *Keeping the largest certified lower bound preserves a gap contraction of at most* $7/8$ *per phase.*

Proof of Proposition 13. Put $h(u) = A\|u - c\|_s^2/2$ and $\mu = A(s-1)$. For each tangent model m_j, compute a feasible $\widetilde{x}_j$ and lower bound l_j with $m_j(\widetilde{x}_j) + h(\widetilde{x}_j) \leq l_j + \zeta$. Write $\ell_j = \min_{\mathcal{Q}}(m_j + h)$, so $l_j \leq \ell_j$, retain $V_j = \min_{i \leq j}(f + h)(\widetilde{x}_i)$, and stop when $V_j - l_j \leq \tau$. Otherwise add the tangent at $\widetilde{x}_j$. For the exact model minimizer x_j, strong convexity gives

$$\|\widetilde{x}_j - x_j\|_s^2 \leq 2\zeta/\mu, \qquad \|\widetilde{x}_{j+1} - \widetilde{x}_j\|_s^2 \leq 4(\ell_{j+1} - \ell_j + 2\zeta)/\mu.$$

The second bound compares both points to x_j. Set $E_j = V_j - \ell_j$, $\omega = \tau/8$, and $a = 2H_\omega/\mu$. The inexact gradient trick at the last tangent and $V_{j+1} \leq V_j$ give

$$E_{j+1} \leq a(\ell_{j+1} - \ell_j) + (2a+1)\zeta + \omega \leq a(E_j - E_{j+1}) + (2a+1)\zeta + \omega.$$

For $\zeta \leq \tau/[16(1+a)]$, this contracts to at most $\tau/4$. Since $E_0 \leq 2H_\omega R^2 + \omega + \zeta$ and $V_j - l_j \leq E_j + \zeta$, it takes $O_s(\log(2 + AR^2/\tau))$ calls to certify (11) relative to the exact model minimum; the point computed is $\widetilde{x}$, while the exact minimizer x is used only in the analysis below.

At termination, let x be the exact model minimizer and compute $g = AJ_s(c - \widetilde{x})$ in the stated real-arithmetic model. The duality map satisfies $\|J_s(w) - J_s(w')\|_{s^*} \leq$

$C_s R^{2-s}\|w-w'\|_s^{s-1}$ on $2RB_s^d$; this follows from its coordinate formula and the $(s-1)$-Hölder continuity of $t \mapsto |t|^{s-2}t$. For $g^\circ = AJ_s(c-x)$, choose

$$\zeta \leq \min\left\{\frac{\tau}{16(1+a)},\ \nu,\quad \frac{\mu}{2}\left(\frac{\nu}{4C_s A R^{3-s}}\right)^{2/(s-1)}\right\}, \qquad E_g = \frac{\nu}{4R}.$$

Then $\|g-g^\circ\|_{s^*} \leq E_g$. The exact cut $a^\circ(v) = m(x) + \langle g^\circ, v-x\rangle$ is valid, while the model gap and convexity of h give $0 \leq m(\widetilde{x}) - m(x) - \langle g^\circ, \widetilde{x}-x\rangle \leq \zeta$. Thus the computable cut $\widehat{a}(v) = m(\widetilde{x}) + \langle g, v-\widetilde{x}\rangle - \zeta - 2RE_g$ satisfies

$$a^\circ(v) - \zeta - 4RE_g \leq \widehat{a}(v) \leq a^\circ(v) \leq f(v) \qquad (v \in \mathcal{Q}). \tag{24}$$

By (22), g is a $(\tau + 2RE_g)$-subgradient of M_A on $\mathcal{Q}$. The proof of Lemma 10 therefore applies with that error, giving $M_A(c) \leq M_A(y^{\text{old}}) + \eta + \tau + 2RE_g$ and $\langle g, z-c\rangle \geq 0$ in $O(\log(2+AR^2/\eta))$ trials. Consequently $\langle g^\circ, z-c\rangle \geq -2RE_g$. Apply the proof of Lemma 11 to g°, then use (24) and $\|g^\circ\|_{s^*}^2 \geq G^2/2 - E_g^2$. Since $E_g^2/A \leq \nu/16$, this gives the stated descent and cut bounds. Retain the witness corresponding to $U^+ = \min\{U, f(y)\}$.

For the outer test, let $V_m = m(v_m)$ with $v_m \in \mathcal{Q}$. If $L_m \geq U - 5\Gamma/4$, the new gap is at most $5\Gamma/4$, which is $5/8$ of the starting gap. Otherwise $m(v_m) < U - 19\Gamma/16$, leaving slack $3\Gamma/16$ in every model cut of $K = \{m \leq U - \Gamma\}$. Upper progress still contracts the gap by $7/8$. To apply the companion's implementation, normalize $R = L = 1$ and let B be the largest norm of the stored cut normals. Here $B \leq 2\max A$. With $\theta = \min\{1/2, \Gamma/[32(1+B)]\}$,

$$(1-\theta)v_m + \frac{\theta}{2\sqrt{d}}B_2^d \subseteq K.$$

This follows from the cut slack and $\|v_m\|_p \leq 1$. The phase parameters give inverse-polynomial radii and objective tolerances for fixed κ, s. The transfer to the companion's sample problems is proved in Proposition 14. Inner tangent slopes are bounded by $\|\nabla f(0)\|_{s^*} + 1$ in normalized units, giving the logarithmic dependence stated in Section 9. ■

Proposition 14 (Feasible selector implementation) *Let $1 \leq p < \min\{q,2\}$ and $s = \min\{q,2\}$. The numerical version of Algorithm 1, initialized by Lemma 16, admits feasible randomized selectors with ℓ_s-movement $\widetilde{O}_{p,s}(RN_\Gamma^{1-\rho_{p,q}})$ in phase $k \geq 0$, conditionally on its initial history, with probability at least $1-\delta_k$. Their implementation uses no additional first-order calls. For $0 < \varepsilon \leq LR^\kappa$, allocating $\delta_k = \delta/[2(k+1)^2]$ gives simultaneous success with probability at least $1-\delta$ and the additional-work bound in Section 9 on every run. On a sampling cutoff, return the retained feasible upper witness.*

Proof Normalize $R = L = 1$ by replacing f with $f(R\,\cdot)/(LR^\kappa)$. Fix a phase and write $N = N_\Gamma$, $\rho = \rho_{p,q}$. Apply MG26, Proposition 11 with their $(q,T) = (s,N)$, sampling and solver errors $E = E_{\text{sol}} = N^{-\rho}/4$, and failure probability $\delta_k/2$. Use exactly their sample count (17), objective tolerance (32), and feasible empirical average. These give exact feasibility and, by their path-error estimate (33) and Theorem 9, the stated movement bound. Their conditional concentration argument already covers adaptive bodies and fresh samples; shorter phases can be padded by their last body. Our coupling is evaluated at the computed feasible center itself, so no additional cut error is needed.

We check only the geometric input to the companion's Appendix E. The preceding proof supplies an interior ball of radius $\theta/(2\sqrt{d})$ in each queried K_t, where $\theta = \min\{1/2, \Gamma/[32(1+B)]\}$ and the stored cut normals satisfy $B \le 2\max A$. Shrinking the head-tail decomposition of v_m by $1-\theta$, exactly as there, gives an interior ball in its lift of radius $\theta\min\{1, R_1, R_s\}/(4\sqrt{d})$, with $R_1 = N^{1-1/p}$ and $R_s = N^{1/s-1/p}$; an outer radius is at most $4N$. Norm constraints and stored cuts supply the same separation oracles without queries to f. Hence the certified feasible solves of Appendix E apply to our aggregate cuts. The fixed powers are permitted directly in our model, without rational-exponent approximation.

Also use the Gaussian cutoff in Appendix E with planned draw budget $(N+1)M$, where M is the sample count above, and phase failure parameter δ_k. It costs probability at most $\delta_k/2$ and bounds the work of every sample solve. Since $\Gamma > \varepsilon/2$ in active phases, our planned horizons and A_Γ make all inverse radii, inverse tolerances, and sample counts polynomial in the parameters of Section 9. The query guard bounds the number of stored cuts and center calls by J, and phase contraction bounds the number of phases. Model and tangent solves use the same certified procedure on $\mathcal{Q}$, with the logarithmic affine-term dependence established above. Thus work is bounded on every run, and $\sum_{k\ge 0}\delta_k \le \delta$ gives simultaneous success. Rescaling restores R, L. ■

Appendix F. Phase tuning, initialization, and final assembly

Lemma 15 (Quadratic phase and fixed point) *Consider N nonterminal transitions in one phase with fixed $A, \Gamma > 0$, nonempty bodies $K_0, \ldots, K_N \subseteq RB_p^d$, and $U_0 - U_N < \Gamma/4$. Suppose the selector satisfies $\sum_{t=0}^{N-1}\|z_{t+1}-z_t\|_s = \widetilde{O}(RN^{1-\rho})$. Let G_t be the aggregate residual norms. Assume the cumulative contribution of the line-search, bundle, and numerical errors is a sufficiently small multiple of Γ, and each cut error is at most $\Gamma/2$. Then*

$$N^{1+2\rho} = \widetilde{O}_s\left(\frac{AR^2}{\Gamma}\right).$$

At the planned horizon $N = N_\Gamma$, with the parameters in (17), this is equivalent, up to constants and logs, to (18).

Proof Before upper progress, sum (14) to obtain $\sum_{t<N} G_t^2 \le CA\Gamma$. The weaker constants in Proposition 13 give the same estimate. The cut depth from (15), after absorbing its error, and the unpacked movement bound $\sum_{t<N}\|z_{t+1}-z_t\|_s \le CR\Lambda(d,N)N^{1-\rho}$ give

$$\frac{\Gamma}{2}\sum_{t<N}\frac{1}{G_t} \le R\Lambda(d,N)N^{1-\rho}.$$

Hölder with exponents 3 and 3/2 gives

$$N \le \left(\sum_{t<N} G_t^2\right)^{1/3}\left(\sum_{t<N}\frac{1}{G_t}\right)^{2/3}.$$

Substitution proves the quadratic phase inequality.

To test the planned horizon, set $N = N_\Gamma$ and use $A \asymp L^{2/\kappa}(N/\Gamma)^{(2-\kappa)/\kappa}$. The phase inequality becomes

$$N^{1+2\rho-(2-\kappa)/\kappa} \lesssim L^{2/\kappa} R^2 \Gamma^{-2/\kappa}.$$

Raising to $\kappa/2$ yields

$$N^{\kappa-1+\kappa\rho} \lesssim \frac{LR^\kappa}{\Gamma},$$

with the movement logarithms restored as in (18). ■

To specify the horizon in Algorithm 1, fix $\Lambda(d,N) = 1 + \log(2d+2)\log(2N)$. The selector estimates in MG26, including their feasible approximations on the success event, give movement at most $C_{p,s} R\Lambda(d,N)N^{1-\rho}$ over N transitions. Fix the constants in (17) as in the proof below, and a constant $C_{\kappa,p,s}$ large enough for the preceding phase bound. The computable horizon rule is

$$N_\Gamma = \min\left\{2^j : j \in \mathbb{Z}_{\geq 0}, \quad (2^j)^{\kappa-1+\kappa\rho} > C_{\kappa,p,s}\frac{LR^\kappa}{\Gamma}\Lambda(d,2^j)^\kappa\right\}. \tag{25}$$

Thus N_Γ counts transitions, and a phase uses at most $N_\Gamma + 1$ selector points.

Lemma 16 (Affine-invariant initialization) *One certified envelope call at $c = 0$ with*

$$A_0 \asymp LR^{\kappa-2}, \qquad \tau_0 \asymp LR^\kappa$$

produces a feasible upper witness and a valid affine lower cut whose certified gap over $\mathcal{Q}$ is $O_{\kappa,s}(LR^\kappa)$, independently of any affine term in f.

Proof Let (x_0, y_0, m_0) be the bundle certificate and $g_0 = A_0 J_s(-x_0)$. The bundle gap and $\|x_0\|_s \leq R$ give

$$f(y_0) - m_0(x_0) \leq \frac{A_0 R^2}{2} + \tau_0.$$

Moreover $\|g_0\|_{s^*} \leq A_0 R$. The aggregate cut $a_0(v) = m_0(x_0) + \langle g_0, v - x_0\rangle$ therefore satisfies

$$f(y_0) - \min_{v\in\mathcal{Q}} a_0(v) \leq \frac{5}{2}A_0R^2 + \tau_0 = O(LR^\kappa).$$

The bound on the initial tangent gap in Proposition 9 is independent of an affine term in f, so the first-order call bound is affine-invariant as well. With numerical solves, the buffered cut and a lower-bound tolerance $O(LR^\kappa)$ add only $O(LR^\kappa)$ to this initial gap. ■

Proof of Theorem 12. For randomized selectors, assign phase $k \geq 0$ failure probability $\delta_k = \delta/[2(k+1)^2]$, using Proposition 14, including its Gaussian cutoff. Condition on simultaneous selector success. Fresh samples make each phase guarantee valid conditionally on the past, so this event has probability at least $1 - \sum_{k\geq 0}\delta_k \geq 1 - \delta$. Initialize with Lemma 16. At a phase of scale Γ, set τ_Γ, η_Γ to a sufficiently small constant times Γ/N_Γ and choose A_Γ as in (17). Then the cumulative errors in (14) and (15) are small enough for Lemma 15. Choose N_Γ by (25). If a phase contained N_Γ nonterminal transitions, Lemma 15 would give the opposite inequality, a contradiction. Thus the phase terminates earlier; one final empty-body step changes the count by at most one.

Take $\omega = \tau_\Gamma/8$ and a sufficiently large $A_\Gamma \asymp_{\kappa,s} H_\omega$. By Propositions 9 and 13, every envelope call is logarithmic in $2+AR^2/\tau_\Gamma$. For numerical solves, take $\nu_\Gamma = \frac{1}{2}\min\{\tau_\Gamma, A_\Gamma R^2\}$ and use the early lower-bound test in Proposition 13; its contraction factor $5/8$ is also at most $7/8$. The line search is logarithmic as well. Put $\alpha = \kappa - 1 + \kappa\rho > 0$. To reach error ε, every active phase has $\Gamma > \varepsilon/2$. Reading the gaps backwards from the final active phase bounds the oracle costs by a geometric series with ratio $(7/8)^{1/\alpha}$, up to logarithms. Including initialization, choose the deterministic first-order budget

$$\begin{aligned} J(\varepsilon,\delta) &= \left\lceil C_{\kappa,p,q}\left(1+(LR^\kappa/\varepsilon)^{1/\alpha}\right)[1+\log(2d)+\log(2+LR^\kappa/\varepsilon)+\log(2/\delta)]^8\right\rceil, \\ \alpha &= \kappa-1+\kappa\rho. \end{aligned} \tag{26}$$

Here the fixed constant covers all the preceding bounds. The logarithmic power 8 suffices: $\rho > 1/2$ gives $\kappa/\alpha \le 2$, so the horizon rule costs at most four logarithmic powers, the inner solves and line search add two, and geometric phase summation preserves this order. In particular, $J(\varepsilon,\delta) = \widetilde{O}_{\kappa,p,q}(1+(LR^\kappa/\varepsilon)^{1/\alpha})$. On simultaneous success, the certified gap reaches ε within this budget, and each phase uses at most $N_\Gamma + 1$ selector calls. On every run, the guards in Algorithm 1 enforce these budgets and retain a feasible output. Only within-phase edges are charged to movement; restarting the selector introduces no charged edge between phases. For a prescribed first-order budget $T \ge 1$, set

$$\varepsilon_T = \min\left\{LR^\kappa 2^{-j} : j \in \mathbb{Z}_{\ge 0}, \quad J(LR^\kappa 2^{-j},\delta) \le T\right\}, \qquad J = J(\varepsilon_T,\delta). \tag{27}$$

If this set is nonempty, run Algorithm 1 with accuracy ε_T. Minimality gives $J(\varepsilon_T/2,\delta) > T$, while $J(\varepsilon_T,\delta) \le T$ bounds $\log(2+LR^\kappa/\varepsilon_T)$ by $O_{\kappa,p,q}(\log(2+T))$. Hence $\varepsilon_T = \widetilde{O}_{\kappa,p,q}(LR^\kappa/T^\alpha)$. If the set is empty, use one query at 0 and minimize its tangent over $\mathcal{Q}$. This gives error at most LR^κ/κ; since then $T < J(LR^\kappa,\delta)$, the same rate holds after increasing its logarithmic factor. Every retained upper witness is feasible and matches its upper certificate. ∎

Proof of Theorem 1. If $p < q \le 2$, use Theorem 12 with $s = q$. If $p < 2 \le q$, use the same theorem with $s = 2$ and the $p/2$ movement selector. These two cases cover $p < \min\{q,2\}$ and give (4). For $p = q \le 2$, the classical universal accelerated method (Nes15) in the p-geometry gives error $\widetilde{O}_{\kappa,p}(LR^\kappa/T^{3\kappa/2-1})$, including the logarithmic loss at $p = q = 1$. For this endpoint, take $r = 1 + 1/\log(2d+2)$ and the prox function $u \mapsto e^2\|u\|_r^2/[2(r-1)]$, which is 1-strongly convex in ℓ_1 and has range $O(R^2\log(2d))$ on $\mathcal{Q}$. Since $\kappa(1+1/p-(1/p-1/2)_+)-1 = 3\kappa/2-1$ for $p \le 2$, this is also (4). ∎

References

[ABJ+24] Deeksha Adil, Brian Bullins, Arun Jambulapati, and Aaron Sidford. *Convex Optimization with p-Norm Oracles*. 2024 (cit. on p. 5).

[AO14] Zeyuan Allen-Zhu and Lorenzo Orecchia. *Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent*. 2014 (cit. on p. 5).

[CGM24] Juan Pablo Contreras, Cristóbal Guzmán, and David Martínez-Rubio. *Non-Euclidean High-Order Smooth Convex Optimization*. 2024 (cit. on p. 5).

[CHJ+22] Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. *Optimal and Adaptive Monteiro–Svaiter Acceleration*. 2022 (cit. on p. 5).

[Dan67] John M. Danskin. *The Theory of Max-Min and Its Application to Weapons Allocation Problems*. Vol. 5. Econometrics and Operations Research. Berlin, Heidelberg: Springer, 1967 (cit. on p. 15).

[DGN14] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. In: *Mathematical Programming* 146.1 (2014), pp. 37–75 (cit. on p. 9).

[GN15] Cristóbal Guzmán and Arkadi Nemirovski. On Lower Complexity Bounds for Large-Scale Smooth Convex Optimization. In: *Journal of Complexity* 31.1 (2015), pp. 1–14 (cit. on pp. 3, 11).

[Guz15] Cristóbal Guzmán. Open Problem: The Oracle Complexity of Smooth Convex Optimization in Nonstandard Settings. In: *Proceedings of the 28th Conference on Learning Theory*. Vol. 40. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1761–1763 (cit. on pp. 1, 3, 11).

[Lev17] Kfir Y. Levy. Online to Offline Conversions, Universality and Adaptive Minibatch Sizes. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 1613–1622 (cit. on p. 5).

[LNN95] Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New Variants of Bundle Methods. In: *Mathematical Programming* 69 (July 1995), pp. 111–147 (cit. on p. 3).

[MG26] David Martínez-Rubio and Cristóbal Guzmán. *Stable Movement for Nondual Lipschitz Convex Optimization: Efficiency and Nearly Optimal Oracle Rates*. 2026 (cit. on pp. 1–4, 7, 12, 18, 20).

[Nes15] Yurii Nesterov. Universal Gradient Methods for Convex Optimization Problems. In: *Mathematical Programming* 152.1–2 (2015), pp. 381–404 (cit. on pp. 3, 21).

[NN85] Arkadii S. Nemirovskii and Yurii E. Nesterov. Optimal Methods of Smooth Convex Optimization. In: *USSR Computational Mathematics and Mathematical Physics* 25.2 (1985), pp. 21–30 (cit. on p. 3).

[Ora23] Francesco Orabona. Normalized gradients for all. In: *arXiv preprint arXiv:2308.05621* (2023) (cit. on p. 5).

[Ouy26] Yuyuan Ouyang. *An $O(1/T^3)$ Algorithm for Minimizing Convex Quadratic Functions over the L_1 Ball*. 2026 (cit. on p. 3).