# Dimension-Free Gradient-Norm Minimization in $\ell_p$ and Schatten-$p$ Spaces

**Abstract**

We establish dimension-free gradient-query bounds without multiplicative logarithmic loss for convex optimization in $\ell_p$ and Schatten-$p$ spaces. For fixed $2 \le p < \infty$, $1 < \kappa \le 2$, and $q = p/(p-1)$, suppose the gradient is $(\kappa-1)$-Hölder continuous from $\ell_p$ to $\ell_q$ with constant $L$, and the initial point lies within distance $R$ of a minimizer. In the deterministic exact-real oracle model, our accumulative shifted-power regularization method finds $x$ with $\|\nabla f(x)\|_q \le \varepsilon$ using $O_{p,\kappa}((LR^{\kappa-1}/\varepsilon)^{p/(\kappa p+\kappa-p)})$ gradient queries when $0 < \varepsilon < LR^{\kappa-1}$, independently of dimension. For Lipschitz gradients, the exponent $p/(p+2)$ improves the earlier $2(p-1)/(p+2)$ upper-bound exponent for $p > 2$, addressing the complexity gap identified by Diakonikolas and Guzmán [6]. Contemporaneous work by Pelleriti et al. [18] obtains the same vector rate through accumulated regularization. Our framework also yields a dimension-free Schatten-$p$ bound for matrix objectives and an accuracy-oblivious checkpoint scheme that retains the target-accuracy rate. The analysis combines minimizer transport, localization, and a summable path bound to convert approximate composite minimization into a final-gradient guarantee. The matrix extension permits exact multicenter proximal computations between oracle calls. We further give a matching minimax comparison under an explicit simultaneous hard-family assumption.

**Note.** This paper was generated entirely by AI, including MiMo, using an automated research pipeline developed by Chenghua Liu and Hanyu Li.

## 1 Introduction

Small gradient norm is both a stationarity criterion and a checkable stopping certificate, but fast objective decrease does not automatically give the desired last-gradient rate. This section introduces the exact-real oracle model, states the main vector and matrix guarantees, and separates the vector result shared with contemporaneous work from the additional Schatten and oracle-model analyses presented here.

In Euclidean geometry, gradient-norm minimization has motivated regularization, optimized-gradient, potential-function, and duality-based methods [7, 12, 13, 15, 17]. In a normed space the intrinsic certificate is the dual gradient norm, and the geometry can introduce dimension dependence. In particular, a quadratic prox that is dimension-free in a Hilbert space can lose a polynomial factor in dimension in $\ell_p$ when $p > 2$.

### 1.1 Problem, oracle model, and literature

Fix an integer $d \ge 1$, a point $x_0 \in \mathbb{R}^d$, $2 \le p < \infty$, $1 < \kappa \le 2$, $q = p/(p-1)$, and $L, R \ge 0$. We consider differentiable convex functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\|\nabla f(x) - \nabla f(y)\|_q \le L \|x - y\|_p^{\kappa-1} \qquad (x, y \in \mathbb{R}^d), \tag{1.1}$$

and assume that $f$ has a minimizer $x^*$ with

$$\|x_0 - x^*\|_p \le R. \tag{1.2}$$

We write $B_p(c,r) = \{x \in \mathbb{R}^d : \|x-c\|_p \le r\}$ for the closed $\ell_p$ ball. The case $\kappa = 2$ is Lipschitz-gradient smoothness.

**Definition 1.1** (Deterministic exact-real query model)**.** The method knows $x_0, L, R, p, \kappa$ and, in the target-accuracy version, $\varepsilon$. After any finite exact transcript, its next query, stopping decision, and output are deterministic single-valued functions of that transcript and the known parameters. Each oracle invocation, including a repeated query, counts once; the output need not be a previously queried point. Exact-real operations involving only the known parameters and the transcript are uncharged. We make no finite-precision, arithmetic, or bit-complexity claim.

The upper bound uses only a gradient oracle. For comparison with lower bounds, a local value-gradient query at $x$ returns $f(x)$ and $\nabla f(x)$; the cited lower bound applies to a still broader class of local oracles. The internal composite maps used by the method are specified in Section 5. Throughout, a generic $C_{p,\kappa}$ may change from line to line and depends only on $p, \kappa$; constants used as fixed budgets receive separate superscripts.

At $\kappa = 2$, mirror duality gives a dimension-free $LR/N^2$ gradient rate for $1 < p \le 2$, but its quadratic geometry does not extend dimension-freely to $p > 2$ [14, Corollary 3 and Section 4.1]. A previous general upper bound for $p > 2$ is

$$\widetilde{O}_p\left(\left(\frac{LR}{\varepsilon}\right)^{2(p-1)/(p+2)}\right) \tag{1.3}$$

[6, Theorem 3], whereas its local-oracle lower bound has exponent

$$\frac{p}{\kappa p + \kappa - p} \tag{1.4}$$

in $LR^{\kappa-1}/\varepsilon$, subject to explicit small-accuracy and large-dimension conditions [6, Corollary 2]. Contemporaneous work by Pelleriti et al. [18, Theorem 5.3] reports this same vector-space exponent and a closely related accumulated-regularization construction. We therefore do not claim sole priority for the core vector rate or that mechanism. The version-specific one-stage comparison is recorded in Section 7.2.

For a fixed $x_0$, let $\mathcal{F}^{\kappa}_{p,d}(L, R; x_0)$ be the class of functions satisfying (1.1) and $\arg\min f \cap B_p(x_0, R) \neq \varnothing$. Define $\mathrm{Comp}^{\mathrm{det}}_{p,d,\kappa}(L, R, \varepsilon)$ as the infimum, over deterministic local value-gradient methods in Definition 1.1, of the worst-case query count needed to return $x$ with $\|\nabla f(x)\|_q \le \varepsilon$, where the worst case ranges over this class. Translation invariance makes the value independent of the fixed $x_0$.

## 1.2 Main results and scope

The following theorem is the central vector-space guarantee. It states the target-accuracy bound in the exact-real model; the matrix result below uses the same outer geometric interface.

**Theorem 1.2** (Exact-real gradient-query upper bound)**.** *Let $d \ge 1$, $2 \le p < \infty$, $1 < \kappa \le 2$, $q = p/(p-1)$, $L, R \ge 0$, $\varepsilon > 0$, and $x_0 \in \mathbb{R}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, satisfy* (1.1)*, and have a minimizer satisfying* (1.2)*. In the model of Definition 1.1, the accumulative shifted-power regularization method in Section 3 returns $x$ with $\|\nabla f(x)\|_q \le \varepsilon$ after at most*

$$C_{p,\kappa}\left[1 + \left(\frac{LR^{\kappa-1}}{\varepsilon}\right)^{\beta}\right] \quad \textit{gradient queries}, \qquad \beta = \frac{p}{\kappa p + \kappa - p}. \tag{1.5}$$

*Here $C_{p,\kappa} < \infty$ is independent of $d, L, R, \varepsilon$. If $\varepsilon \geq LR^{\kappa-1}$, the method returns $x_0$ without a query. In the nontrivial regime $0 < \varepsilon < LR^{\kappa-1}$, the additive 1 can be absorbed into $C_{p,\kappa}(LR^{\kappa-1}/\varepsilon)^{\beta}$.*

At $(p, \kappa) = (2, 2)$, the theorem gives the $O(\sqrt{LR/\varepsilon})$ Hilbert-space query rate. For $p > 2$ and $\kappa = 2$, it attains exponent $p/(p+2)$, the exponent highlighted by the earlier lower bound and also obtained by Pelleriti et al. [18].

The lower comparison requires a simultaneous hard-family interface that is not established by the individual cited statements. We therefore formulate that interface as assumption 6.1 and state the minimax consequence conditionally; this qualification does not affect Theorem 1.2.

**Corollary 1.3** (Conditional minimax comparison)**.** *Assume the simultaneous normalized hard-family realization in assumption 6.1. Fix $2 \leq p < \infty$, $1 < \kappa \leq 2$, and $L, R > 0$. There are $c_{p,\kappa}, C^{\mathrm{lb}}_{p,\kappa} > 0$, depending only on $p, \kappa$. For $\beta = p/(\kappa p + \kappa - p)$, every integer $d \geq 1$ and $\varepsilon > 0$ with $0 < \varepsilon \leq c_{p,\kappa} LR^{\kappa-1}$, $\log d \geq p$, and*

$$d \geq C^{\mathrm{lb}}_{p,\kappa} \left( \frac{LR^{\kappa-1}}{\varepsilon} \right)^{\beta}$$

*obey the matching relation*

$$\mathrm{Comp}^{\mathrm{det}}_{p,d,\kappa}(L, R, \varepsilon) = \Theta_{p,\kappa}\left( \left( \frac{LR^{\kappa-1}}{\varepsilon} \right)^{\beta} \right).$$

*The upper-bound half is unconditional.*

The outer proof depends only on a norm-geometric power-atom interface. This yields the following matrix extension; its nonseparable prox is an exact internal map, not a finite-precision implementation.

**Corollary 1.4** (Schatten-$p$ upper bound)**.** *Let $m, n \geq 1$, $k = \min\{m, n\}$, $2 \leq p < \infty$, $1 < \kappa \leq 2$, $q = p/(p-1)$, $L, R \geq 0$, and $\varepsilon > 0$. Fix $X_0 \in \mathbb{R}^{m \times n}$. On $\mathbb{R}^{m \times n}$, define*

$$\|X\|_{\mathsf{S}_p} = \left( \sum_{j=1}^{k} \sigma_j(X)^p \right)^{1/p}$$

*and take gradients with respect to the trace pairing. Suppose that a convex differentiable $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ satisfies*

$$\|\nabla f(X) - \nabla f(Y)\|_{\mathsf{S}_q} \leq L \|X - Y\|^{\kappa-1}_{\mathsf{S}_p}$$

*and has a minimizer $X^*$ with $\|X_0 - X^*\|_{\mathsf{S}_p} \leq R$. In the matrix analogue of Definition 1.1, with $X_0, L, R, p, \kappa, \varepsilon$ known, there is a gradient-query method returning $X$ with $\|\nabla f(X)\|_{\mathsf{S}_q} \leq \varepsilon$ within the bound (1.5), independently of $m, n$. For $p > 2$, this statement treats the nonseparable multicenter prox as an exact uncharged map and makes no finite-precision, arithmetic, or bit-complexity claim.*

Under assumption 6.1, diagonal instances transfer the conditional vector lower bound to Schatten geometry; see Section 5.3. Sections 5 and 7 also give an accuracy-oblivious checkpoint wrapper, schedule variants, and finite-dimensional diagnostics.

**Contributions and proof mechanism.** The vector upper bound is presented with explicit exact-real accounting and a short transport–path–stationarity proof; no sole-priority claim is made in light of Pelleriti et al. [18]. The additional geometric result is the Schatten-$p$ extension together with its exact multicenter-prox qualification. We state the simultaneous hard-family premise explicitly and give the objective-to-gradient localization reduction conditional on it.

At scale $s$, the method approximately solves $F_s = f + H_s$, where $H_s$ contains every shifted $p$-power regularizer introduced so far. Comparing consecutive exact regularized minimizers keeps the new minimizer in the preceding target radius, and the computed iterate reduces its distance to that minimizer by a constant factor. Exact stationarity at $x_S^*$ and a path bound turn the historical regularizer gradients into a geometric series; Hölder continuity then transfers the estimate to $x_S$. The same base ratio controls $B_{s+1}/B_s$, so the power-dependent stage terms have ratio $r^\beta$. The additive stage count is absorbed by the main power term, and no multiplicative logarithmic loss appears in the final query rate.

## 2 Power geometry and the one-stage routine

This section supplies the two ingredients used at every stage. We first normalize a shifted $p$-power atom under the Jensen convention and then specialize the complementary-composite solver to the contraction required by the outer method.

We use the convention of Diakonikolas and Guzmán [6]: a convex function $h$ is $(\mu, p)$-*uniformly convex* with respect to $\|\cdot\|_p$ if, for every $x, y \in \mathbb{R}^d$ and $\theta \in [0,1]$,

$$h((1-\theta)x + \theta y) \le (1-\theta)h(x) + \theta h(y) - \frac{\mu}{p}\theta(1-\theta)\,\|x-y\|_p^p\,. \tag{2.1}$$

The Jensen inequality has the following subgradient consequence: for every $g \in \partial h(x)$,

$$h(y) \ge h(x) + \langle g, y-x\rangle + \frac{\mu}{p}\,\|y-x\|_p^p \qquad (x, y \in \mathbb{R}^d). \tag{2.2}$$

Indeed, apply the subgradient inequality at $x$ to $h((1-\theta)x+\theta y)$, combine it with (2.1), divide by $\theta$, and let $\theta \downarrow 0$. For differentiable $h$, take $g = \nabla h(x)$.

For $v \in \mathbb{R}^d$, define the $p$-duality map coordinatewise, with $\operatorname{sign}(0) = 0$, by

$$J_p(v)_j = \operatorname{sign}(v_j)\,|v_j|^{p-1}\,. \tag{2.3}$$

In particular, $J_2$ is the identity, including at zero.

The normalization below gives the atom unit $p$-uniform-convexity modulus and the dual-gradient growth used in the terminal estimate.

**Lemma 2.1** (Normalized shifted-power atom)**.** *Let $d \ge 1$, $2 \le p < \infty$, $q = p/(p-1)$, and $c \in \mathbb{R}^d$. Define*

$$a_p = 2^{p-2}, \qquad \omega_c(x) = \frac{a_p}{p}\,\|x-c\|_p^p\,. \tag{2.4}$$

*Then $\omega_c$ is $(1,p)$-uniformly convex in the convention* (2.1)*, and*

$$\nabla\omega_c(x) = a_p J_p(x-c), \qquad \|J_p(v)\|_q = \|v\|_p^{p-1}\,. \tag{2.5}$$

*Consequently, every finite sum $\sum_{i=1}^m b_i\omega_{c_i}$ with $b_i > 0$ is $(\sum_{i=1}^m b_i, p)$-uniformly convex.*

The proof below also checks that the factor $a_p$ is necessary under this convention; see Section 2.1.

The only algorithmic black box is the nonlogarithmic branch of the complementary-composite acceleration theorem of Diakonikolas and Guzmán [6, Theorem 2]. The next lemma records precisely the specialization used here. We include the substitution to make the exponents and the oracle interface transparent.

Fix a sufficiently large source constant $C^{\mathrm{cc}}_{p,\kappa}$ for which that theorem's specialized bound holds. For fixed problem parameters, the update rule, coefficients, and iteration budget are fixed before the run; the iterates remain deterministic functions of the observed gradient transcript. This is an oracle-complexity statement with a sufficiently large source constant; it does not provide a numerically calibrated finite-precision implementation.

**Lemma 2.2** (Complementary-composite contraction)**.** *Let $d \geq 1$, $L > 0$, $2 \leq p < \infty$, $1 < \kappa \leq 2$, $q = p/(p-1)$, and $c \in \mathbb{R}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable and satisfy* (1.1)*. Let $H$ be a known $(\mu, p)$-uniformly convex function with $\mu > 0$. Suppose that, for every $g \in \mathbb{R}^d$ and $A, m > 0$, the internal minimization of $\langle g, x\rangle + AH(x) + m\omega_c(x)$ is available. Let $z^*$ minimize $F = f + H$, and suppose $\|c - z^*\|_p \leq D$ for $D > 0$. Put*

$$\Delta = \kappa p + \kappa - p, \qquad \beta = \frac{p}{\Delta}. \tag{2.6}$$

*For $0 < \delta \leq D$, Generalized AGD+ with its coefficients and budget fixed from these parameters returns a transcript-dependent point $z$ with $\|z - z^*\|_p \leq \delta$ using at most*

$$C^{\mathrm{cc}}_{p,\kappa}\left[1 + \left(\frac{L}{\mu\delta^{p-\kappa}}\right)^{\beta}\left(\frac{D}{\delta}\right)^{\kappa\beta}\right] \tag{2.7}$$

*gradient queries to $f$. The same fixed budget guarantees*

$$F(z) - F(z^*) \leq \frac{\mu}{p}\delta^p. \tag{2.8}$$

The proof below gives the complete source-parameter map and the endpoint convention. The lemma is invoked by the outer method only under (3.1), which forces $L, R > 0$.

## 2.1 Proofs of the one-stage ingredients

We verify the power-atom normalization and then give the complete source-parameter substitution for the contraction guarantee.

*Proof of Lemma 2.1.* A weighted form of Clarkson's inequality [3] states that, for real $u, v$ and $\theta \in [0, 1]$,

$$|(1-\theta)u + \theta v|^p \leq (1-\theta)\,|u|^p + \theta\,|v|^p - 2^{2-p}\theta(1-\theta)\,|u - v|^p\,. \tag{2.9}$$

Apply this inequality coordinatewise to $x - c$ and $y - c$, sum, divide by $p$, and multiply by $a_p = 2^{p-2}$. This gives (2.1) with $\mu = 1$. Differentiating coordinatewise gives the gradient formula, and

$$\|J_p(v)\|_q^q = \sum_j |v_j|^{(p-1)q} = \sum_j |v_j|^p = \|v\|_p^p\,.$$

Since $p/q = p - 1$, this proves (2.5). Summing the Jensen inequalities proves the finite-sum claim. The factor $a_p$ is necessary for this normalization: for $p = 4$, $u = 1$, $v = -1$, and $\theta = 1/2$, the unscaled atom $|\cdot|^4/4$ does not have unit modulus in (2.1). □

*Proof of Lemma 2.2.* We first map every parameter to the cited theorem and obtain the objective guarantee; uniform convexity then converts that guarantee to distance. In the second, nonlogarithmic branch of Diakonikolas and Guzmán [6, Theorem 2], take

$$q_{\mathrm{src}} = p, \quad \lambda = \mu, \quad \psi = H, \quad x_{\mathrm{init}} = c, \quad \phi = \omega_c, \quad m_0 = A_0 M_0, \quad \eta = \frac{\mu}{p}\delta^p, \quad X = \mathbb{R}^d,$$

and take $z = y_T$, the theorem's output after the fixed budget $T$. Here $q_{\mathrm{src}}$ is the source's uniform-convexity order, not the dual exponent $q = p/(p-1)$. The internal oracle requested by that specialization is exactly the minimization assumed in Lemma 2.2. At $(p, \kappa) = (2, 2)$, define the source recurrence quantity $M_t = L$ by continuous extension; this removes the formal $0^0$ in the generic expression without changing the endpoint method.

By Lemma 2.1, $\phi(u) \geq \|u - c\|_p^p / p$ and $\phi(z^*) \leq a_p D^p / p$. The cited branch uses at most

$$C_{p,\kappa} \left(\frac{L}{\eta}\right)^{p/\Delta} \phi(z^*)^{\kappa/\Delta}$$

iterations and guarantees (2.8). Substitution gives (2.7), because

$$\delta^{-p\beta} D^{\kappa\beta} = \delta^{-(p-\kappa)\beta} \left(\frac{D}{\delta}\right)^{\kappa\beta}.$$

All remaining factors depend only on $p, \kappa$ and are absorbed into $C^{\mathrm{cc}}_{p,\kappa}$. Finally, $F = f + H$ is $(\mu, p)$-uniformly convex and $0 \in \partial F(z^*)$, so

$$F(z) - F(z^*) \geq \frac{\mu}{p} \|z - z^*\|_p^p.$$

Together with (2.8), this yields $\|z - z^*\|_p \leq \delta$. Only $\nabla f$ is queried; $H$ and the composite map are known. □

## 3 Accumulative shifted-power regularization

We now combine the normalized atom with the one-stage contraction. After disposing of the zero-query regime, we specify the radii and weights, state the method, and introduce the exact regularized minimizers used only in the analysis.

If $\varepsilon \geq LR^{\kappa-1}$, then $\nabla f(x^*) = 0$, so (1.1)–(1.2) imply $\|\nabla f(x_0)\|_q \leq LR^{\kappa-1} \leq \varepsilon$. This also covers $L = 0$ or $R = 0$. Henceforth assume

$$0 < \varepsilon < LR^{\kappa-1}. \tag{3.1}$$

Set

$$\begin{aligned} &\rho = 2, && \tau = \rho^{p-(\kappa+1)/2}, \\ &r = \frac{\tau}{\rho^{p-1}} = \rho^{-(\kappa-1)/2} = 2^{-(\kappa-1)/2}, && \vartheta = \rho^{-(p-1)}, \\ &U = \frac{\rho+1}{\rho-1} = 3. \end{aligned} \tag{3.2}$$

Define target radii and the number of stages by

$$\delta_s = R\rho^{-s} \quad (s \geq 0), \qquad S = \left\lceil \frac{1}{\kappa - 1} \log_\rho \left(\frac{2LR^{\kappa-1}}{\varepsilon}\right) \right\rceil. \tag{3.3}$$

The cumulative regularization weights are

$$\sigma_0 = 0, \qquad \sigma_1 = \frac{(1-r)\varepsilon}{2a_p U^{p-1}(1-\vartheta)R^{p-1}}, \qquad \sigma_s = \sigma_1 \tau^{s-1} \quad (s \geq 1), \tag{3.4}$$

and their increments are

$$\alpha_s = \sigma_s - \sigma_{s-1} > 0. \tag{3.5}$$

Here $\alpha_1 = \sigma_1 > 0$; for $s \geq 2$, positivity follows from $\tau > 1$, which holds because $p \geq 2$ and $\kappa \leq 2$.

**Accumulative shifted-power regularization. Input:** $x_0, L, R, \varepsilon$, $2 \leq p < \infty$, and $1 < \kappa \leq 2$. If $\varepsilon \geq LR^{\kappa-1}$, return $x_0$. Otherwise compute (3.2)–(3.5), set $H_0 = 0$, and for $s = 1, \ldots, S$ do the following.

1. Form the known composite regularizer and objective

$$H_s(x) = H_{s-1}(x) + \alpha_s \omega_{x_{s-1}}(x), \qquad F_s(x) = f(x) + H_s(x). \tag{3.6}$$

2. Starting from $x_{s-1}$, apply Lemma 2.2 with $H = H_s$, $c = x_{s-1}$, $D = \delta_{s-1}$, $\delta = \delta_s$, and $\mu = \sigma_s$. Use the fixed integer query budget

$$T_s = \left\lceil C^{\mathrm{cc}}_{p,\kappa} \left[ 1 + \rho^{\kappa\beta} \left( \frac{L}{\sigma_s \delta_s^{p-\kappa}} \right)^{\beta} \right] \right\rceil, \tag{3.7}$$

which guarantees a point $x_s$ satisfying

$$F_s(x_s) - \min_x F_s(x) \leq \eta_s, \qquad \eta_s = \frac{\sigma_s}{p} \delta_s^p. \tag{3.8}$$

Return $x_S$.

For analysis only, let $x_s^*$ be the unique minimizer of $F_s$ for $s = 1, \ldots, S$. Set $F_0 = f$ and choose $x_0^* = x^*$, where $x^*$ is any minimizer satisfying (1.2). These exact minimizers are not algorithmic inputs. Since $\sum_{i=1}^{s} \alpha_i = \sigma_s$, Lemma 2.1 makes $H_s$ $(\sigma_s, p)$-uniformly convex. It is also coercive from the first stage. As $f$ is bounded below by its attained minimum, $F_s$ is coercive and uniformly convex, so $x_s^*$ exists and is unique.

# 4 Proof of the upper bound

The proof has three ingredients: transport and contraction give valid warm starts, a path estimate controls every historical center, and stationarity followed by Hölder continuity gives the computed output's gradient bound. We then sum the stage query budgets.

**Lemma 4.1** (Minimizer transport and stage contraction)**.** *For every $s = 1, \ldots, S$,*

$$\|x_s^* - x_{s-1}\|_p \leq \delta_{s-1}, \qquad \|x_s - x_s^*\|_p \leq \delta_s. \tag{4.1}$$

*Proof.* The induction begins with $\|x_0 - x_0^*\|_p \leq R = \delta_0$. Suppose $\left\|x_{s-1} - x_{s-1}^*\right\|_p \leq \delta_{s-1}$. Optimality of the exact minimizers for the consecutive objectives gives

$$\begin{gathered} F_{s-1}(x_s^*) + \alpha_s \omega_{x_{s-1}}(x_s^*) \leq F_{s-1}(x_{s-1}^*) + \alpha_s \omega_{x_{s-1}}(x_{s-1}^*), \\ F_{s-1}(x_{s-1}^*) \leq F_{s-1}(x_s^*). \end{gathered}$$

Subtracting and using $\alpha_s > 0$,

$$\omega_{x_{s-1}}(x_s^*) \le \omega_{x_{s-1}}(x_{s-1}^*).$$

Because both sides are the same positive multiple of a $p$-th power of distance from $x_{s-1}$,

$$\|x_s^* - x_{s-1}\|_p \le \left\|x_{s-1}^* - x_{s-1}\right\|_p \le \delta_{s-1}.$$

Thus Lemma 2.2 applies with $D = \delta_{s-1} = \rho\delta_s$. The objective guarantee (3.8) and $(\sigma_s, p)$-uniform convexity imply

$$\frac{\sigma_s}{p}\|x_s - x_s^*\|_p^p \le F_s(x_s) - F_s(x_s^*) \le \frac{\sigma_s}{p}\delta_s^p,$$

which proves the second inequality and closes the induction. □

The accumulated centers remain close enough to the final exact minimizer to make every historical regularizer gradient summable.

**Lemma 4.2** (Path bound)**.** *For every* $1 \le i \le S$*,*

$$\|x_S^* - x_{i-1}\|_p \le U\delta_{i-1}, \qquad U = \frac{\rho+1}{\rho-1} = 3. \tag{4.2}$$

*Proof.* By Lemma 4.1, for $k \ge 1$,

$$\|x_k - x_{k-1}\|_p \le \|x_k - x_k^*\|_p + \|x_k^* - x_{k-1}\|_p \le \delta_k + \delta_{k-1}. \tag{4.3}$$

Let $n = S - i \ge 0$. The triangle inequality, (4.1), and (4.3) yield

$$\begin{aligned}\|x_S^* - x_{i-1}\|_p &\le \delta_{S-1} + \sum_{k=i}^{S-1}(\delta_k + \delta_{k-1}) \\ &= \delta_{i-1}\left[\rho^{-n} + \frac{\rho+1}{\rho-1}(1-\rho^{-n})\right] \\ &\le \frac{\rho+1}{\rho-1}\delta_{i-1}.\end{aligned}$$

This formula also covers $i = S$, when the sum is empty. □

*Proof of Theorem 1.2.* We first control the gradient at the final exact regularized minimizer $x_S^*$ by stationarity and the path bound, and then transfer the estimate to the computed output $x_S$ by Hölder continuity. We finally sum the power-dependent stage terms and absorb the additive stage count. Exact stationarity of $F_S$ at $x_S^*$, together with (2.5), gives

$$\nabla f(x_S^*) = -\nabla H_S(x_S^*) = -a_p \sum_{i=1}^{S} \alpha_i J_p(x_S^* - x_{i-1}). \tag{4.4}$$

By Lemma 4.2, $\delta_{i-1} = R\rho^{-(i-1)}$, and the exact geometric increments satisfy

$$\sum_{i=1}^{\infty} \alpha_i \vartheta^{i-1} = \sigma_1 \frac{1-\vartheta}{1-r}, \qquad \vartheta = \rho^{-(p-1)}, \quad r = \tau\vartheta. \tag{4.5}$$

Indeed, the first term is $\sigma_1$, while for $i \geq 2$, $\alpha_i = \sigma_1(\tau-1)\tau^{i-2}$. Therefore

$$\begin{aligned}\|\nabla f(x_S^*)\|_q &\leq a_p \sum_{i=1}^{S} \alpha_i \|x_S^* - x_{i-1}\|_p^{p-1} \\ &\leq a_p U^{p-1} R^{p-1} \sum_{i=1}^{\infty} \alpha_i \vartheta^{i-1} \\ &= \frac{a_p U^{p-1} R^{p-1} \sigma_1 (1-\vartheta)}{1-r} = \frac{\varepsilon}{2}. \end{aligned} \tag{4.6}$$

Moreover, (3.3) implies $\delta_S^{\kappa-1} \leq \varepsilon/(2L)$. Hence Hölder continuity and Lemma 4.1 give

$$\|\nabla f(x_S) - \nabla f(x_S^*)\|_q \leq L \|x_S - x_S^*\|_p^{\kappa-1} \leq L\delta_S^{\kappa-1} \leq \frac{\varepsilon}{2}. \tag{4.7}$$

Combining (4.6) and (4.7) proves $\|\nabla f(x_S)\|_q \leq \varepsilon$.

For the query count, put

$$B_s = \frac{L}{\sigma_s \delta_s^{p-\kappa}}, \qquad \beta = \frac{p}{\kappa p + \kappa - p}. \tag{4.8}$$

The fixed factor $\rho^{\kappa\beta}$ in (3.7) is absorbed into a $(p,\kappa)$-dependent constant. Thus stage $s$ uses at most

$$T_s \leq C_{p,\kappa}(1 + B_s^{\beta}) \tag{4.9}$$

queries. The balanced schedule gives

$$\frac{B_{s+1}}{B_s} = \frac{\rho^{p-\kappa}}{\tau} = \rho^{-(\kappa-1)/2} = r, \tag{4.10}$$

and direct substitution at the first stage gives

$$B_1 = \frac{2a_p U^{p-1}(1-\vartheta)\rho^{p-\kappa}}{1-r} \frac{LR^{\kappa-1}}{\varepsilon}. \tag{4.11}$$

Therefore

$$\begin{aligned}\sum_{s=1}^{S} T_s &\leq C_{p,\kappa} \left[ S + B_1^{\beta} \sum_{s=1}^{\infty} r^{\beta(s-1)} \right] \\ &\leq C_{p,\kappa} \left[ \log\left(2 + \frac{LR^{\kappa-1}}{\varepsilon}\right) + \left(\frac{LR^{\kappa-1}}{\varepsilon}\right)^{\beta} \right]. \end{aligned} \tag{4.12}$$

Under (3.1), the logarithm is bounded by a $(p,\kappa)$-dependent multiple of the power term. This proves (1.5). The case $\varepsilon \geq LR^{\kappa-1}$ was handled before the algorithm, completing the proof. □

# 5 Exact-real implementation and extensions

This section makes the internal composite maps explicit and distinguishes query complexity from arithmetic cost. We then give accuracy-oblivious checkpoints and extend the vector argument to Schatten geometry.

### 5.1 Internal composite computations

Every internal minimization used by Lemma 2.2 at stage $s$ has the form

$$\arg\min_{x\in\mathbb{R}^d} \left\{ \langle g, x\rangle + AH_s(x) + m\omega_{x_{s-1}}(x) \right\}. \tag{5.1}$$

All centers and coefficients in this expression are already known. The problem separates by coordinate. For each coordinate, its unique minimizer is the root of

$$g_j + a_p \sum_i w_i \operatorname{sign}(t - c_{i,j})\, |t - c_{i,j}|^{p-1} = 0, \tag{5.2}$$

where, explicitly,

$$c_i = x_{i-1}, \qquad w_i = A\alpha_i + m\mathbf{1}_{\{i=s\}} \quad (i = 1, \dots, s).$$

Here $\mathbf{1}_{\{i=s\}}$ is the indicator of $i = s$. Thus $w_i \geq 0$ and $\sum_i w_i > 0$. The left-hand side is continuous and strictly increasing from $-\infty$ to $+\infty$, so the root exists and is unique. Thus the internal problem is a well-defined computation from known data and does not query $f$.

Each Generalized AGD+ iteration uses one gradient of $f$ and one such known composite minimization, up to an inessential initialization call. Hence the number of composite calls has the same order as (1.5). If a direct implementation scans all $s$ historical centers in a stage-$s$ root, then

$$\sum_{s=1}^{S} sT_s \leq C_{p,\kappa} \left( S^2 + B_1^{\beta} \sum_{s\geq 1} s r^{\beta(s-1)} \right) = O_{p,\kappa}\left[ 1 + \left( \frac{LR^{\kappa-1}}{\varepsilon} \right)^{\beta} \right].$$

Thus even the number of shifted scalar atom contributions is $O_{p,\kappa}(d[1 + (LR^{\kappa-1}/\varepsilon)^{\beta}])$. This representation count is not a bit-complexity bound.

At $p = 2$, history can be compressed exactly. Since $a_2 = 1$,

$$H_s(x) = \frac{1}{2} \sum_{i=1}^{s} \alpha_i \|x - x_{i-1}\|_2^2 = \frac{\sigma_s}{2} \|x - \bar{c}_s\|_2^2 + \text{constant}, \tag{5.3}$$

where

$$\bar{c}_1 = x_0, \qquad \bar{c}_s = \frac{\sigma_{s-1}\bar{c}_{s-1} + \alpha_s x_{s-1}}{\sigma_s} \quad (s \geq 2).$$

Consequently, the minimizer in (5.1) is the closed-form point

$$\frac{A\sigma_s \bar{c}_s + m x_{s-1} - g}{A\sigma_s + m}. \tag{5.4}$$

The Hilbert endpoint therefore needs only $O(d)$ storage and standard real-arithmetic work per composite call; no nonlinear root primitive is needed.

This makes the theorem's internal-operation assumption explicit: the unique coordinate-root map is an admissible deterministic function of the real transcript. This is not a claim that the root is obtainable by a fixed finite sequence of algebraic real-number operations, nor is it a bit-complexity statement. Computing (5.2) to finite precision and propagating such errors through all stages requires an inexact-composite analysis. The base method uses $L, R, \varepsilon, p, \kappa$ in its schedule; Corollary 5.2 removes prior knowledge of $\varepsilon$, but adaptation to the other parameters is not part of the theorem.

## 5.2 Dual targets and accuracy-oblivious checkpoints

Two short consequences clarify the target norm and the role of the requested accuracy. They do not change the base method or its oracle model.

*Remark* 5.1 (Why the dual norm is the dimension-free target). For $q \le s < \infty$, norm monotonicity gives $\|g\|_s \le \|g\|_q$, so Theorem 1.2 also certifies every such weaker $\ell_s$ target. No dimension-free conversion is possible in the opposite direction: if $1 \le s < q$ and $g = d^{-1/q}(1, \dots, 1)$, then

$$\|g\|_q = 1, \qquad \|g\|_s = d^{1/s-1/q}.$$

Thus a stronger $\ell_s$ target is a different oracle problem.

**Corollary 5.2** (Accuracy-oblivious checkpoints)**.** *Under the hypotheses of Theorem 1.2, except for a prescribed target $\varepsilon$, there is a deterministic procedure producing checkpoints $z_1, z_2, \dots$. If $Q_j$ denotes the total number of gradient queries through checkpoint $j$, then*

$$\|\nabla f(z_j)\|_q \le LR^{\kappa-1}2^{-j}, \qquad Q_j \le C_{p,\kappa}2^{j\beta},$$

*where $\beta = p/(\kappa p + \kappa - p)$. Hence prior knowledge of $\varepsilon$ is unnecessary, with no multiplicative logarithmic loss in the final query rate; the wrapper still uses $L, R, p, \kappa$.*

*Proof.* If $LR^{\kappa-1} = 0$, set every $z_j = x_0$ and $Q_j = 0$. Otherwise, restart Theorem 1.2 from $x_0$ in phase $j$ with target $\varepsilon_j = LR^{\kappa-1}2^{-j}$. The phase costs $O_{p,\kappa}(2^{j\beta})$, and the geometric sum through phase $j$ has the same order. □

## 5.3 Schatten geometry and the multicenter matrix prox

*Proof of Corollary 1.4.* We first construct a normalized Schatten power atom and reuse the outer transport–path–stationarity argument. We then verify that the multicenter prox is a single-valued exact internal map in Definition 1.1. Equip $\mathbb{R}^{m\times n}$ with the trace pairing $\langle G, X\rangle = \operatorname{tr}(G^\top X)$. The dual of $\|\cdot\|_{\mathsf{S}_p}$ is $\|\cdot\|_{\mathsf{S}_q}$. For

$$\Psi_C(X) = \frac{1}{p}\,\|X - C\|_{\mathsf{S}_p}^p$$

and $D_{\Psi_C}(Y, X) = \Psi_C(Y) - \Psi_C(X) - \langle \nabla\Psi_C(X), Y - X\rangle$, the Clarkson–McCarthy inequality for trace ideals gives [2, 16]

$$\Psi_C\left(\frac{X+Y}{2}\right) + \frac{1}{p}\left\|\frac{X-Y}{2}\right\|_{\mathsf{S}_p}^p \le \frac{\Psi_C(X) + \Psi_C(Y)}{2}.$$

For completeness, the rectangular case follows from the square one by the self-adjoint dilation $\mathcal{D}(Z) = \left(\begin{smallmatrix} 0 & Z \\ Z^\top & 0 \end{smallmatrix}\right)$: each singular value of $Z$ occurs twice in $\mathcal{D}(Z)$, so every term in the square inequality acquires the same factor two. Convexity at $X$ also gives $\Psi_C((X+Y)/2) \ge \Psi_C(X) + \frac{1}{2}\langle \nabla\Psi_C(X), Y - X\rangle$. Combining the two displays and multiplying by two yields the dimension-free estimate

$$D_{\Psi_C}(Y, X) \ge \frac{2^{1-p}}{p}\,\|Y - X\|_{\mathsf{S}_p}^p\,. \tag{5.5}$$

For $Z = (1-\theta)X + \theta Y$, cancellation of the linear Bregman terms and (5.5) give

$$\begin{aligned}
&(1-\theta)\Psi_C(X) + \theta\Psi_C(Y) - \Psi_C(Z) \\
&\quad = (1-\theta)D_{\Psi_C}(X,Z) + \theta D_{\Psi_C}(Y,Z) \\
&\quad \geq \frac{2^{1-p}}{p}\theta(1-\theta)\big(\theta^{p-1} + (1-\theta)^{p-1}\big)\,\|X-Y\|_{\mathsf{S}_p}^p \\
&\quad \geq \frac{2^{3-2p}}{p}\theta(1-\theta)\,\|X-Y\|_{\mathsf{S}_p}^p .
\end{aligned}$$

Thus $\Psi_C$ is $(2^{3-2p}, p)$-uniformly convex in the Jensen convention (2.1). Normalize it as

$$\Omega_C(X) = 2^{2p-3}\Psi_C(X).$$

Then $\Omega_C$ is $(1,p)$-uniformly convex and is radial about $C$. If $X - C = P\,\mathrm{Diag}(\varsigma)Q^\top$ is a compact singular-value decomposition, differentiation of the spectral power gives

$$\nabla\Omega_C(X) = 2^{2p-3}P\,\mathrm{Diag}(\varsigma^{p-1})Q^\top, \qquad \|\nabla\Omega_C(X)\|_{\mathsf{S}_q} = 2^{2p-3}\,\|X-C\|_{\mathsf{S}_p}^{p-1}. \tag{5.6}$$

The formula extends continuously across repeated and zero singular values.

Run the algorithm of Section 3 with $\omega_c$ replaced by $\Omega_C$, and replace $a_p$ in (3.4) by $2^{2p-3}$. The transport proof uses only radial monotonicity; the path proof uses only the triangle inequality; and the terminal stationarity estimate uses (5.6). Apply the norm-agnostic theorem underlying Lemma 2.2 on the matrix space with norm $\mathsf{S}_p$, with the same $q_{\mathrm{src}} = p$, $\lambda = \mu$, objective error, and stage parameters, and with $\phi = \Omega_C$. The required prox bounds are

$$\frac{1}{p}\,\|X-C\|_{\mathsf{S}_p}^p \leq \Omega_C(X) \leq \frac{2^{2p-3}}{p}\,\|X-C\|_{\mathsf{S}_p}^p .$$

Consequently every display in Section 4 carries over, with constants depending only on $p$. This proves the asserted upper bound.

One oracle-model qualification differs from the vector case. The internal problem, for known $G \in \mathbb{R}^{m\times n}$ and $A, \xi > 0$, is now

$$\arg\min_X \big\{\langle G, X\rangle + AH_s(X) + \xi\Omega_{X_{s-1}}(X)\big\}. \tag{5.7}$$

It has a unique minimizer because its known nonlinear part is coercive and uniformly convex. Thus (5.7) defines a single-valued map of the exact transcript and is an admissible internal operation in the unrestricted exact-real query model. In general, shifted spectral powers around different, noncommuting centers cannot be simultaneously diagonalized, so (5.7) is not a separable singular-value computation. We make no finite-precision, arithmetic, or bit-complexity claim for this matrix extension. At the Hilbert endpoint $p = 2$, however, the chosen normalization gives $\Omega_C(X) = \|X-C\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}} = \|\cdot\|_{\mathsf{S}_2}$, and the entire history compresses to a weighted center. In that case (5.7) has the explicit solution

$$X = \frac{A\sum_{i=1}^s \alpha_i X_{i-1} + \xi X_{s-1} - G/2}{A\sigma_s + \xi}.$$

Thus the nonseparable exact-multicenter-prox qualification is needed only for the genuinely nonquadratic case $p > 2$. This completes the proof of the unconditional upper bound. □

**Conditional diagonal lower transfer.** Assume assumption 6.1. Let $k = \min\{m, n\}$, let $\mathrm{Diag}_{m,n} : \mathbb{R}^k \to \mathbb{R}^{m\times n}$ place its argument on the rectangular main diagonal, and let diag be its adjoint. Given a vector hard instance $h : \mathbb{R}^k \to \mathbb{R}$, define

$$\widetilde{h}(X) = h(\operatorname{diag} X).$$

Diagonal extraction is contractive from Schatten $p$ to $\ell_p$, since

$$\|\operatorname{diag} X\|_p = \sup_{\|u\|_q\le 1} \langle \mathrm{Diag}_{m,n}\, u, X\rangle \le \|X\|_{\mathsf{S}_p}.$$

Moreover,

$$\nabla\widetilde{h}(X) = \mathrm{Diag}_{m,n}(\nabla h(\operatorname{diag} X)), \qquad \left\|\nabla\widetilde{h}(X)\right\|_{\mathsf{S}_q} = \|\nabla h(\operatorname{diag} X)\|_q.$$

Choose $X_0 = \mathrm{Diag}_{m,n}(x_0) = 0$ and, for a vector minimizer $x^*$, choose $X^* = \mathrm{Diag}_{m,n}(x^*)$. Then $X^*$ minimizes $\widetilde{h}$ and $\|X_0 - X^*\|_{\mathsf{S}_p} = \|x_0 - x^*\|_p$. Hence $\widetilde{h}$ preserves the Hölder constant, the initial radius, and the gradient target. A matrix value-gradient query is simulated by one vector query at $\operatorname{diag} X$. Therefore any faster matrix method would contradict the conditional vector lower bound in Corollary 1.3. Hence the same exponent follows in the corresponding small-accuracy and large-$k$ regime, conditional on assumption 6.1. This diagonal transfer is also the standard route from vector to Schatten oracle lower bounds [5, 10].

*Remark* 5.3 (Other Schatten gradient targets). Let $k = \min\{m, n\}$. If $q \le s < \infty$, then $\|G\|_{\mathsf{S}_s} \le \|G\|_{\mathsf{S}_q}$, so Corollary 1.4 certifies the $\mathsf{S}_s$ target without any change. If $1 \le s < q$, the sharp norm conversion is

$$\|G\|_{\mathsf{S}_s} \le k^{1/s-1/q}\, \|G\|_{\mathsf{S}_q}.$$

Running the corollary with dual-norm tolerance $\varepsilon k^{-(1/s-1/q)}$ therefore gives the valid bound

$$O_{p,\kappa}\left[1 + \left(\frac{LR^{\kappa-1}k^{1/s-1/q}}{\varepsilon}\right)^{p/(\kappa p+\kappa-p)}\right].$$

The dimensional norm factor cannot be improved. Equality is attained by

$$G = k^{-1/q}\,\mathrm{Diag}_{m,n}(1,\dots,1).$$

This is a sharpness statement for norm conversion, not a matching oracle lower bound for the stronger $\mathsf{S}_s$ target.

# 6 Conditional minimax comparison

The proof of Theorem 1.2 is independent of this section. Here we separate the simultaneous hard-family premise needed by Corollary 1.3, prove the localization lemma, and rescale the normalized instance to $(L, R, \varepsilon)$.

There is a small interface point in the one-line reduction from objective error to gradient norm. Convexity gives

$$f(x) - f(x^*) \le \langle \nabla f(x), x - x^*\rangle \le \|\nabla f(x)\|_q\, \|x - x^*\|_p, \tag{6.1}$$

so one must control the distance of the *output* from a minimizer, not only the distance of a minimizer from the starting point. The coercive norm branch in the hard family of Diakonikolas and Guzmán [5] supplies exactly this localization. We record the short argument.

The individual cited statements do not establish the full simultaneous hard-family interface used below. We therefore state that premise explicitly and make the objective-to-gradient reduction conditional on it.

**Assumption 6.1** (Simultaneous normalized hard-family interface)**.** Fix $2 \le p < \infty$, $1 < \kappa \le 2$, and put $q = p/(p-1)$ and $\beta = p/(\kappa p + \kappa - p)$. There are positive constants $c^{\mathrm{obj}}_{p,\kappa}, C^{\mathrm{obj}}_{p,\kappa}, c^{\mathrm{qry}}_{p,\kappa}$, depending only on $p, \kappa$, with the following property. If an integer $d \ge 1$ and $\zeta$ satisfy

$$0 < \zeta \le c^{\mathrm{obj}}_{p,\kappa}, \qquad \log d \ge p, \qquad d \ge C^{\mathrm{obj}}_{p,\kappa}\zeta^{-\beta},$$

then, for every deterministic local-oracle method starting from the origin, there is a normalized unconstrained $\ell_p^d$ instance $F$ such that no queried point has objective gap at most $\zeta$ before $c^{\mathrm{qry}}_{p,\kappa}\zeta^{-\beta}$ calls. The instance satisfies

$$\|\nabla F(x) - \nabla F(y)\|_q \le \|x - y\|_p^{\kappa - 1}.$$

Moreover, the same instance is realized as $F = \mathcal{S}h/\mu$ for an integer $M \ge 1$, parameters $\bar{\delta}, \mu > 0$ and $\eta, r_{\mathrm{h}} \ge 0$, and vectors $z_1, \dots, z_M \in \mathbb{R}^d$ satisfying $\|z_i\|_q \le 1$.

All hypotheses of Lemma 6.2 hold: $h$ has the form (6.3); $\mathcal{S}$ is convexity-preserving and $r_{\mathrm{h}}$-local; $\mathcal{S}h$ is differentiable; and $\|\mathcal{S}h - h\|_\infty \le \eta$. These parameters satisfy

$$r_{\mathrm{h}} \le \frac{\bar{\delta}}{8}, \qquad M\bar{\delta} \le \mu\zeta, \qquad \eta \le \frac{\mu\zeta}{4}, \qquad 4\mu\zeta \le 1. \tag{6.2}$$

The queried-point convention above is the one used in the reduction; an arbitrary transcript-measurable output is handled by one appended query.

The rate and dimension components are stated in Diakonikolas and Guzmán [6, Theorem 4, specialized to smoothness order $\kappa$]; its construction note invokes $p$-norm smoothing together with the coercive norm branch in Diakonikolas and Guzmán [5, Eq. (3)]. In the notation of the formal JMLR Diakonikolas and Guzmán [5, Theorem 3 and Lemma 40], with its target accuracy renamed $\zeta$, its smoothness parameter set to $\kappa_{\mathrm{old}} = \kappa - 1$, and its locality radius renamed $r_{\mathrm{h}}$, condition (a) gives $r_{\mathrm{h}} \le \bar{\delta}/8$ and $\eta \le \mu\zeta/4$, while condition (d), $\bar{\delta} \le \mu\zeta/M$, is equivalent to $M\bar{\delta} \le \mu\zeta$. For the $p \ge 2$ realization of the hard function in their Eq. (3), the verification of condition (b) imposes $M \le (4\mu\zeta)^{-p}$; since $M \ge 1$, this also yields $4\mu\zeta \le 1$. These citations support the individual components of assumption 6.1; the manuscript does not claim that they independently verify their simultaneous realization with the optimal dimension threshold.

**Lemma 6.2** (Localization in the hard family)**.** *Adopt the hard-family notation of Diakonikolas and Guzmán [5, Eq. (3)]. In particular, let $d, M \ge 1$ be integers, $2 \le p < \infty$, $q = p/(p-1)$, $\bar{\delta}, \mu > 0$, $\eta, r_{\mathrm{h}} \ge 0$, and $z_1, \dots, z_M \in \mathbb{R}^d$ with $\|z_i\|_q \le 1$. Let*

$$h(x) = \max\left\{\frac{1}{2}\max_{1 \le i \le M}\big(\langle z_i, x\rangle - i\bar{\delta}\big),\ \|x\|_p - C\right\}, \qquad C = \frac{1}{2}\big(3(1 + r_{\mathrm{h}}) + M\bar{\delta}\big), \tag{6.3}$$

*where $r_{\mathrm{h}}$ is the locality radius. Let $\mathcal{S}$ be the convexity-preserving local smoothing used in that construction, assume that $\mathcal{S}h$ is differentiable and $\|\mathcal{S}h - h\|_\infty \le \eta$, and set $F = \mathcal{S}h/\mu$. Then, at every point $x$ where $F$ is differentiable,*

$$\big(1 - \mu\,\|\nabla F(x)\|_q\big)\,\|x\|_p \le C - \frac{\bar{\delta}}{2} + 2\eta. \tag{6.4}$$

*Proof.* The norm branch gives $\mathcal{S}h(x) \geq \|x\|_p - C - \eta$. At the origin the affine branch equals $-\bar{\delta}/2$, and $C > \bar{\delta}/2$, so $h(0) = -\bar{\delta}/2$ and hence $\mathcal{S}h(0) \leq -\bar{\delta}/2 + \eta$. Convexity of $F$ yields

$$F(x) \leq F(0) + \langle \nabla F(x), x \rangle .$$

Combining these three inequalities and applying Hölder's inequality proves (6.4). □

*Proof of Corollary 1.3.* We first localize both a sufficiently small-gradient output and a normalized minimizer to a fixed ball. We then rescale to $(L, R)$, convert the gradient guarantee into objective accuracy, and invoke assumption 6.1.

Write $\varepsilon_{\mathrm{obj}}$ for the target objective accuracy in the normalized hard instance and apply assumption 6.1 with $\zeta = \varepsilon_{\mathrm{obj}}$. Its simultaneous realization satisfies

$$r_{\mathrm{h}} \leq \frac{\bar{\delta}}{8}, \qquad M\bar{\delta} \leq \mu\varepsilon_{\mathrm{obj}}, \qquad \eta \leq \frac{\mu\varepsilon_{\mathrm{obj}}}{4}, \qquad 4\mu\varepsilon_{\mathrm{obj}} \leq 1. \tag{6.5}$$

Because $M \geq 1$, these inequalities give $3r_{\mathrm{h}} \leq 3\mu\varepsilon_{\mathrm{obj}}/8$, $(M-1)\bar{\delta} \leq \mu\varepsilon_{\mathrm{obj}}$, and $4\eta \leq \mu\varepsilon_{\mathrm{obj}}$. Consequently,

$$\begin{aligned} 2\left(C - \frac{\bar{\delta}}{2} + 2\eta\right) &= 3(1 + r_{\mathrm{h}}) + (M-1)\bar{\delta} + 4\eta \\ &\leq 3 + \frac{19}{8}\mu\varepsilon_{\mathrm{obj}} < 4. \end{aligned} \tag{6.6}$$

Thus Lemma 6.2 places every point satisfying $\mu \|\nabla F(x)\|_q \leq 1/2$ in $B_p(0, 4)$. Moreover, $\mathcal{S}h(x) \geq \|x\|_p - C - \eta$ makes $F$ coercive, so it has a global minimizer. Choose $y^* \in \arg\min F$. Differentiability gives $\nabla F(y^*) = 0$, and (6.4) yields $\|y^*\|_p \leq C - \bar{\delta}/2 + 2\eta < 2$, so $y^* \in B_p(0, 4)$.

The scaling can now be tracked explicitly. The normalized $F$ has Hölder-gradient constant one and exponent $\kappa - 1$ in the construction. For prescribed $L, R > 0$ and starting point $x_0 = 0$, define

$$\widehat{f}(x) = \frac{LR^\kappa}{4^\kappa} F\left(\frac{4x}{R}\right), \qquad \varepsilon_{\mathrm{obj}} = \frac{2 \cdot 4^\kappa \varepsilon}{LR^{\kappa-1}}. \tag{6.7}$$

Set $\widehat{x}^* = (R/4)y^*$, which minimizes $\widehat{f}$ and lies in $B_p(0, R)$. Normalized radius-four points map into $B_p(0, R)$, and

$$\nabla \widehat{f}(x) = \frac{LR^{\kappa-1}}{4^{\kappa-1}} \nabla F\left(\frac{4x}{R}\right).$$

Consequently, for all $x, x'$,

$$\left\|\nabla \widehat{f}(x) - \nabla \widehat{f}(x')\right\|_q \leq \frac{LR^{\kappa-1}}{4^{\kappa-1}} \left\|\frac{4(x - x')}{R}\right\|_p^{\kappa-1} = L\left\|x - x'\right\|_p^{\kappa-1},$$

which verifies (1.1) after rescaling. A value-gradient query to $\widehat{f}$ at $x$ is simulated by one local query to $F$ at $4x/R$, so this rescaling preserves both the number of queries and the local-oracle information structure. If an oracle convention allows the method to return a point that it never queried, append one query at that output before applying assumption 6.1. If the original method uses $Q$ queries, then $Q + 1 \geq c^{\mathrm{qry}}_{p,\kappa} \varepsilon^{-\beta}_{\mathrm{obj}}$. After decreasing the small-accuracy constant so that the right-hand side is at least two, this implies $Q \geq (c^{\mathrm{qry}}_{p,\kappa}/2)\varepsilon^{-\beta}_{\mathrm{obj}}$.

Let $\widehat{x}$ be an $\varepsilon$-small-gradient output for $\widehat{f}$ and set $y = 4\widehat{x}/R$. Then $\|\nabla F(y)\|_q \leq \varepsilon_{\mathrm{obj}}/8$. By (6.5), its scaled gradient obeys $\mu \|\nabla F(y)\|_q \leq 1/32 < 1/2$, closing the localization bootstrap. Thus

$y \in B_p(0,4)$ and $\widehat{x} \in B_p(0,R)$. The points $\widehat{x}$ and $\widehat{x}^*$ are therefore at distance at most $2R$. Equation (6.1) gives

$$\widehat{f}(\widehat{x}) - \widehat{f}(\widehat{x}^*) \le 2\varepsilon R.$$

Using $y^* = 4\widehat{x}^*/R$ and $\widehat{f} = (LR^\kappa/4^\kappa)F(4\,\cdot\,/R)$, this is exactly

$$F(y) - F(y^*) \le \frac{4^\kappa}{LR^\kappa}(2\varepsilon R) = \frac{2\cdot 4^\kappa \varepsilon}{LR^{\kappa-1}} = \varepsilon_{\mathrm{obj}}.$$

Thus an $\varepsilon$-small-gradient method for $\widehat{f}$ would solve the conditional objective hard family to accuracy $\varepsilon_{\mathrm{obj}}$. Using objective target $2\varepsilon R$, rather than $\varepsilon R$, changes the lower bound only by a $(p,\kappa)$-dependent constant factor. Moreover, with $\beta = p/(\kappa p + \kappa - p)$,

$$\varepsilon_{\mathrm{obj}}^{-\beta} = (2\cdot 4^\kappa)^{-\beta}\left(\frac{LR^{\kappa-1}}{\varepsilon}\right)^{\beta}.$$

Hence the accuracy restriction is enforced by $\varepsilon \le c_{p,\kappa} LR^{\kappa-1}$, and the dimension requirement is enforced by $d \ge C^{\mathrm{lb}}_{p,\kappa}(LR^{\kappa-1}/\varepsilon)^\beta$ together with $\log d \ge p$, after enlarging $C^{\mathrm{lb}}_{p,\kappa}$. This proves the conditional lower-bound half of Corollary 1.3; its upper half is Theorem 1.2. □

# 7 Schedule analysis and diagnostics

The main proof fixes one convenient balance, but the admissible schedule is not unique. This section derives the schedule family, compares the upper bound with one-stage regularization, and records finite-dimensional special cases and limitations.

## 7.1 The balanced schedule is structural

The choice in (3.2) is one point in a transparent family. Choose $\rho > 1$ and $\tau > 0$, set $\delta_s = R\rho^{-s}$, and let $\sigma_s = \sigma_1 \tau^{s-1}$. The historical regularizer gradients have ratio

$$r_{\mathrm{hist}} = \frac{\tau}{\rho^{p-1}}, \tag{7.1}$$

whereas the stage complexity quantities $B_s$ have ratio

$$r_{\mathrm{cost}} = \frac{\rho^{p-\kappa}}{\tau}. \tag{7.2}$$

Both series decrease precisely when

$$\rho^{p-\kappa} < \tau < \rho^{p-1}. \tag{7.3}$$

Since $r_{\mathrm{hist}} r_{\mathrm{cost}} = \rho^{-(\kappa-1)}$, the choice

$$\tau = \rho^{p-(\kappa+1)/2} \tag{7.4}$$

equalizes the two ratios at $\rho^{-(\kappa-1)/2}$ and minimizes $\max\{r_{\mathrm{hist}}, r_{\mathrm{cost}}\}$ for fixed $\rho$. The path argument gives $U = (\rho+1)/(\rho-1)$. Thus $\rho = 2$ yields the convenient values $r_{\mathrm{hist}} = r_{\mathrm{cost}} = 2^{-(\kappa-1)/2}$ and $U = 3$.

One can also optimize the dominant query constant rather than the worse of the two ratios. Using the exact increment sum (4.5), the $\tau$-dependent factor in the leading power term is proportional to

$$\frac{1}{(1-r_{\mathrm{hist}})^\beta (1-r_{\mathrm{cost}}^\beta)}, \qquad r_{\mathrm{hist}} r_{\mathrm{cost}} = \rho^{-(\kappa-1)}. \tag{7.5}$$

Differentiation gives the unique minimizer

$$r^*_{\mathrm{hist}} = \rho^{-(\kappa-1)\beta/(\beta+1)}, \qquad r^*_{\mathrm{cost}} = \rho^{-(\kappa-1)/(\beta+1)}, \tag{7.6}$$

or equivalently $\tau^* = \rho^{p-1-(\kappa-1)\beta/(\beta+1)}$. Replacing the balanced choice by this one, and using $r^*_{\mathrm{hist}}$ in $\sigma_1$, improves the asymptotic leading constant without changing any exponent. We keep (7.4) in the main algorithm because it makes the two invariants identical. Neither choice asserts that $\rho = 2$ minimizes the complete $(p, \kappa)$-dependent constant.

### 7.2 Comparison with one-stage regularization

In the original Lipschitz-gradient specialization $\kappa = 2$, write $\Lambda = LR/\varepsilon$. The exponent comparison is

| Method or bound | Query scale | Role |
|---|---|---|
| One shifted $p$-power regularizer [6] | $\widetilde{O}_p(\Lambda^{2(p-1)/(p+2)})$ | Previous dimension-free general upper bound |
| Accumulated shifted-power continuation | $O_p(\Lambda^{p/(p+2)})$ | Vector upper bound here and in Pelleriti et al. [18] |
| Deterministic local-oracle hard family | $\Omega_p(\Lambda^{p/(p+2)})$ | Conditional large-dimensional comparison |

For general $1 < \kappa \le 2$, put $\Lambda_\kappa = LR^{\kappa-1}/\varepsilon$. The continuation upper bound has exponent $\beta = p/(\kappa p + \kappa - p)$ in $\Lambda_\kappa$; under assumption 6.1, the large-dimensional lower comparison has the same exponent. By contrast, direct substitution into the unsimplified one-shot analysis in the proof of Diakonikolas and Guzmán [6, Theorem 3] gives exponent

$$\gamma = \frac{\kappa(p-1)}{(\kappa-1)(\kappa p + \kappa - p)}. \tag{7.7}$$

The comparison is exact:

$$\gamma - \beta = \frac{p-\kappa}{(\kappa-1)(\kappa p + \kappa - p)} \ge 0,$$

with equality in our parameter range only at $p = \kappa = 2$. Thus the continuation gain persists throughout the non-Hilbert Hölder regime, not only at the smooth endpoint.

This substitution uses the unsimplified nonlogarithmic branch in the proof of the cited theorem. The displayed simplification in arXiv v2 appears to omit a factor $1/(\kappa-1)$ in its $p > 2$ general-Hölder branch; the $\kappa = 2$ specialization and the source expression actually used here are unaffected.

The lower-bound exponent has a parallel geometric explanation. In the signed-vector template underlying the cited hard family, the extremal $\ell_q$ margin is $M^{-1/p}$. After local smoothing in the Hölder-gradient case, the normalized gradient scale is $LR^{\kappa-1}/M^{\kappa-1+\kappa/p}$ [5, 6]. Inverting this relation gives $M \asymp \Lambda_\kappa^\beta$, the same exponent as the continuation upper bound. At $\kappa = 2$, this is $LR/M^{1+2/p}$.

In the smooth case, a single regularizer must simultaneously be weak enough that its gradient bias is at most $\varepsilon$ and strong enough that one highly accurate regularized solve is affordable. That tension produces the earlier exponent. Concretely, the one-shot comparator is

$$f(x) + \frac{\lambda a_p}{p}\,\|x - x_0\|_p^p, \qquad \lambda \asymp \frac{\varepsilon}{R^{p-1}},$$

where the fixed $p$-dependent normalization $a_p$ could equivalently be absorbed into $\lambda \asymp \varepsilon/R^{p-1}$. This scale is forced by the terminal regularization bias. Applying the complementary-composite

analysis of Diakonikolas and Guzmán [6, Theorem 3] to this one-shot regularization yields the earlier $\widetilde{O}_p((LR/\varepsilon)^{2(p-1)/(p+2)})$ complexity. Accumulation changes the task at each stage: the method only needs a constant-factor distance contraction, while the new center transports the regularized minimizer along a controlled path. Old regularizers are retained rather than discarded, so uniform convexity grows geometrically; nevertheless their gradients at the final point remain summable. This recovers for small gradients the same exponent $p/(p+2)$ that power-uniform geometry permits for accelerated objective minimization [4].

Still in the smooth case, optimal $p$-uniform objective acceleration followed by ordinary steepest descent gives the valid but weaker gradient scale $O_p(LR/N^{(p+1)/p})$, corresponding to query exponent $p/(p+1)$. The continuation argument gains the missing factor $N^{-1/p}$ by converting objective progress into a controlled path of regularized minimizers rather than a single terminal descent phase.

### 7.3 Fixed Hilbert diagnostics and a quadratic special case

We return in this subsection to $\kappa = 2$. Suppose there is a known set $I \subset [d]$, $|I| = k$, such that $f(x) = \bar{f}(x_I)$; then every oracle gradient is supported on $I$, and the iterates may be restricted to $x_0 + \operatorname{span}\{e_i : i \in I\}$. On this subspace,

$$\|v\|_p \le \|v\|_2 \le k^{1/2-1/p}\,\|v\|_p\,, \qquad \|g\|_2 \le \|g\|_q \le k^{1/2-1/p}\,\|g\|_2\,.$$

Thus (1.1) implies Euclidean $L$-smoothness, the Euclidean initial radius is at most $k^{1/2-1/p}R$, and the Hilbert-space $N^{-2}$ small-gradient bound [14] gives an $N$-query method whose output $x_N^{\mathrm{H}}$ satisfies

$$\left\|\nabla f(x_N^{\mathrm{H}})\right\|_q = O\left(\frac{LR\,k^{1-2/p}}{N^2}\right). \tag{7.8}$$

This reaches the dimension-free target when $k \le N$, but a universal fixed quadratic geometry pays $d^{1-2/p}$ in the full space. Combining the $k = d$ case with the rate form of Theorem 1.2 gives, for $N$ above a $p$-dependent constant, an $N$-query method with output $x_N^{\mathrm{env}}$ satisfying the upper envelope

$$\|\nabla f(x_N^{\mathrm{env}})\|_q = O_p\left(LR\min\left\{\frac{d^{1-2/p}}{N^2}, \frac{1}{N^{(p+2)/p}}\right\}\right). \tag{7.9}$$

This comparison does not claim a matching finite-$d$ minimax interpolation.

For a fixed positive-semidefinite quadratic, an instance-dependent Hilbert metric can do better still. The following self-contained observation records precisely what is available and why it does not solve the general local-oracle problem.

**Proposition 7.1** (A supplied diagonal metric for quadratics)**.** *Let $d, N \ge 1$, $2 \le p < \infty$, $q = p/(p-1)$, $A \in \mathbb{R}^{d\times d}$, $b, x_0 \in \mathbb{R}^d$, and $R \ge 0$, where $d$ and $N$ are integers. Suppose $A \succeq 0$. Define*

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x, \qquad L_A = \|A\|_{p\to q} := \sup_{\|v\|_p \le 1} \|Av\|_q\,.$$

*Suppose that $f$ has a minimizer $x^*$ with $\|x_0 - x^*\|_p \le R$. Put*

$$s_2 = \infty, \qquad s_p = \frac{p}{p-2} \quad (p > 2), \qquad \gamma_p = \left(\mathbb{E}\,|G|^p\right)^{2/p}, \quad G \sim \mathcal{N}(0,1).$$

*We use the endpoint conventions $1/s_2 = 0$ and $d^{-1/s_2} = 1$. If $L_A = 0$, then $\nabla f \equiv 0$. If $L_A > 0$, there exists a positive diagonal matrix $D = \operatorname{Diag}(w) \succ 0$ such that*

$$D \succeq A, \qquad \|w\|_{s_p} \le 2\gamma_p L_A. \tag{7.10}$$

*If $L_A > 0$ and this metric is supplied to the algorithm, Hilbert-space mirror-dual concatenation has, after at most $N$ queries, an output $x_N^D$ satisfying*

$$\left\|\nabla f(x_N^D)\right\|_q = O_p\!\left(\frac{L_A R}{N^2}\right). \tag{7.11}$$

*For $L_A = 0$, the gradient vanishes identically, so the same estimate holds with output $x_0$ and no queries.*

*Proof.* We first construct a diagonal majorant by semidefinite duality and a Gaussian-moment bound. We then work in the induced Hilbert geometry and convert its gradient estimate back to $\ell_q$.

If $L_A = 0$, then $A = 0$; existence of a minimizer of $f(x) = -b^\top x$ forces $b = 0$, proving the first assertion and (7.11). Hence assume $L_A > 0$.

For $A \succeq 0$, Cauchy–Schwarz in the seminorm induced by $A$ shows

$$L_A = \sup_{\|u\|_p \le 1} u^\top A u.$$

The upper inequality follows from Hölder's inequality. For the converse, Cauchy–Schwarz in the $A$-seminorm gives, for all $u, v \in \mathbb{R}^d$,

$$\left|v^\top A u\right| \le (u^\top A u)^{1/2} (v^\top A v)^{1/2}.$$

Taking the supremum over the two unit balls proves the reverse inequality.

Since the exponent dual to $s_p$ is $p/2$, semidefinite duality, with strict feasibility provided by a sufficiently large multiple of the identity, gives

$$\inf_{w:\,\mathrm{Diag}(w) \succeq A} \|w\|_{s_p} = \sup\left\{\mathrm{tr}(AX) : X \succeq 0,\ \|\mathrm{diag}\, X\|_{p/2} \le 1\right\}. \tag{7.12}$$

The primal infimum is attained. Indeed, feasibility implies $w_j \ge A_{jj} \ge 0$, and intersecting the feasible set with any objective sublevel $\{w : \|w\|_{s_p} \le t\}$ gives a closed bounded set. For a dual-feasible $X$, let $Z \sim \mathcal{N}(0, X)$. Then

$$\begin{aligned}\mathrm{tr}(AX) &= \mathbb{E}[Z^\top A Z] \le L_A\, \mathbb{E}\,\|Z\|_p^2 \\ &\le L_A \big(\mathbb{E}\,\|Z\|_p^p\big)^{2/p} = \gamma_p L_A \,\|\mathrm{diag}\, X\|_{p/2} \le \gamma_p L_A.\end{aligned}$$

The equality uses $\mathbb{E}\,|Z_j|^p = \mathbb{E}\,|G|^p\, X_{jj}^{p/2}$. Thus (7.12) produces a diagonal semidefinite majorant with $s_p$-norm at most $\gamma_p L_A$. Adding $\gamma_p L_A d^{-1/s_p} I$ makes it positive definite and at most doubles the norm, because $\left\|\gamma_p L_A d^{-1/s_p}(1, \dots, 1)\right\|_{s_p} = \gamma_p L_A$. This proves (7.10).

Write $\|v\|_D = (v^\top D v)^{1/2}$ and $\|g\|_{D^{-1}} = (g^\top D^{-1} g)^{1/2}$ for its dual norm. Since $A \preceq D$, the matrix $B = D^{-1/2} A D^{-1/2}$ satisfies $0 \preceq B \preceq I$. Hence $A D^{-1} A \preceq A \preceq D$, and therefore

$$\|Av\|_{D^{-1}}^2 = v^\top A D^{-1} A v \le v^\top D v,$$

so the quadratic is 1-smooth from $\|\cdot\|_D$ to its dual norm. Hölder's inequality and norm duality give

$$\|x_0 - x^*\|_D \le \|w\|_{s_p}^{1/2} R, \qquad \|g\|_q \le \|w\|_{s_p}^{1/2} \|g\|_{D^{-1}}.$$

Apply the Hilbert-space $N^{-2}$ small-gradient result [14] in the $D$-geometry and combine the last two inequalities with (7.10) to obtain (7.11). □

The metric in Proposition 7.1 depends on the full Hessian. Nothing in its existence proof learns that metric from a dimension-free number of local queries; recovering a general quadratic Hessian can itself require dimension-dependent information. The proposition is therefore a special-case diagnostic, not an alternative proof of Theorem 1.2.

# 8 Related work

We compare the accumulated-regularization mechanism with prior vector-space results and explain why a trajectory-dependent steepest-descent guarantee does not directly imply the present bound.

The Euclidean literature develops regularization, performance-estimation, potential-function, and duality mechanisms [7, 8, 12, 13, 17]. The objective-acceleration benchmark in power-uniform geometry is due to d'Aspremont et al. [4]; the local-oracle hard families descend from large-scale smoothing and its unconstrained extension [5, 10]. The inner engine used here is complementary-composite minimization [6].

Accumulation also has close precedents in accelerated proximal-point, Euclidean, and high-order methods [9, 11, 15]. Most directly, Pelleriti et al. [18] give the same vector exponent with the closely matching quantitative construction described in the main text. Accordingly, neither accumulation in general nor the vector continuation mechanism and rate are claimed as exclusive contributions of this manuscript.

For completeness, consider Hyper-Accelerated Steepest Descent [1, Corollary 9]. If a trajectory reaches a zero gradient, the present stationarity task is already solved. Otherwise define, for every prefix $1 \le t \le N$,

$$G_t = \frac{1}{t} \sum_{j=0}^{t-1} \frac{\|\nabla f(x_{j+1})\|_q}{\|\nabla f(x_{j+1})\|_2}.$$

The cited corollary assumes a uniform $1 \le \widehat{G} \le \min_{1 \le t \le N} G_t$ and an $\ell_2$ radius. Since $q \le 2$, $\widehat{G} = 1$ is always valid on a nonzero-gradient trajectory, but it does not offset the worst-case conversion $\|x_0 - x^*\|_2 \le d^{1/2-1/p} R$. The present assumptions do not provide the stronger dimension-growing prefix bound needed to recover Theorem 1.2 from that result.

# 9 Limitations and extensions

The theorem is deliberately scoped to the model in Section 1. Several extensions require new arguments.

- **Finite precision.** The roots in (5.2) are admissible exact-real maps; arithmetic or bit-complexity results need tolerance schedules and error propagation.
- **Uniformity in $p$.** Constants can grow with fixed $p$: $a_p = 2^{p-2}$ and $U^{p-1}$ already appear explicitly. No uniformity as $p \to \infty$ is claimed.
- **Parameter adaptation.** The schedule knows $L, R, p, \kappa, \varepsilon$. The checkpoint wrapper in Corollary 5.2 removes prior knowledge of $\varepsilon$; adapting to the other parameters must preserve both geometric series without introducing additional losses.
- **Other oracle regimes.** The comparison is deterministic, local, large-dimensional, and conditional on assumption 6.1. The cited construction has a randomized extension under stronger dimension requirements, but neither a randomized matching theorem nor exact finite-$d$ interpolation is formalized here.
- **Other power-uniform geometries.** The proof uses only a dimension-free $p$-uniformly convex shifted atom, nested sublevel sets $\omega_c(u) \le \omega_c(v) \Rightarrow \|u - c\| \le \|v - c\|$, and dual-gradient growth $\|\nabla \omega_c(u)\|_* \lesssim \|u - c\|^{p-1}$, where $\|\cdot\|_*$ denotes the norm dual to the displayed primal norm. A compatible complementary-composite routine and accumulated prox are also

required. The Schatten prox need not be separable; finite precision would require a tractable inexact implementation.

- **Conceptual mirror duality.** A power-uniform analogue of the mirror-duality principle might explain the rate without continuation and expose a broader primal-dual structure; the present proof does not.

For fixed $\kappa$, the gradient decay exponent is $\kappa - 1 + \kappa/p$. At $p = 2$ it equals $(3\kappa - 2)/2$, whose smooth specialization is the familiar $N^{-2}$ scale. As $p$ grows it approaches $\kappa - 1$; at $\kappa = 2$, this is the smooth exponent $(p + 2)/p \to 1$. The explicit constants deteriorate with $p$, consistently with the fixed-$p$ scope of the theorem.

# 10 Conclusion

For every fixed $2 \le p < \infty$ and $1 < \kappa \le 2$, accumulative shifted-power regularization therefore achieves the dimension-free gradient-query bound

$$O_{p,\kappa}\left(\left[\frac{LR^{\kappa-1}}{\varepsilon}\right]^{p/(\kappa p+\kappa-p)}\right)$$

in the nontrivial regime of the deterministic unrestricted exact-real model. The proof combines minimizer transport, a constant-factor path bound, exact regularized stationarity, Hölder transfer to the computed output, and a geometric sum of the power-dependent stage terms. The same upper-bound interface extends to Schatten-$p$ geometry. Finite-precision, arithmetic, and bit-complexity guarantees remain outside the present scope.

# References


[1] C. S. Bai and B. Bullins. Faster acceleration for steepest descent. In *Proceedings of the 38th Annual Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 202–230, 2025. PMLR 291.

[2] K. Ball, E. A. Carlen, and E. H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115(1):463–482, 1994. doi:10.1007/BF01231769.

[3] J. A. Clarkson. Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40(3):396–414, 1936. doi:10.1090/S0002-9947-1936-1501880-4.

[4] A. d'Aspremont, C. Guzmán, and M. Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018. doi:10.1137/17M1116842.

[5] J. Diakonikolas and C. Guzmán. Lower bounds for parallel and randomized convex optimization. *Journal of Machine Learning Research*, 21(5):1–31, 2020. JMLR 21(5); arXiv:1811.01903v3.

[6] J. Diakonikolas and C. Guzmán. Complementary composite minimization, small gradients in general norms, and applications. *Mathematical Programming*, 208(1–2):319–363, 2024. doi:10.1007/s10107-023-02040-5; arXiv:2101.11041v2.

[7] J. Diakonikolas and P. Wang. Potential function-based framework for minimizing gradients in convex and min-max optimization. *SIAM Journal on Optimization*, 32(3):1668–1697, 2022. doi:10.1137/21M1395302.

[8] M. I. Florea. A template for gradient norm minimization. arXiv:2410.23135v1, 2024.

[9] O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992. doi:10.1137/0802032.

[10] C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015. doi:10.1016/j.jco.2014.08.003.

[11] Y. Ji and G. Lan. High-order accumulative regularization for gradient minimization in convex programming. arXiv:2511.03723v2, 2025.

[12] D. Kim and J. A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021. doi:10.1007/s10957-020-01770-2.

[13] J. Kim, A. Ozdaglar, C. Park, and E. K. Ryu. Time-reversed dissipation induces duality between minimizing gradient norm and function value. In *Advances in Neural Information Processing Systems 36*, pages 23389–23440, 2023. doi:10.52202/075280-1014.

[14] J. Kim, C. Park, A. Ozdaglar, J. Diakonikolas, and E. K. Ryu. Mirror duality in convex optimization. arXiv:2311.17296v2, 2024.

[15] G. Lan, Y. Ouyang, and Z. Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. *Mathematical Programming*, 2026. doi:10.1007/s10107-026-02352-2.

[16] C. A. McCarthy. $C_p$. *Israel Journal of Mathematics*, 5(4):249–271, 1967. doi:10.1007/BF02771613.

[17] Y. Nesterov. How to make the gradients small. *Optima: Mathematical Optimization Society Newsletter*, 88:10–11, 2012.

[18] N. Pelleriti, M. Shiran, D. Martínez-Rubio, M. Zimmer, and S. Pokutta. Optimal gradient-norm minimization in non-Euclidean Hölder-smooth convex optimization. arXiv:2609.01122v1, 2026.