ROBUST LEARNING MEETS QUASAR-CONVEX OPTIMIZATION: INEXACT HIGH-ORDER PROXIMAL-POINT METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust representation alignment can lead to nonsmooth and nonconvex objectives when paired features are corrupted or mismatched. We show that several realizable alignment models built from stabilized capped and anisotropic losses are strongly quasar-convex. Motivated by this geometry, we study Inexact HiPPA, an inexact high-order proximal-point method with regularization order $p > 1$. Under summable model-value inexactness, bounded HiPPA sequences converge in objective value to the global minimum, while strong quasar-convexity makes boundedness and uniqueness automatic. Under local stability and suitable forcing of the inexactness, the method converges linearly for $1 < p \leq 2$ and superlinearly with Q-order $p - 1$ for $p > 2$. A capped-quadratic alignment model verifies these local assumptions explicitly. On a controlled CIFAR-100 feature experiment, the mean outer iteration count drops from 14.8 for $p = 2$ to 4 for $p = 3$ and 3 for $p = 4$. A second corrupted-feature experiment uses a computable residual-and-descent rule and shows the clearest benefit from capped $p = 3$ fitting under severe replacement corruption.

1 INTRODUCTION

Robust learning methods are designed to reduce sensitivity to corrupted or mismatched supervision. This issue is especially relevant in representation alignment, where teacher features or paired embeddings may contain outliers or incorrect matches. Robust losses limit the influence of such observations, but capped and redescending constructions often lead to nonsmooth and nonconvex objectives (Huber, 1964; Barron, 2019; Wang et al., 2019; Song et al., 2023).

Quasar-convexity provides a useful middle ground between convexity and general nonconvexity. Although quasar-convex objectives may be nonconvex, their geometry remains directed toward a global minimizer, which rules out spurious local minima and supports global optimization guarantees. Strong quasar-convexity further provides quantitative growth away from the minimizer and can lead to linear or faster local convergence under suitable algorithmic conditions. These properties make the class broad enough to capture several learning problems while retaining favorable optimization structure (Hardt et al., 2018; Hinder et al., 2020; Wang & Wibisono, 2023; Ahookhosh et al., 2026a). Most existing algorithmic results focus on smooth objectives, whereas nonsmooth quasar-convex optimization remains less developed (de Brito et al., 2025; Ahookhosh et al., 2026a). We study robust alignment models whose clean realizable objectives belong to this nonsmooth class. Accordingly, a natural question arises in the optimization and machine learning communities:

"*How can we design optimization algorithms that are both theoretically rigorous and computationally efficient for (strongly) quasar-convex objectives arising in robust learning?*"

We then develop an inexact high-order proximal-point method. At iteration k, the method computes an approximation to a solution of the proximal subproblem

$$\min_{y \in \mathbb{R}^n} \left\{ h(y) + \frac{1}{p\beta_k} \left\| y - x^k \right\|^p \right\}, \tag{1}$$

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

where $p > 1$ is the regularization order and $\beta_k > 0$ is the proximal parameter. Rather than requiring an exact minimizer of (1), the method accepts an approximate solution satisfying a prescribed inexactness criterion. This gives the inexact high-order proximal-point algorithm, or inexact HiPPA.

The main contributions are as follows.

- **Robust alignment models.** We identify nonsmooth robust feature-alignment and anisotropic model-stitching objectives that are strongly quasar-convex under clean realizability and identifiability assumptions. We also give stability and local recovery results when target features are corrupted.
- **Inexact HiPPA method.** We design inexact HiPPA approach which is flexible to choose appropriate regularization parameter p with respect to the local geometry of the objective function and analyze it under a model-value gap criterion and establish global convergence to minimizers under summable inexactness. Under local stability and suitable forcing of the inexactness, the method converges linearly for $1 < p \leq 2$ and superlinearly with Q-order $p - 1$ for $p > 2$. These rates yield outer-iteration complexity bounds of $\mathcal{O}(\log(\varepsilon^{-1}))$ and $\mathcal{O}(\log\log(\varepsilon^{-1}))$, respectively.
- **Verified numerical regimes.** We introduce a capped-quadratic alignment model for which the local assumptions and model-value gap criterion can be verified explicitly. The corresponding order-sensitivity experiment confirms the predicted higher-order behavior and shows substantial gains for $p > 2$, with diminishing returns beyond $p = 4$ as the lower-level work increases. A separate corrupted-feature experiment uses a residual-and-descent acceptance rule and evaluates the practical inexact interface under controlled target corruption.

1.1 Related work

Quasar-convex optimization. Quasar-convex functions arising as objectives in many practical applications are inherently nonsmooth, as nonsmoothness naturally emerges from regularization (e.g., sparsity-inducing penalties) and robust loss formulations designed to mitigate the effect of noise and outliers (Ahookhosh et al., 2026a; de Brito et al., 2025). They generalize convexity while remaining a broad subclass of nonconvex functions that includes star-convex functions (Ahookhosh et al., 2026a; Guminov et al., 2023; Hermant et al., 2024; Hinder et al., 2020; Nesterov & Polyak, 2006). They enjoy useful properties, such as first-order characterizations, calculus rules, error bounds, and the Polyak-Łojasiewicz inequality, in both smooth (Guminov et al., 2023; Hardt et al., 2018; Hermant et al., 2024; Hinder et al., 2020; Pun & Shames, 2024; Wang & Wibisono, 2023) and nonsmooth settings (Ahookhosh et al., 2026a; de Brito et al., 2025). Notably, they have a benign landscape: every local minimum is global, and there are no saddle points (Ahookhosh et al., 2026a, Section 3.3). Moreover, (κ, γ)-strong quasar-convexity with $\gamma > 0$ ensures isolated global minimizers, making this class well-suited for first-order methods. Subgradient methods (Davis et al., 2018; Rahimi et al., 2026) and inexact proximal-point schemes (de Brito et al., 2025; Sra, 2012) are standard optimization methods for such nonsmooth and nonconvex problems. Although subgradient methods are simple and general, they often converge slowly and are sensitive to conditioning; proximal-point methods are typically more stable and faster via implicit regularization.

Proximal-point methods and their inexact and high-order variants. The classical proximal-point method (PPA, (1) with $p = 2$) forms a powerful framework for nonsmooth optimization, based on iteratively solving regularized subproblems that stabilize the original objective. Originating in monotone operator theory, they have strong convergence guarantees even in nonsmooth and nonconvex settings, and can be interpreted as implicit gradient steps that improve robustness to conditioning (Martinet, 1970; 1972; Rockafellar, 1976). In practice, however, the resulting proximal subproblems are often too expensive to solve exactly. Inexact proximal-point methods address this issue by allowing approximate solutions under controlled accuracy requirements, thereby reducing the per-iteration computational burden while maintaining convergence guarantees (Barré et al., 2023; Dvurechensky, 2022; Liu et al., 2012; Rockafellar, 1976; Salzo & Villa, 2012; Solodov & Svaiter, 2001). This flexibility makes them particularly suitable for large-scale settings, where exact proximal evaluations are impractical, and forms the basis for several scalable first-order and operator-splitting schemes. High-order proximal-point methods and their inexact variants have recently attracted significant attention for their ability to capture detailed local geometry of the objective (Ahookhosh et al., 2026a;b; Ahookhosh & Nesterov, 2024; 2025; Kabgani & Ahookhosh, 2025c; 2026a; 2025b; 2026b; 2025a; Nesterov, 2023). This added flexibility allows for more effective steps, improved

complexity, and a better balance between global regularization and local model fidelity, making high-order methods a natural and powerful proximal framework for modern large-scale optimization.

Exact quadratic PPA convergence under quasar-convex geometry has been studied in prior work (de Brito et al., 2025). More recently, Exact HiPPA and its p-dependent convergence rates for strongly quasar-convex objectives have been established (Ahookhosh et al., 2026a;b). While these results provide important theoretical guarantees, Exact HiPPA requires the proximal subproblem to admit a closed-form solution. Such closed-form solutions are generally unavailable for arbitrary objective functions, limiting the direct applicability of the exact method and motivating the development of practically implementable alternatives. Accordingly, our contribution is not to develop a generic exact PPA/HiPPA theory, but rather to address the algorithmic and computational challenges that arise when such proximal subproblems cannot be solved in closed form. In addition, existing smooth first-order methods for quasar-convex optimization, such as Hinder et al. (2020), assume differentiability and therefore do not directly apply to the nonsmooth capped objectives considered in this work.

1.2 Notation

Throughout, $\mathbb{R}^n$ is equipped with the inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$, $\overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$, and $\mathbb{N}_0 := \mathbb{N}\cup\{0\}$. For $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, let $\mathbf{dom}(h) := \{x : h(x) < +\infty\}$, $X^\star := \mathbf{arg\,min}_{x\in\mathbb{R}^n} h(x)$, and $h^\star := \mathbf{inf}_{x\in\mathbb{R}^n} h(x)$, and call h proper when $\mathbf{dom}(h) \neq \emptyset$. For $p > 1$, define $J_p(z) := \|z\|^{p-2} z$ with $J_p(0) := 0$, so that $J_p = \nabla(\frac{1}{p}\|\cdot\|^p)$.

For a locally Lipschitz function h, $\partial^C h(x)$ denotes its Clarke subdifferential. A point $x \in \mathbf{dom}(h)$ is Clarke critical if $0 \in \partial^C h(x)$. Further notation is collected in Appendix A.

Definition 1 (**(Strong) quasar-convexity**). *Let $\kappa \in (0, 1]$, $\gamma \geq 0$, and let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper with convex domain and $X^\star := \mathbf{arg\,min}_{x\in\mathbb{R}^n} h(x) \neq \emptyset$. We say that h is (κ, γ)-strongly quasar-convex with respect to $\overline{x} \in X^\star$ if, for every $x \in \mathbf{dom}(h)$ and every $\lambda \in [0, 1]$,*

$$h\left(\lambda\overline{x} + (1-\lambda)x\right) \leq \kappa\lambda h(\overline{x}) + (1-\kappa\lambda)h(x) - \lambda\left(1 - \tfrac{\lambda}{2-\kappa}\right)\tfrac{\kappa\gamma}{2}\left\|x - \overline{x}\right\|^2. \tag{2}$$

When $\gamma = 0$, we say that h is κ-quasar-convex with respect to $\overline{x}$.

2 Motivational examples from robust learning

We consider robust representation-alignment models whose clean realizable forms exhibit strong quasar-convexity despite nonsmooth and nonconvex residual losses. We first study capped feature alignment and a capped-quadratic variant used later in our numerical validation, and then consider anisotropic robust model stitching. Proofs and geometric properties are deferred to Appendix B.

2.1 Stabilized capped residual loss

Capped and truncated losses are standard robust mechanisms for reducing sensitivity to large residuals (Yang et al., 2023; Barron, 2019). For $\tau, \mu > 0$, define

$$\Phi_{\tau,\mu}(z) := \mathbf{min}\{\|z\|, \tau\} + \mu\|z\|^2. \tag{3}$$

The capped term limits the influence of large residuals and the quadratic term provides global stabilization.

Proposition 2 (**Stabilized capped residual loss**). *Let $\tau, \mu > 0$ satisfy $0 < \mu\tau < 1/2$. Then $\Phi_{\tau,\mu}$ is locally Lipschitz, nonsmooth, and non-star-convex with respect to the origin. Moreover, it is (κ, γ)-strongly quasar-convex with respect to $\overline{z} = 0$, where one may take $\gamma = \mu$ and $\kappa = \frac{\mu\tau}{1+2\mu\tau}$.*

Example 3 (**Robust feature-alignment distillation**). *Feature-based knowledge distillation trains a student model using intermediate teacher representations as supervision. FitNets is a classical example in which teacher hidden representations are used as hints for a thinner student network (Romero et al., 2015). Motivated by this setting, let $Z_s = [z_{s,1}, \ldots, z_{s,N}] \in \mathbb{R}^{d_s\times N}$ and $Z_t = [z_{t,1}, \ldots, z_{t,N}] \in \mathbb{R}^{d_t\times N}$ denote student and teacher features collected from the same batch, respectively. For a trainable alignment map $W \in \mathbb{R}^{d_t\times d_s}$, we define*

$$h_{\mathrm{KD}}(W) := \frac{1}{N}\sum_{i=1}^{N} \Phi_{\tau,\mu}(Wz_{s,i} - z_{t,i}). \tag{4}$$

Assume that the source representations span the feature space used for alignment, so Z_s has full row rank, and assume $W^\star Z_s = Z_t$. Let $0 < \mu\tau < 1/2$ and $\sigma_{\mathbf{min}}(Z_s)$ be the smallest singular value of Z_s. Then h_{KD} is locally Lipschitz, nonsmooth, and (κ, γ)-strongly quasar-convex with respect to $W^\star$, where $\kappa = \frac{\mu\tau}{1+2\mu\tau}$ and $\gamma = \frac{\mu}{N}\sigma_{\mathbf{min}}(Z_s)^2$.

For the local rate study, we use the capped-quadratic loss

$$\Phi^{(2)}_{\tau,\mu}(z) := \frac{1}{2}\,\mathbf{min}\{\|z\|^2, \tau^2\} + \frac{\mu}{2}\|z\|^2, \qquad 0 < \mu < 1. \tag{5}$$

Capped-squared losses are established robust-loss mechanisms (Wang et al., 2024). This loss is locally Lipschitz, nonsmooth, nonconvex, and non-star-convex, while it is $(\frac{\mu}{1+\mu}, 1+\mu)$-strongly quasar-convex with respect to the origin; see Appendix B.3. Let us consider

$$h_{\mathrm{CQ}}(W) := \frac{1}{N}\sum_{i=1}^{N} \Phi^{(2)}_{\tau,\mu}(W z_{s,i} - z_{t,i}).$$

Proposition 4 (Capped-quadratic alignment geometry). *Assume that $W^\star Z_s = Z_t$ and that Z_s has full row rank. Then h_{CQ} is $(\frac{\mu}{1+\mu}, \frac{(1+\mu)\sigma_{\mathbf{min}}(Z_s)^2}{N})$-strongly quasar-convex with respect to $W^\star$. Let*

$$M_s := \max_i \|z_{s,i}\|, \qquad m := \frac{(1+\mu)\sigma_{\mathbf{min}}(Z_s)^2}{N}.$$

If $\|W - W^\star\|_F < \frac{\tau}{M_s}$, then

$$h_{\mathrm{CQ}}(W) = \frac{1+\mu}{2N}\|(W - W^\star)Z_s\|_F^2, \tag{6}$$

and hence h_{CQ} is locally m-strongly convex on this region.

The corresponding proximal-map and inexact model-growth properties are given in Appendix B.3.

2.2 Anisotropic robust residuals

The capped losses above are isotropic because they depend only on the residual norm. However, in neural representation learning, residual directions may have different semantic roles, reliability levels, or alignment importance (Hu et al., 2022; Romero et al., 2015; Maekawa et al., 2026; Csiszárik et al., 2021).

Proposition 5 (A locally Lipschitz nonsmooth master family). *Fix parameters $\tau, \mu > 0$ such that $\mu\tau < \frac{1}{8}$ and define*

$$\psi_\tau(r) := \begin{cases} r, & 0 \le r \le \tau, \\ \sqrt{\tau r}, & r > \tau. \end{cases} \tag{7}$$

Let $f_{\tau,\mu}(r) := \psi_\tau(r) + \mu r^2$ for $r \ge 0$. Let $d \ge 2$ and let $q : \mathbb{S}^{d-1} \to [1, \infty)$ be a nonconstant C^1 function, where $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\| = 1\}$. Define

$$H_{\tau,\mu,q}(z) := \begin{cases} f_{\tau,\mu}(\|z\|)\, q\left(\dfrac{z}{\|z\|}\right), & z \neq 0, \\ 0, & z = 0, \end{cases}$$

Then, $H_{\tau,\mu,q}$ is locally Lipschitz, nonsmooth, and non-star-convex with respect to the origin. Moreover, it is $(\frac{1}{2}, 6\mu)$-strongly quasar-convex with respect to $\overline{z} = 0$.

Example 6 (Anisotropic robust model stitching). *Let $X = [x_1, \ldots, x_N] \in \mathbb{R}^{d_x \times N}$ and $Y = [y_1, \ldots, y_N] \in \mathbb{R}^{d_y \times N}$ denote paired hidden representations from two frozen models. Consider the affine connector $x \mapsto Ax + b$. With $\widetilde{X} := \begin{bmatrix} X \\ \mathbf{1}^\top \end{bmatrix}$ and $\widetilde{A} := [A\ b]$, define the anisotropic robust stitching loss*

$$h_{\mathrm{stitch}}(\widetilde{A}) := H_{\tau,\mu,q}\left(\mathrm{vec}(\widetilde{A}\widetilde{X} - Y)\right),$$

where now $q : \mathbb{S}^{d_y N-1} \to [1, \infty)$ is a nonconstant C^1 angular weight. The angular factor q allows direction-dependent penalties in the hidden representation mismatch, while the kinked radial profile gives a nonsmooth nonconvex discrepancy.

Assume that the affine connector is identifiable, so $\widetilde{X}$ has full row rank, and assume $\widetilde{A}^\star \widetilde{X} = Y$. Then h_{stitch} is locally Lipschitz, non-star-convex, and $\left(\frac{1}{2}, 6\mu\, \sigma_{\mathbf{min}}(\widetilde{X})^2\right)$-strongly quasar-convex with respect to $\widetilde{A}^\star$.

The rank assumptions above express identifiability of the connector on the observed representation space. If the features lie in a lower-dimensional empirical span, the same models can be reparameterized on that span, where the reduced feature matrix has full row rank. Appendix B.6 gives the reduction.

3 Inexact high-order proximal-point algorithm

We now turn to the algorithmic framework. The goal is to solve

$$\min_{x \in \mathbb{R}^n} \; h(x), \tag{8}$$

under the following standing assumptions:

Assumption 7 (Basic assumptions). *The function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and lower semicontinuous, $\mathbf{dom}(h)$ is convex, and $X^\star \neq \emptyset$. Unless otherwise stated, h is κ-quasar-convex or (κ, γ)-strongly quasar-convex with respect to a fixed minimizer $\overline{x} \in X^\star$. The proximal parameters satisfy*

$$0 < \beta_{\mathbf{min}} \leq \beta_k \leq \beta_{\mathbf{max}} < \infty. \tag{9}$$

For $x \in \mathbb{R}^n$, $\beta > 0$, and $p > 1$, let us define

$$Q_{x,\beta}(y) := h(y) + \frac{1}{p\beta} \|y - x\|^p, \qquad T_{\beta,p}(x) := \operatorname*{\mathbf{arg\,min}}_{y \in \mathbb{R}^n} Q_{x,\beta}(y), \tag{10}$$

and write $Q_k := Q_{x^k,\beta_k}$. Since h is bounded below and the regularizer is coercive, $T_{\beta,p}(x) \neq \emptyset$. For a given $\beta_k > 0$, the exact HiPPA chooses $x^{k+1} \in T_{\beta_k,p}(x^k)$. We allow this subproblem to be solved inexactly.

3.1 Model-value gap, residual, and metric inexactness

We introduce three inexactness criteria for approximate proximal steps.

Definition 8 (Inexact proximal step). *Given $x^k \in \mathbf{dom}(h)$, $\beta_k > 0$, and $p > 1$, a point $x^{k+1} \in \mathbf{dom}(h)$ is called*

(a) *a model-value gap inexact proximal step with accuracy $\varepsilon_k \geq 0$ if*

$$Q_k(x^{k+1}) \leq \inf_{y \in \mathbb{R}^n} Q_k(y) + \varepsilon_k. \tag{11}$$

(b) *a residual inexact proximal step with tolerance $\eta_k \geq 0$ if h is locally Lipschitz around x^{k+1} and there exists $r^{k+1} \in \mathbb{R}^n$ such that*

$$r^{k+1} \in \partial^C h(x^{k+1}) + \frac{1}{\beta_k} J_p(x^{k+1} - x^k), \qquad \textit{with} \qquad \left\|r^{k+1}\right\| \leq \eta_k. \tag{12}$$

(c) *a metric inexact proximal step with accuracy $\delta_k \geq 0$ if*

$$\mathbf{dist}\big(x^{k+1}, T_{\beta_k,p}(x^k)\big) \leq \delta_k. \tag{13}$$

The three notions serve different purposes. The model-value gap is convenient for descent and global convergence, residual inexactness provides a stationarity interface, and metric inexactness transfers local properties of the exact proximal map to inexact iterates. They are not equivalent without additional regularity

Algorithm 1 Inexact HiPPA (Inexact High-Order Proximal-Point Algorithm)

1: **Input:** $x^0 \in \mathbf{dom}(h)$, order $p > 1$, bounds $0 < \beta_{\mathbf{min}} \leq \beta_{\mathbf{max}} < \infty$, tolerances $\{\varepsilon_k\}_{k\in\mathbb{N}_0}$.
2: **while** stopping criteria are not satisfied **do**
3: Choose $\beta_k \in [\beta_{\mathbf{min}}, \beta_{\mathbf{max}}]$.
4: Compute $x^{k+1} \in \mathbf{dom}(h)$ such that

$$Q_{x^k,\beta_k}(x^{k+1}) \leq \inf_y Q_{x^k,\beta_k}(y) + \varepsilon_k.$$

5: **end while**

3.2 Inexact high-order proximal-point algorithm

We now state Inexact HiPPA in Algorithm 1. This method is analyzed under the model-value gap condition (11). Section 5 states separately which interface is used in each experiment.

We now analyze Algorithm 1.

Theorem 9 (Global convergence under model-value gap inexactness). *Let Assumption 7 hold and let $\{x^k\}_{k\in\mathbb{N}_0}$ be generated by Algorithm 1. Suppose that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Then the following statements hold.*

- **(a)** *For every $k \in \mathbb{N}_0$, $Q_k(x^{k+1}) \leq Q_k(x^k) + \varepsilon_k$ and $h(x^{k+1}) \leq h(x^k) + \varepsilon_k$;*
- **(b)** *We have $\sum_{k=0}^{\infty} \left\|x^{k+1} - x^k\right\|^p < \infty$ and in particular, $\left\|x^{k+1} - x^k\right\| \to 0$.*
- **(c)** *If $\{x^k\}_{k\in\mathbb{N}_0}$ is bounded, then $h(x^k) \to h^\star$ and every cluster point belongs to $X^\star$. If $X^\star = \{\overline{x}\}$, then $x^k \to \overline{x}$.*
- **(d)** *If h is (κ,γ)-strongly quasar-convex with $\gamma > 0$, then $\{x^k\}_{k\in\mathbb{N}_0}$ is bounded, $X^\star = \{\overline{x}\}$, and $x^k \to \overline{x}$.*

The following assumption describes the local stability of the inexact proximal step.

Assumption 10 (Local inexact stability). *Set $r_k := \left\|x^k - \overline{x}\right\|$. Assume that h is (κ,γ)-strongly quasar-convex with respect to $\overline{x}$, with $\gamma > 0$.*

- **(a)** *For $p \in (1, 2]$, there exist constants $\eta \in (0, 1)$ and $C > 0$ such that, for all sufficiently large k,*

$$r_{k+1}^2 \leq \eta r_k^2 + C\varepsilon_k. \tag{14}$$

- **(b)** *For $p > 2$, there exist constants $a, b > 0$ such that, for all sufficiently large k,*

$$r_{k+1} \leq a r_k^{p-1} + b\sqrt{\varepsilon_k}. \tag{15}$$

Proposition 32 in the appendix gives a sufficient route to Assumption 10 from a local rate of the exact proximal map and a metric inexactness estimate.

Theorem 11 (Convergence rates of inexact HiPPA). *Let Assumption 10 hold and let $x^k \to \overline{x}$. Then, for all sufficiently large k, there exist $A, B > 0$ such that*

$$\begin{cases} \left\|x^k - \overline{x}\right\| \leq A\rho^k, & p \in (1,2], \ \text{if } \exists \rho \in (\sqrt{\eta}, 1), E > 0, \text{ s.t. } \varepsilon_k \leq E\,\rho^{2k}, \\ \|x^{k+1} - \overline{x}\| \leq B\|x^k - \overline{x}\|^{p-1}, & p > 2, \ \text{if } \exists M > 0, \delta > 0, \text{ s.t. } \sqrt{\varepsilon_k} \leq M\left\|x^k - \overline{x}\right\|^{p-1+\delta}. \end{cases}$$

Moreover, suppose that h is locally Lipschitz near $\overline{x}$. If $1 < p \leq 2$, then there exists $C_h > 0$ such that

$$h(x^k) - h^\star \leq C_h \rho^k, \tag{16}$$

for all sufficiently large k. If $p > 2$, then there exists $L_h > 0$ such that

$$h(x^k) - h^\star \leq L_h \left\|x^k - \overline{x}\right\|,$$

for all sufficiently large k, while the distance sequence satisfies

$$\left\|x^{k+1} - \overline{x}\right\| \leq B\left\|x^k - \overline{x}\right\|^{p-1}.$$

Remark 12. *For $p \in (1, 2]$, Theorem 11 gives R-linear convergence of the iterates. For $p > 2$, it gives*

$$\frac{\left\|x^{k+1} - \overline{x}\right\|}{\|x^k - \overline{x}\|} \to 0,$$

and hence the convergence is superlinear.

The local rates yield the following outer-iteration bounds, excluding proximal subproblem costs.

Theorem 13 (Outer-iteration complexity). *Define, for $\varepsilon > 0$,*

$$N_r(\varepsilon) := \mathbf{min}\left\{N \in \mathbb{N}_0 : \ \|x^k - \overline{x}\| \le \varepsilon \text{ for all } k \ge N\right\}.$$

Under the assumptions of Theorem 11, the following outer-iteration complexity bounds hold:

$$N_r(\varepsilon) = \begin{cases} \mathcal{O}\left(\log(\varepsilon^{-1})\right), & 1 < p \le 2, \quad \varepsilon_k = \mathcal{O}(\rho^{2k}), \\ \mathcal{O}\left(\log\log(\varepsilon^{-1})\right), & p > 2, \quad \varepsilon_k = \mathcal{O}\left(r_k^{2p-2+2\delta}\right), \end{cases} \tag{17}$$

If h is locally Lipschitz near $\overline{x}$, then the same outer-iteration orders guarantee $h(x^k) - h(\overline{x}) \le \varepsilon$.

4 Corruption, recovery, and a practical residual interface

The strong quasar-convexity statements in Section 2 concern the clean realizable models. We now record what can still be said when the capped feature-alignment targets are corrupted, without claiming that the corrupted empirical objective inherits the clean global geometry.

Let us write

$$Z_t = W^\star Z_s + E, \tag{18}$$

where the columns e_i of E encode mismatched or noisy supervision. Let $S = Z_s Z_s^\top / N$, $\lambda_- = \lambda_{\mathbf{min}}(S) > 0$, $\delta_2 = \|E\|_F / \sqrt{N}$, and

$$A_E = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{min}\{\|e_i\|, \tau\} + \mu \|e_i\|^2\right).$$

Proposition 14 (Approximate-realizability stability). *Every global minimizer $\widehat{W}$ of the corrupted capped objective satisfies*

$$\left\|\widehat{W} - W^\star\right\|_F \le \frac{\delta_2 + \sqrt{A_E/\mu}}{\sqrt{\lambda_-}}, \tag{19}$$

$$\frac{1}{2N}\left\|(\widehat{W} - W^\star) Z_s\right\|_F^2 \le \frac{1}{2}\left(\delta_2 + \sqrt{A_E/\mu}\right)^2. \tag{20}$$

For bounded replacements affecting a fraction ϱ, these bounds scale as $\mathcal{O}(\sqrt{\varrho})$ in parameter error and $\mathcal{O}(\varrho)$ in clean alignment MSE.

For a local recovery result, let $\mathcal{I} = \{i : e_i = 0\}$ and $\mathcal{E} = \{i : e_i \ne 0\}$, assume $\|e_i\| \ne \tau$ for $i \in \mathcal{E}$, and restrict attention to a neighborhood in which this cap pattern does not change. Define

$$L_0(U) := \frac{1}{N} \sum_{i \in \mathcal{I}} \|U z_{s,i}\|, \qquad a_i := 2\mu + \frac{\mathbf{1}\{\|e_i\| < \tau\}}{\|e_i\|}, \qquad B_E = \frac{1}{N} \sum_{i \in \mathcal{E}} a_i e_i z_{s,i}^\top.$$

Theorem 15 (Fixed-pattern local recovery). *Within the fixed-pattern neighborhood, $W^\star$ is a local minimizer if and only if $B_E \in \partial L_0(0)$, equivalently if*

$$\alpha_E := \inf_{\|U\|_F = 1} \left(L_0'(0; U) - \langle B_E, U \rangle_F\right) \ge 0.$$

Appendix D gives an explicit conic sufficient certificate, a quantitative margin, and the corresponding local proximal-identification result.

The model-value gap in (11) is generally unavailable computationally because $\mathbf{inf}\, Q_k$ is unknown. In the corrupted experiment, we therefore use the separate residual-and-descent rule

$$Q_k(x^{k+1}) \le h(x^k), \qquad \mathbf{dist}\left(0, \partial^C h(x^{k+1}) + \beta_k^{-1} J_p(x^{k+1} - x^k)\right) \le \eta_k. \tag{21}$$

For bounded accepted sequences, $\eta_k \to 0$ and (9) imply vanishing steps and Clarke-critical cluster points. Proposition 29 gives the precise statement. Global optimality follows only when quasar geometry is also available.

5 APPLICATION TO ROBUST FEATURE-ALIGNMENT DISTILLATION

5.1 LOCAL THEOREM VALIDATION AND ORDER SENSITIVITY FOR $p > 2$

We test the local theory on the capped-quadratic alignment model of Proposition 4, whose proximal properties are verified in Proposition 22. Using the official torchvision ResNet-18 with ImageNet-1K weights (Paszke et al., 2019), we extract 512-dimensional penultimate features from a deterministic balanced set of 5,000 CIFAR-100 training images (Krizhevsky, 2009) and pair them with the same network's logits. After adding an intercept, the 513×5000 design matrix has numerical rank 513, and the logits agree with the affine classifier head to relative Frobenius error 9.29×10^{-8}. Thus the realizability and identifiability assumptions are directly satisfied.

We set $\mu = 0.1$, choose τ as the median logit norm, take $\beta = 1/m$, and use the same L-BFGS lower solver from five prescribed perturbations initialized at half of the conservative quadratic-basin radius. For $p = 2$, $\varepsilon_k = 10^{-6}(0.75)^{2k}$, while for $p > 2$,

$$\sqrt{\varepsilon_k} = 0.05\, r_k^{p-1+0.25}, \qquad r_k := \left\|W^k - W^\star\right\|_F .$$

Because $W^\star$ and the proximal model minimum are available in this controlled construction, the model-value gap can be certified directly. Full protocol and diagnostic details are given in Appendix E.

At the common endpoint $r_k \le 10^{-3}$, $p = 2$ requires 14.8 ± 0.45 outer iterations, compared with 4 for $p = 3$ and 3 for $p = 4$. Increasing the order to $p = 5$ or $p = 6$ gives no further reduction in outer iterations, while the mean lower-level work rises from 416.2 evaluations for $p = 4$ to 776.2 for $p = 6$. All reported runs satisfy their prescribed model-value tolerances and remain inside the verified quadratic basin. Thus higher order improves the local outer convergence, but the computational benefit saturates as the lower-level accuracy requirement becomes more demanding. Since the forcing schedules depend on p, this is a rate-validation and order-sensitivity study rather than a generic runtime ranking.

Table 1: Local capped-quadratic order sensitivity. Means over five perturbation directions at the endpoint $\left\|W^k - W^\star\right\|_F \le 10^{-3}$.

p	outer steps	lower-level evaluations
2	14.8 ± 0.45	571.2 ± 109.9
3	4.0 ± 0.0	529.8 ± 64.6
4	3.0 ± 0.0	416.2 ± 26.8
5	3.0 ± 0.0	523.0 ± 96.1
6	3.0 ± 0.0	776.2 ± 160.3

5.2 CORRUPTED REAL-FEATURE STRESS TEST

We next evaluate the capped alignment model under controlled target corruption. We use frozen CIFAR-100 representations from a fine-label ResNet-18 as source features and a coarse-label ResNet-18 as target features (Zenodo DOI 10.5281/zenodo.21131944). The fit set contains 5,000 images and evaluation uses the untouched 10,000-image test split. Source and target PCA32 and whitening transforms are fitted before corruption and then held fixed. For $\varrho \in \{.1, .2, .4\}$ and five seeds, a fraction ϱ of the target fit vectors is replaced by vectors from a different coarse class. This is a controlled gross-mismatch stress test rather than a model of arbitrary natural noise or distribution shift.

We use the same parameters for $p = 2$ and $p = 3$, with cap quantile .7, $\mu\tau = .05$, $\beta = .05$, residual tolerance

$$\eta_k = \frac{0.1}{(k+1)^{3/2}},$$

and at most 40 outer iterations. Each accepted step satisfies the residual-and-descent rule (21). The complete endpoint definition, certificate audit, and implementation details are reported in Appendix E.2.

Performance is assessed over 16 prescribed objective/stationarity tolerance pairs on each of the 15 corrupted instances. The reference value h_{ref} is the smallest finite objective attained by the $p = 2$

and $p = 3$ trajectories on the same instance and is not assumed to be the global optimum. Across the resulting 240 instance–endpoint combinations, $p = 3$ reaches every prescribed endpoint, whereas $p = 2$ reaches none within the 40-iteration budget under the same fixed parameters. Table 2 reports test MSE at the strictest endpoint $(\varepsilon_f, \varepsilon_s) = (5 \times 10^{-4}, 5 \times 10^{-3})$.

Table 2: Real-feature corruption results averaged over five seeds at the strictest prescribed endpoint.

ϱ	LS MSE	Group-LAD MSE	$p = 3$ MSE	lower sweeps
.10	.600994	.601645	.610063	128.8
.20	.615811	.612848	.612575	124.8
.40	.665884	.654759	.638229	144.8

At the highest corruption level, capped fitting with $p = 3$ yields lower test MSE than the LS and Group-LAD baselines, whereas at lighter corruption the simpler baselines are competitive or better. Under the common fixed protocol, $p = 3$ is also certified substantially more readily than $p = 2$. These observations are specific to the present data, corruption mechanism, and solver configuration and are not a universal ranking of proximal orders or optimization methods.

6 Discussion

We connected nonsmooth robust representation-alignment models with inexact high-order proximal optimization. For the clean realizable models in Section 2, the objectives admit explicit strong-quasar geometry. Inexact HiPPA converges globally under summable model-value inexactness. Under local stability and suitable forcing, it converges linearly for $1 < p \leq 2$ and superlinearly with Q-order $p-1$ for $p > 2$. The capped-quadratic model makes these local conditions explicit, and the numerical results show clear gains from $p > 2$ followed by diminishing returns at larger orders.

The clean strong-quasar geometry is not assumed under target corruption, which is treated through separate stability and recovery results. The model-value gap criterion is directly certifiable in the controlled capped-quadratic experiment, while the corrupted experiment uses the residual-and-descent rule in (21). The complexity bounds count outer proximal iterations only, and therefore the numerical study reflects a tradeoff between higher local order and increasing lower-level work rather than a universally optimal choice of p.

AI use statement

In this work, we used generative AI tools for language editing.

References

M. Ahookhosh and Y. Nesterov. High-order methods beyond the classical complexity bounds: Inexact high-order proximal-point methods. *Mathematical Programming*, 208:365–407, 2024. doi: 10.1007/s10107-023-02041-4.

M. Ahookhosh and Y. Nesterov. High-order methods beyond the classical complexity bounds: Inexact high-order proximal-point methods with segment search, 2025.

M. Ahookhosh, J. de Brito, A. Kabgani, F. Lara, and J. Yuan. Quasar-convex optimization: Fundamental properties and high-order proximal-point methods, 2026a.

M. Ahookhosh, A. Iusem, A. Kabgani, and F. Lara. Asymptotic convergence analysis of high-order proximal-point methods beyond sublinear rates. *SIAM Journal on Optimization (accepted)*, pp. 1–30, 2026b.

M. Barré, A. B. Taylor, and F. Bach. Principled analyses and design of first-order methods with inexact proximal operators. *Mathematical Programming*, 201:185–230, 2023. doi: 10.1007/s10107-022-01903-7.

J. T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4326–4334, 2019.

F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

A. Csiszárik, P. Kőrösi-Szabó, Á. K. Matszangosz, G. Papp, and D. Varga. Similarity and matching of neural network representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5656–5668, 2021.

D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

J. de Brito, F. Lara, and D. Liu. Extending linear convergence of the proximal point algorithm: The quasar-convex case, 2025.

P. E. Dvurechensky. A gradient method with inexact oracle for composite nonconvex optimization. *Computer Research and Modeling*, 14(2):321–334, 2022.

S. Guminov, A. Gasnikov, and I. Kuruzov. Accelerated methods for weakly-quasi-convex optimization problems. *Computational Management Science*, 20:36, 2023.

N. Hadjisavvas, S. Komlósi, and S. Schaible (eds.). *Handbook of Generalized Convexity and Generalized Monotonicity*. Springer-Verlag, 2005.

M. Hardt, T. Ma, and B. Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

J. Hermant, J.-F. Aujol, C. Dossal, and A. Rondepierre. Study of the behaviour of Nesterov accelerated gradient in a non-convex setting: The strongly quasar-convex case, 2024.

O. Hinder, A. Sidford, and N. Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1894–1938, 2020.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

A. Kabgani and M. Ahookhosh. First-order majorization-minimization meets high-order majorant: Boosted inexact high-order forward-backward method, 2025a.

A. Kabgani and M. Ahookhosh. ItsDEAL: Inexact two-level smoothing descent algorithms for weakly convex optimization, 2025b.

A. Kabgani and M. Ahookhosh. Moreau envelope and proximal-point methods under the lens of high-order regularization. *Set-Valued and Variational Analysis*, 33:47, 2025c.

A. Kabgani and M. Ahookhosh. ItsOPT: An inexact two-level smoothing framework for nonconvex optimization via high-order Moreau envelope. *SIAM Journal on Optimization (accepted)*, pp. 1–30, 2026a.

A. Kabgani and M. Ahookhosh. On fundamental properties of high-order forward-backward envelope. *Journal of Optimization Theory and Applications*, 210(1):1–34, 2026b. doi: 10.1007/s10957-026-03047-6.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

F. Lara. On strongly quasiconvex functions: Existence results and proximal point algorithms. *Journal of Optimization Theory and Applications*, 192(3):891–911, 2022.

Y.-J. Liu, D. Sun, and K.-C. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical Programming*, 133:399–436, 2012.

S. Maekawa, M. Aminnaseri, P. Pezeshkpour, and E. Hruschka. Align then train: Efficient retrieval adapter learning, 2026.

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Revue Française d'Informatique et de Recherche Opérationnelle*, 4(Ser. R-3):154–158, 1970.

B. Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante: Cas de l'application prox. *Comptes Rendus de l'Académie des Sciences de Paris, Série A-B*, 274: A163–A165, 1972.

Y. Nesterov. Inexact accelerated high-order proximal-point methods. *Mathematical Programming*, 197:1–26, 2023.

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

Y.-M. Pun and I. Shames. Online non-stationary stochastic quasar-convex optimization, 2024.

Morteza Rahimi, Susan Ghaderi, Yves Moreau, and Masoud Ahookhosh. Projected subgradient methods for paraconvex optimization: Application to robust low-rank matrix recovery. *Journal of Global Optimization*, pp. 1–40, 2026.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2009. 3rd printing.

A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex Analysis*, 19(4):1167–1192, 2012.

M. V. Solodov and B. F. Svaiter. A unified framework for some inexact proximal point algorithms. *Numerical Functional Analysis and Optimization*, 22(7–8):1013–1035, 2001.

H. Song, M. Kim, D. Park, Y. Shin, and J. G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8135–8153, 2023.

S. Sra. Scalable nonconvex inexact proximal splitting. In *Advances in Neural Information Processing Systems*, volume 25, pp. 539–547, 2012.

C. Wang, Q. Ye, P. Luo, N. Ye, and L. Fu. Robust capped L_1-norm twin support vector machine. *Neural Networks*, 114:47–59, 2019.

Huajun Wang, Hongwei Zhang, and Wenqian Li. Sparse and robust support vector machine with capped squared loss for large-scale pattern classification. *Pattern Recognition*, 153:110544, 2024.

J.-K. Wang and A. Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization. In *International Conference on Learning Representations*, 2023.

L. Yang, Y. Wang, and G. Li. Robust capped L_1-norm projection twin support vector machine. *Journal of Industrial and Management Optimization*, 19(8):5797–5815, 2023. doi: 10.3934/jimo.2022195.

CONTENTS

1 Introduction **1**

1.1 Related work . . . 2

1.2 Notation . . . 3

2 Motivational examples from robust learning **3**

2.1 Stabilized capped residual loss . . . 3

2.2 Anisotropic robust residuals . . . 4

3 Inexact high-order proximal-point algorithm **5**

3.1 Model-value gap, residual, and metric inexactness . . . 5

3.2 Inexact high-order proximal-point algorithm . . . 6

4 Corruption, recovery, and a practical residual interface **7**

5 Application to robust feature-alignment distillation **8**

5.1 Local theorem validation and order sensitivity for $p > 2$ . . . 8

5.2 Corrupted real-feature stress test . . . 8

6 Discussion **9**

A Preliminaries and further properties of quasar-convexity **14**

A.1 Supplemental Notation . . . 14

A.2 Further properties of quasar-convexity . . . 14

B Proofs of Section 2 **16**

B.1 Proof of Proposition 2 . . . 16

B.2 Proof of Example 3 . . . 17

B.3 Capped-quadratic alignment: proofs and proximal properties . . . 18

B.4 Proof of Proposition 5 . . . 20

B.5 Proof of Example 6 . . . 22

B.6 Identifiable feature spans . . . 23

C Proofs of Section 3 and further properties of inexact high-order oracle **23**

C.1 Auxiliary one-step estimates . . . 24

C.2 Restricted capped-loss residual certificate . . . 26

C.3 Proof of Theorem 9 . . . 26

C.4 Sufficient conditions for Assumption 10 . . . 27

C.5 Proof of Theorem 11 . . . 28

C.6 Proof of Theorem 13 . . . 29

D Proofs for the corruption and recovery results **30**

E Additional numerical details **31**

E.1 Order-sensitivity diagnostics 31

E.2 Corrupted-feature protocol and certificate audit 32

F Further examples: Strong quasar-convexity beyond local Lipschitzness **33**

F.1 Nonconvex robust LAD regression 34

F.2 Nonconvex robust multi-task regression 34

F.3 Nonconvex robust matrix sensing 35

A PRELIMINARIES AND FURTHER PROPERTIES OF QUASAR-CONVEXITY

A.1 SUPPLEMENTAL NOTATION

A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *lower semicontinuous* (lsc henceforth) at $\overline{x} \in \mathbb{R}^n$ if for every sequence $\{x^k\}_{k\in\mathbb{N}_0} \subseteq \mathbb{R}^n$ with $x^k \to \overline{x}$, one has $h(\overline{x}) \leq \liminf_{k\to+\infty} h(x^k)$. A point $\widehat{x} \in \mathbb{R}^n$ is called a *cluster point* of a sequence $\{x^k\}_{k\in\mathbb{N}_0}$ if there exists an infinite subset $J \subseteq \mathbb{N}_0$ such that $x^j \to \widehat{x}$ as $j \to \infty$ with $j \in J$.

A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ with convex domain is said to be:

(a) *strongly convex* on $\mathbf{dom}(h)$ if there exists $\gamma > 0$ such that, for all $x, y \in \mathbf{dom}(h)$ and all $\lambda \in [0, 1]$,

$$h(\lambda y + (1-\lambda)x) \leq \lambda h(y) + (1-\lambda)h(x) - \lambda(1-\lambda)\frac{\gamma}{2}\|x-y\|^2, \tag{22}$$

(b) *strongly star-convex* with respect to $\overline{x} \in \mathbf{arg\,min}_{\mathbb{R}^n} h$ if there exists $\gamma > 0$ such that, for all $x \in \mathbf{dom}(h)$ and all $\lambda \in [0, 1]$,

$$h(\lambda\overline{x} + (1-\lambda)x) \leq \lambda h(\overline{x}) + (1-\lambda)h(x) - \lambda(1-\lambda)\frac{\gamma}{2}\|x-\overline{x}\|^2. \tag{23}$$

Setting $\gamma = 0$ in (22) (resp. (23) yields the notions of *convexity* (resp. *star-convexity*). In particular, every (strongly) convex function is (strongly) star-convex with respect to any point $\overline{x} \in \mathbf{dom}(h)$, whereas the converse implication does not hold in general (Hadjisavvas et al., 2005; Lara, 2022; Nesterov & Polyak, 2006).

Let h be locally Lipschitz around $x \in \mathbf{dom}(h)$. The *Clarke subdifferential* of h at x is defined by

$$\partial^C h(x) := \mathrm{cl}\,\mathrm{conv}\left\{v \in \mathbb{R}^n : \exists\, x^j \to x,\ h \text{ is differentiable at } x^j,\ \nabla h(x^j) \to v\right\},$$

where $\mathrm{cl}\,\mathrm{conv}$ denotes the *closed convex hull*. If, in addition, h is convex, then the Clarke subdifferential coincides with the standard convex subdifferential. A point $x \in \mathbf{dom}(h)$ is *Clarke critical* if $0 \in \partial^C h(x)$. We use the standard fact that, for locally Lipschitz functions, the Clarke subdifferential is locally bounded and outer semicontinuous (Clarke, 1990; Rockafellar & Wets, 2009).

A.2 FURTHER PROPERTIES OF QUASAR-CONVEXITY

We first recall a first-order characterization of strong quasar-convexity in terms of the Clarke subdifferential. This form is useful because it converts the geometric definition into inequalities involving generalized gradients.

Fact 16 (Clarke first-order characterization). *(Ahookhosh et al., 2026a, Theorem 31) Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function with convex domain, and let $\overline{x} \in \mathbf{arg\,min}_{\mathbb{R}^n} h$. Assume that h is locally Lipschitz on $\mathbf{dom}(h)$. Then h is (κ, γ)-strongly quasar-convex with respect to $\overline{x}$, where $\gamma \geq 0$, if and only if*

$$h(\overline{x}) \geq h(x) + \frac{1}{\kappa}\langle v, \overline{x} - x\rangle + \frac{\gamma}{2}\|x-\overline{x}\|^2, \qquad \forall\, x \in \mathbf{dom}(h),\ \forall\, v \in \partial^C h(x). \tag{24}$$

The next consequence explains why the quasar-convexity guarantees a benign nonsmooth landscape: Clarke criticality cannot occur away from the set of global minimizers.

Fact 17 (Clarke critical points are global minimizers). *(Ahookhosh et al., 2026a, Proposition 33) Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function with convex domain, and let $\overline{x} \in \mathbf{arg\,min}_{\mathbb{R}^n} h$. Assume that h is (κ, γ)-strongly quasar-convex with respect to $\overline{x}$, where $\gamma \geq 0$, and that h is locally Lipschitz on $\mathbf{dom}(h)$. If $0 \in \partial^C h(x)$, then $x \in \mathbf{arg\,min}_{\mathbb{R}^n} h$.*

When $\gamma > 0$, strong quasar-convexity also provides quantitative control near the minimizer. The subsequent growth condition and error-bound estimate will be used later to convert residual information into distance and function-value bounds.

Fact 18 (Quadratic growth and error bound). *(Ahookhosh et al., 2026a, Corollary 32) Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function and let $\overline{x} \in \mathbf{arg\,min}_{\mathbb{R}^n} h$. Assume that h is (κ, γ)-strongly*

quasar-convex with respect to $\overline{x}$, where $\kappa \in (0, 1]$ and $\gamma > 0$, and that h is locally Lipschitz on $\mathbf{dom}(h)$. Then, for every $x \in \mathbf{dom}(h)$, it holds that

$$h(x) - h(\overline{x}) \geq \frac{\kappa\gamma}{2(2-\kappa)} \|x - \overline{x}\|^2, \tag{25}$$

and

$$\|x - \overline{x}\| \leq \frac{2}{\kappa\gamma} \mathbf{dist}(0, \partial^C h(x)). \tag{26}$$

The next estimate is the main tool for converting an approximate stationarity condition into quantitative control of both the distance to the minimizer and the objective gap. In particular, it will be used with w equal to the proximal regularization term and r equal to the residual error of an inexact proximal step.

Proposition 19 (Residual-to-gap inequality). *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, locally Lipschitz on $\mathbf{dom}(h)$, and (κ, γ)-strongly quasar-convex with respect to $\overline{x} \in \arg\min_{\mathbb{R}^n} h$. Let $x \in \mathbf{dom}(h)$ and suppose that*

$$r \in \partial^C h(x) + w, \tag{27}$$

for some $r, w \in \mathbb{R}^n$. Then,

$$\kappa\left(h(x) - h(\overline{x})\right) + \frac{\kappa\gamma}{2} \|x - \overline{x}\|^2 \leq \langle w, \overline{x} - x\rangle + \|r\| \, \|x - \overline{x}\|. \tag{28}$$

If moreover $\gamma > 0$, then

$$\|x - \overline{x}\| \leq \frac{2}{\kappa\gamma} \left(\|w\| + \|r\|\right). \tag{29}$$

Furthermore,

$$h(x) - h(\overline{x}) \leq \frac{\|w\| + \|r\|}{\kappa} \|x - \overline{x}\| \leq \frac{2}{\kappa^2\gamma} \left(\|w\| + \|r\|\right)^2. \tag{30}$$

Moreover, the sharper bound

$$h(x) - h(\overline{x}) \leq \frac{1}{2\kappa^2\gamma} \left(\|w\| + \|r\|\right)^2 \tag{31}$$

holds.

Proof. Let us choose $v \in \partial^C h(x)$ such that $r = v + w$. It follows from (24) that

$$h(\overline{x}) \geq h(x) + \frac{1}{\kappa} \langle v, \overline{x} - x\rangle + \frac{\gamma}{2} \|x - \overline{x}\|^2.$$

Multiplying by κ and rearranging, it can be concluded that

$$\kappa\left(h(x) - h(\overline{x})\right) + \frac{\kappa\gamma}{2} \|x - \overline{x}\|^2 \leq \langle v, x - \overline{x}\rangle.$$

Substituting $v = r - w$ in this inequality results in

$$\kappa\left(h(x) - h(\overline{x})\right) + \frac{\kappa\gamma}{2} \|x - \overline{x}\|^2 \leq \langle r - w, x - \overline{x}\rangle = \langle r, x - \overline{x}\rangle + \langle w, \overline{x} - x\rangle.$$

Using the Cauchy–Schwarz inequality on the first term yields (28). Now, let us assume $\gamma > 0$. Since $h(x) \geq h(\overline{x})$, the first term on the left-hand side of (28) is nonnegative. Hence

$$\frac{\kappa\gamma}{2} \|x - \overline{x}\|^2 \leq \langle w, \overline{x} - x\rangle + \|r\| \, \|x - \overline{x}\| \leq \left(\|w\| + \|r\|\right) \|x - \overline{x}\|,$$

If $x = \overline{x}$, the claim is trivial. Otherwise, dividing by $\|x - \overline{x}\|$ gives (29). Dropping the quadratic term in (28) and applying the Cauchy–Schwarz inequality gives

$$\kappa\left(h(x) - h(\overline{x})\right) \leq \left(\|w\| + \|r\|\right) \|x - \overline{x}\|,$$

which implies the first inequality in (30). Combining this estimate with (29) yields the second inequality in (30). For the sharper quadratic bound, set $d := \|x - \overline{x}\|$ and $M := \|w\| + \|r\|$. From (28) and the Cauchy–Schwarz inequality, we obtain

$$\kappa\left(h(x) - h(\overline{x})\right) + \frac{\kappa\gamma}{2} d^2 \leq Md,$$

i.e.,

$$\kappa\left(h(x)-h(\overline{x})\right)\leq Md-\frac{\kappa\gamma}{2}d^2.$$

The right-hand side is maximized over $d\geq 0$ at $d=M/(\kappa\gamma)$, and the maximal value is $M^2/(2\kappa\gamma)$, i.e.,

$$h(x)-h(\overline{x})\leq\frac{1}{2\kappa^2\gamma}\left(\|w\|+\|r\|\right)^2,$$

which consequently adjusts the inequality (31). □

The previous estimate applies to any approximate stationarity condition. We now specialize it to the residual generated by a high-order proximal subproblem. This gives a one-step stability inequality that links the proximal residual, the regularization displacement $y-x$, and the distance from the new point y to the reference minimizer.

Proposition 20 (Stability of approximate proximal points). *Let $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ be proper, locally Lipschitz on $\mathbf{dom}(h)$, and (κ,γ)-strongly quasar-convex with respect to $\overline{x}\in\mathbf{arg\,min}_{\mathbb{R}^n}h$. Fix $x\in\mathbb{R}^n$, $\beta>0$, and $p>1$. Suppose that $y\in\mathbf{dom}(h)$ satisfies*

$$e\in\partial^C h(y)+\frac{1}{\beta}\|y-x\|^{p-2}(y-x),\tag{32}$$

for some $e\in\mathbb{R}^n$. Then,

$$\kappa\left(h(y)-h(\overline{x})\right)+\frac{\kappa\gamma}{2}\|y-\overline{x}\|^2\leq\frac{1}{\beta}\|y-x\|^{p-2}\langle y-x,\overline{x}-y\rangle+\|e\|\,\|y-\overline{x}\|.\tag{33}$$

Moreover,

$$\kappa\left(h(y)-h(\overline{x})\right)+\frac{\kappa\gamma}{2}\|y-\overline{x}\|^2\leq\frac{1}{\beta}\|x-y\|^{p-1}\|y-\overline{x}\|+\|e\|\,\|y-\overline{x}\|.\tag{34}$$

Proof. Let us choose $v\in\partial^C h(y)$ such that

$$e=v+\frac{1}{\beta}\|y-x\|^{p-2}(y-x).$$

Applying the Clarke characterization (24) at y and rearranging ensures

$$\kappa\left(h(y)-h(\overline{x})\right)+\frac{\kappa\gamma}{2}\|y-\overline{x}\|^2\leq\langle v,y-\overline{x}\rangle.$$

Substituting $v=e-\frac{1}{\beta}\|y-x\|^{p-2}(y-x)$ into the last inequality, we come to

$$\kappa\left(h(y)-h(\overline{x})\right)+\frac{\kappa\gamma}{2}\|y-\overline{x}\|^2\leq\langle e,y-\overline{x}\rangle+\frac{1}{\beta}\|y-x\|^{p-2}\langle y-x,\overline{x}-y\rangle,$$

Applying the Cauchy-Schwarz inequality to the first term leads to (33). Employing the Cauchy-Schwarz inequality again on the proximal term results in (34). □

B Proofs of Section 2

B.1 Proof of Proposition 2

Proof. Local Lipschitz continuity follows from the Lipschitz continuity of $z\mapsto\|z\|$, the stability of locally Lipschitz functions under a pointwise minimum, and the smoothness of $z\mapsto\mu\|z\|^2$. The function is nonsmooth at $\overline{z}=0$ and along the sphere $\{z:\|z\|=\tau\}$. We next show that $\Phi_{\tau,\mu}$ is not star-convex with respect to $\overline{z}=0$. Fix a unit vector u and set $z=2\tau u$. Then

$$\Phi_{\tau,\mu}(\tau u)=\tau+\mu\tau^2,$$

whereas

$$\frac{1}{2}\Phi_{\tau,\mu}(2\tau u)=\frac{1}{2}(\tau+4\mu\tau^2)=\frac{\tau}{2}+2\mu\tau^2.$$

Since $\mu\tau < 1/2$, we have $\tau + \mu\tau^2 > \frac{\tau}{2} + 2\mu\tau^2$. Thus

$$\Phi_{\tau,\mu}\left(\frac{1}{2}z\right) > \frac{1}{2}\Phi_{\tau,\mu}(z),$$

which violates the star-convexity with respect to $\overline{z} = 0$. It remains to prove strong quasar-convexity. We use Clarke's first-order characterization (24). We claim that, with $\gamma = \mu$ and $\kappa = \frac{\mu\tau}{1+2\mu\tau}$, for every $z \in \mathbb{R}^d$ and every $v \in \partial^C \Phi_{\tau,\mu}(z)$,

$$0 \geq \Phi_{\tau,\mu}(z) + \frac{1}{\kappa}\langle v, -z\rangle + \frac{\gamma}{2}\|z\|^2. \tag{35}$$

For $z = 0$, this inequality is immediate. Let $z \neq 0$ be arbitrary. Set $r := \|z\|$, and write $u = z/r$. If $0 < r < \tau$, then $\Phi_{\tau,\mu}(z) = \|z\| + \mu\|z\|^2$ is continuously differentiable at z, and $\partial^C \Phi_{\tau,\mu}(z) = \{v_1 := (1 + 2\mu r)u\}$. Since $0 < \kappa \leq 1$,

$$\langle v_1, z\rangle = r + 2\mu r^2 \geq \kappa(r + \mu r^2) + \frac{\kappa\mu}{2}r^2 = \kappa\left(\Phi_{\tau,\mu}(z) + \frac{\gamma}{2}\|z\|^2\right).$$

Hence, the inequality (35) holds in this case. If $r > \tau$, then $\Phi_{\tau,\mu}(z) = \tau + \mu\|z\|^2$ is continuously differentiable at z, and $\partial^C \Phi_{\tau,\mu}(z) = \{v_2 := 2\mu r u\}$, and consequently we come to

$$\langle v_2, z\rangle = 2\mu r^2.$$

It remains to consider the kink $r = \tau$. In this case, the Clarke subdifferential is the convex hull of the two radial limits, i.e., $\partial^C \Phi_{\tau,\mu}(z) = [2\mu r^2, r + 2\mu r^2]$, i.e., the smallest possible value of $\langle v, z\rangle$ is $2\mu\tau^2$. Consequently, it is sufficient to verify the estimate for $r \geq \tau$ using the lower radial slope. We have

$$\mu\tau\left(2 - \frac{3\kappa}{2}\right) \geq \frac{\mu\tau}{1 + 2\mu\tau} = \kappa.$$

Since $r \geq \tau$, this implies

$$\mu\left(2 - \frac{3\kappa}{2}\right)r^2 \geq \kappa\tau.$$

Consequently, for every $v \in \partial^C \Phi_{\tau,\mu}(z)$ with $r \geq \tau$,

$$\langle v, z\rangle \geq 2\mu r^2 \geq \kappa(\tau + \mu r^2) + \frac{\kappa\mu}{2}r^2 = \kappa\left(\Phi_{\tau,\mu}(z) + \frac{\gamma}{2}\|z\|^2\right).$$

Equivalently, (35) holds. By Fact 16, $\Phi_{\tau,\mu}$ is (κ, γ)-strongly quasar-convex with respect to $\overline{z} = 0$. □

B.2 Proof of Example 3

Proof. Let us define

$$R_i(W) := Wz_{s,i} - z_{t,i}, \qquad i = 1, \ldots, N.$$

By the realizability assumption $W^\star Z_s = Z_t$, we have $R_i(W^\star) = 0$. Moreover,

$$R_i(W) = (W - W^\star)z_{s,i}.$$

The local Lipschitz continuity of h_{KD} follows from Proposition 2. The nonsmoothness comes from the fact that $R_i(W^\star) = 0$ for all i and the residual loss $\Phi_{\tau,\mu}$ is nonsmooth at the origin. In particular, along any direction ΔW with $\Delta W Z_s \neq 0$, the restriction $t \mapsto h_{\mathrm{KD}}(W^\star + t\Delta W)$ has a kink at $t = 0$.

It remains to verify the strong quasar-convexity. Let $W \in \mathbb{R}^{d_t \times d_s}$ and $\lambda \in [0, 1]$, and set

$$W_\lambda := \lambda W^\star + (1 - \lambda)W.$$

For every i, using $W^\star z_{s,i} = z_{t,i}$, it holds that

$$R_i(W_\lambda) = W_\lambda z_{s,i} - z_{t,i} = (1 - \lambda)(Wz_{s,i} - z_{t,i}) = (1 - \lambda)R_i(W).$$

By Proposition 2, $\Phi_{\tau,\mu}$ is (κ, μ)-strongly quasar-convex with respect to $\overline{z} = 0$, where $\kappa = \frac{\mu\tau}{1+2\mu\tau}$. Therefore, for each i,

$$\Phi_{\tau,\mu}(R_i(W_\lambda)) \leq (1 - \kappa\lambda)\Phi_{\tau,\mu}(R_i(W)) - \lambda\left(1 - \frac{\lambda}{2 - \kappa}\right)\frac{\kappa\mu}{2}\|R_i(W)\|^2.$$

Averaging over $i = 1, \dots, N$ ensures

$$h_{\mathrm{KD}}(W_\lambda) \le (1 - \kappa\lambda) h_{\mathrm{KD}}(W) - \lambda \left(1 - \frac{\lambda}{2 - \kappa}\right) \frac{\kappa\mu}{2N} \sum_{i=1}^{N} \|R_i(W)\|^2.$$

Since $R_i(W) = (W - W^\star) z_{s,i}$, it can be concluded that

$$\sum_{i=1}^{N} \|R_i(W)\|^2 = \|(W - W^\star) Z_s\|_F^2.$$

In light of the fact that Z_s has full row rank, we get

$$\|(W - W^\star) Z_s\|_F^2 \ge \sigma_{\mathbf{min}}(Z_s)^2 \|W - W^\star\|_F^2,$$

which consequently implies

$$h_{\mathrm{KD}}(W_\lambda) \le (1 - \kappa\lambda) h_{\mathrm{KD}}(W) - \lambda \left(1 - \frac{\lambda}{2 - \kappa}\right) \frac{\kappa}{2} \left(\frac{\mu}{N} \sigma_{\mathbf{min}}(Z_s)^2\right) \|W - W^\star\|_F^2.$$

Since $h_{\mathrm{KD}}(W^\star) = 0$, this is precisely $\left(\kappa, \frac{\mu}{N}\sigma_{\mathbf{min}}(Z_s)^2\right)$-strong quasar-convexity with respect to $W^\star$. □

Remark 21 (**Robust feature-alignment distillation is not star-convex**). *Proposition 2 shows that the residual loss $\Phi_{\tau,\mu}$ is not star-convex with respect to the origin. This property does not automatically transfer to the averaged composed objective h_{KD}, because different residual terms may lie in different regions of the capped loss and the one-dimensional violation can be averaged out.*

A simple sufficient condition is the existence of a direction $\Delta W \in \mathbb{R}^{d_t \times d_s}$ such that

$$\sum_{i=1}^{N} \left[\Phi_{\tau,\mu}\left(\frac{1}{2}\Delta W z_{s,i}\right) - \frac{1}{2}\Phi_{\tau,\mu}(\Delta W z_{s,i})\right] > 0. \tag{36}$$

Under this condition, h_{KD} is not star-convex with respect to $W^\star$. Indeed, assume that there exists $\Delta W \in \mathbb{R}^{d_t \times d_s}$ satisfying (36). Set

$$W := W^\star + \Delta W.$$

Since $W^\star Z_s = Z_t$, we have

$$W z_{s,i} - z_{t,i} = \Delta W z_{s,i}, \qquad i = 1, \dots, N.$$

Moreover,

$$\left(\frac{1}{2}W^\star + \frac{1}{2}W\right) z_{s,i} - z_{t,i} = \frac{1}{2}\Delta W z_{s,i}.$$

Therefore, from (36)

$$h_{\mathrm{KD}}\left(\frac{1}{2}W^\star + \frac{1}{2}W\right) = \frac{1}{N}\sum_{i=1}^{N} \Phi_{\tau,\mu}\left(\frac{1}{2}\Delta W z_{s,i}\right) > \frac{1}{2N}\sum_{i=1}^{N} \Phi_{\tau,\mu}(\Delta W z_{s,i}) = \frac{1}{2} h_{\mathrm{KD}}(W).$$

Since $h_{\mathrm{KD}}(W^\star) = 0$, the star-convexity inequality with center $W^\star$ would require

$$h_{\mathrm{KD}}\left(\frac{1}{2}W^\star + \frac{1}{2}W\right) \le \frac{1}{2}h_{\mathrm{KD}}(W^\star) + \frac{1}{2}h_{\mathrm{KD}}(W) = \frac{1}{2}h_{\mathrm{KD}}(W),$$

which contradicts the strict inequality above. Hence, h_{KD} is not star-convex with respect to $W^\star$.

B.3 Capped-quadratic alignment: proofs and proximal properties

Proof of Proposition 4. For $r = \|z\|$, the scalar loss has the piecewise form

$$\Phi^{(2)}_{\tau,\mu}(z) = \begin{cases} \frac{1+\mu}{2} r^2, & r \le \tau, \\ \frac{\tau^2}{2} + \frac{\mu}{2} r^2, & r \ge \tau. \end{cases}$$

Hence, it is locally Lipschitz and nonsmooth on $\|z\| = \tau$, with

$$\partial^C \Phi^{(2)}_{\tau,\mu}(z) = \begin{cases} \{(1+\mu)z\}, & 0 < r < \tau, \\ \{az : a \in [\mu, 1+\mu]\}, & r = \tau, \\ \{\mu z\}, & r > \tau, \end{cases}$$

and $\partial^C \Phi^{(2)}_{\tau,\mu}(0) = \{0\}$. Set $\kappa = \frac{\mu}{1+\mu}$ and $\gamma = 1 + \mu$. By Fact 16, it is enough to verify

$$0 \geq \Phi^{(2)}_{\tau,\mu}(z) - \frac{1}{\kappa}\langle v, z\rangle + \frac{\gamma}{2}\|z\|^2, \qquad v \in \partial^C \Phi^{(2)}_{\tau,\mu}(z).$$

For $0 < r < \tau$ the right-hand side equals $-\frac{1+\mu}{\mu} r^2 \leq 0$. For $r \geq \tau$, the largest right-hand side is obtained at the smallest radial coefficient $a = \mu$, and it is bounded above by $\frac{\tau^2 - r^2}{2} \leq 0$. Thus $\Phi^{(2)}_{\tau,\mu}$ is (κ, γ)-strongly quasar-convex with respect to the origin. To see that it is not star-convex, take $r \in (\tau, \tau/\mu)$ and $s = \tau/r$. The star inequality at a point with norm r would require

$$\mu r^2 - (1+\mu)\tau r + \tau^2 = \mu(r - \tau)(r - \tau/\mu) \geq 0,$$

which is false. Hence the loss is non-star-convex and therefore nonconvex. For the alignment composition, write $D = W - W^\star$. Along the segment from W to $W^\star$, each residual equals $(1-\lambda)Dz_{s,i}$. Summing the scalar strong-quasar inequalities and using

$$\|DZ_s\|_F^2 \geq \sigma_{\mathbf{min}}(Z_s)^2 \|D\|_F^2$$

gives the claimed composite strong-quasar constant. Now suppose $\|W - W^\star\|_F < \tau/M_s$. Then

$$\|(W - W^\star)z_{s,i}\| \leq \|W - W^\star\|_F \|z_{s,i}\| < \tau$$

for every i, and (6) follows. In particular, the local Hessian operator is

$$H[D] = \frac{1+\mu}{N} DZ_s Z_s^\top,$$

whose smallest eigenvalue with respect to the Frobenius norm is

$$m = \frac{(1+\mu)\sigma_{\mathbf{min}}(Z_s)^2}{N}.$$

Thus h_{CQ} is m-strongly convex on the stated region. □

Proposition 22 (Local proximal properties of the capped-quadratic model). *Under the assumptions of Proposition 4, let*

$$\mathcal{B} := \left\{ W : \|W - W^\star\|_F < \frac{\tau}{M_s} \right\}.$$

For every $X \in \mathcal{B}$, $\beta > 0$, and $p > 1$, the exact proximal point $Y = T_{\beta,p}(X)$ is unique, belongs to $\mathcal{B}$, and satisfies

$$\|Y - W^\star\|_F \leq \frac{1}{\beta m}\|X - W^\star\|_F^{p-1}. \tag{37}$$

Moreover, for every $W \in \mathcal{B}$,

$$Q_{X,\beta}(W) - \mathbf{inf}\, Q_{X,\beta} \geq \frac{m}{2}\|W - Y\|_F^2. \tag{38}$$

Consequently, if $W^+ \in \mathcal{B}$ satisfies

$$Q_{X,\beta}(W^+) - \mathbf{inf}\, Q_{X,\beta} \leq \varepsilon,$$

then

$$\|W^+ - W^\star\|_F \leq \frac{1}{\beta m}\|X - W^\star\|_F^{p-1} + \sqrt{\frac{2\varepsilon}{m}}. \tag{39}$$

Proof. Fix $X \in \mathcal{B}$ and let $Y \in T_{\beta,p}(X)$. Set $D_Y := Y - W^\star$. For every $V \in \partial^C h_{\mathrm{CQ}}(Y)$, Clarke calculus yields coefficients $a_i \in [\mu, 1+\mu]$ such that

$$\langle V, D_Y \rangle_F = \frac{1}{N} \sum_{i=1}^{N} a_i \left\| D_Y z_{s,i} \right\|^2 \geq 0.$$

Since Y minimizes the proximal model, Fermat's rule gives

$$V + \frac{1}{\beta} J_p(Y - X) = 0,$$

for some such V. Taking the Frobenius inner product with D_Y gives

$$\langle X - Y, Y - W^\star \rangle_F \geq 0.$$

Therefore,

$$\|X - W^\star\|_F^2 = \|X - Y\|_F^2 + \|Y - W^\star\|_F^2 + 2\langle X - Y, Y - W^\star \rangle_F,$$

and hence

$$\|Y - W^\star\|_F \leq \|X - W^\star\|_F < \frac{\tau}{M_s}.$$

Thus, every exact proximal minimizer belongs to $\mathcal{B}$. Since the proximal model is m-strongly convex on $\mathcal{B}$, the exact proximal point is unique. Inside $\mathcal{B}$, first-order optimality becomes

$$H[Y - W^\star] = \frac{1}{\beta} J_p(X - Y).$$

Using the smallest eigenvalue m of H gives

$$m \left\| Y - W^\star \right\|_F \leq \frac{1}{\beta} \left\| X - Y \right\|_F^{p-1}.$$

The preceding norm identity also yields $\|X - Y\|_F \leq \|X - W^\star\|_F$, which proves (37). Finally, $Q_{X,\beta}$ is m-strongly convex on $\mathcal{B}$ and its unique global minimizer Y lies in $\mathcal{B}$. Hence, for every $W \in \mathcal{B}$,

$$Q_{X,\beta}(W) - \mathbf{inf}\, Q_{X,\beta} \geq \frac{m}{2} \left\| W - Y \right\|_F^2,$$

which is (38). If the model-value gap at W^+ is at most ε, then

$$\left\| W^+ - Y \right\|_F \leq \sqrt{\frac{2\varepsilon}{m}}.$$

Combining this estimate with (37) and the triangle inequality proves (39). □

B.4 Proof of Proposition 5

Proof. We first analyze the scalar radial core $f_{\tau,\mu}$. Since ψ_τ is continuous and piecewise C^1 with slope 1 on $[0, \tau)$ and slope $\frac{1}{2}\sqrt{\tau/r} \leq \frac{1}{2}$ on $(\tau, +\infty)$, it is globally Lipschitz on $[0, +\infty)$. Hence, $f_{\tau,\mu}$ is locally Lipschitz and nonsmooth at $r = \tau$. We claim that

$$\psi_\tau((1-\lambda)r) \leq \left(1 - \frac{\lambda}{2}\right) \psi_\tau(r), \qquad \forall r \geq 0,\ \forall \lambda \in [0,1]. \tag{40}$$

If $r \leq \tau$, then $(1-\lambda)r \leq \tau$ and (40) become $(1-\lambda)r \leq (1-\lambda/2)r$, which is obvious. If $r > \tau$ and $(1-\lambda)r > \tau$, then

$$\psi_\tau((1-\lambda)r) = \sqrt{1-\lambda}\, \psi_\tau(r) \leq \left(1 - \frac{\lambda}{2}\right) \psi_\tau(r).$$

It remains to consider the mixed case $r > \tau$ and $(1-\lambda)r \leq \tau$. Set $s := 1 - \lambda \in [0,1]$ and $a := \sqrt{r/\tau} > 1$. The condition $sr \leq \tau$ is equivalent to $s \leq a^{-2}$. We need to prove

$$sr \leq \frac{1+s}{2} \sqrt{\tau r},$$

or, equivalently,

$$sa \leq \frac{1+s}{2}.$$

The function $\varphi(s) := (1+s)/2 - sa$ decreases on $[0, a^{-2}]$, because $a > 1$. Hence, its minimum on this interval is attained at $s = a^{-2}$ and for each $s \in [0, a^{-2}]$,

$$\frac{1+s}{2} - sa = \varphi(s) \geq \varphi(a^{-2}) = \frac{1+a^{-2}}{2} - \frac{1}{a} = \frac{(a-1)^2}{2a^2} \geq 0.$$

Thus, we have $sa \leq (1+s)/2$, which proves (40) in the mixed case and therefore completes the proof of (40). Using (40), for all $r \geq 0$ and $\lambda \in [0,1]$ we get

$$f_{\tau,\mu}((1-\lambda)r) \leq \left(1 - \frac{\lambda}{2}\right) \psi_\tau(r) + \mu(1-\lambda)^2 r^2.$$

Since

$$\mu(1-\lambda)^2 = \mu\left(1 - \frac{\lambda}{2}\right) - \lambda\left(1 - \frac{\lambda}{2-\kappa}\right)\frac{\kappa\gamma}{2},$$

with $\kappa = 1/2$ and $\gamma = 6\mu$, it follows that

$$f_{\tau,\mu}((1-\lambda)r) \leq \left(1 - \frac{\lambda}{2}\right) f_{\tau,\mu}(r) - \lambda\left(1 - \frac{\lambda}{2-\kappa}\right)\frac{\kappa\gamma}{2}\, r^2.$$

Due to $f_{\tau,\mu}(0) = 0$, this is exactly the (κ,γ)-strong quasar-convexity of $f_{\tau,\mu}$ with respect to $\overline{z} = 0$.

This scalar inequality directly produces the strong quasar-convexity of $H_{\tau,\mu,q}$. Indeed, fix $z \neq 0$ and write $z = ru$ with $r = \|z\|$ and $u \in \mathbb{S}^{d-1}$. Since the segment from 0 to z maintains the same direction u, multiplying the scalar inequality by $q(u)$ gives

$$H_{\tau,\mu,q}((1-\lambda)z) \leq \left(1 - \frac{\lambda}{2}\right) H_{\tau,\mu,q}(z) - \lambda\left(1 - \frac{\lambda}{2-\kappa}\right)\frac{\kappa\gamma}{2}\, q(u)\, \|z\|^2.$$

Together with $q(u) \geq 1$ and $H_{\tau,\mu,q}(0) = 0$, this implies

$$H_{\tau,\mu,q}((1-\lambda)z) \leq \kappa\lambda H_{\tau,\mu,q}(0) + (1-\kappa\lambda) H_{\tau,\mu,q}(z) - \lambda\left(1 - \frac{\lambda}{2-\kappa}\right)\frac{\kappa\gamma}{2}\, \|z\|^2,$$

which is exactly (κ,γ)-strong quasar-convexity with respect to $\overline{z} = 0$.

Next, we prove the local Lipschitz continuity of $H_{\tau,\mu,q}$. Since q is C^1 on the compact sphere, it is Lipschitz and bounded there. Hence, there exist constants $L_q, M_q > 0$ such that

$$|q(u_1) - q(u_2)| \leq L_q \|u_1 - u_2\|, \qquad 1 \leq q(u) \leq M_q, \qquad \forall u, u_1, u_2 \in \mathbb{S}^{d-1}.$$

Let us write

$$H_{\tau,\mu,q}(z) = F_1(z) + F_2(z),$$

where

$$F_1(z) := \begin{cases} \psi_\tau(\|z\|)\, q\left(\dfrac{z}{\|z\|}\right), & z \neq 0, \\ 0, & z = 0, \end{cases} \qquad F_2(z) := \begin{cases} \mu\, \|z\|^2\, q\left(\dfrac{z}{\|z\|}\right), & z \neq 0, \\ 0, & z = 0. \end{cases}$$

Let $z_1, z_2 \neq 0$, set $r_i := \|z_i\|$ and $u_i := z_i/r_i$, and assume without loss of generality that $r_1 \geq r_2$. Since ψ_τ is 1-Lipschitz and $\psi_\tau(r_2) \leq r_2$, we have

$$\begin{aligned} |F_1(z_1) - F_1(z_2)| &\leq |\psi_\tau(r_1) - \psi_\tau(r_2)|\, |q(u_1)| + \psi_\tau(r_2)|q(u_1) - q(u_2)| \\ &\leq M_q|r_1 - r_2| + r_2 L_q \|u_1 - u_2\|. \end{aligned}$$

In addition, it holds that

$$\begin{aligned} r_2 \|u_1 - u_2\| = \|r_2 u_1 - r_2 u_2\| &\leq \|r_2 u_1 - r_1 u_1\| + \|r_1 u_1 - r_2 u_2\| \\ &= |r_1 - r_2| + \|z_1 - z_2\| \leq 2\|z_1 - z_2\|, \end{aligned}$$

i.e.,

$$|F_1(z_1) - F_1(z_2)| \le (M_q + 2L_q)\,\|z_1 - z_2\|\,.$$

If one of the points is 0, the same bound follows from $\psi_\tau(r) \le r$ and $q \le M_q$, i.e., F_1 is globally Lipschitz.

Now, let us fix $R > 0$ and let $z_1, z_2 \in \mathbf{B}(0,R) \setminus \{0\}$. Together with $r_1 + r_2 \le 2R$, the estimate above for $r_2\,\|u_1 - u_2\|$ leads to

$$\begin{aligned}|F_2(z_1) - F_2(z_2)| &\le \mu M_q |r_1^2 - r_2^2| + \mu r_2^2 L_q\,\|u_1 - u_2\| \\ &\le 2\mu M_q R |r_1 - r_2| + \mu r_2 L_q\,(r_2\,\|u_1 - u_2\|) \\ &\le 2\mu(M_q + L_q) R\,\|z_1 - z_2\|\,.\end{aligned}$$

If one point is zero, say $z_2 = 0$, then for $z_1 \in \mathbf{B}(0,R)$,

$$|F_2(z_1) - F_2(0)| = \mu\,\|z_1\|^2\, q(z_1/\,\|z_1\|) \le \mu M_q R\,\|z_1\| = \mu M_q R\,\|z_1 - z_2\|\,.$$

As a consequence, F_2 is Lipschitz on all bounded balls, i.e., locally Lipschitz on $\mathbb{R}^d$. Since $H_{\tau,\mu,q} = F_1 + F_2$, the map $H_{\tau,\mu,q}$ is locally Lipschitz on $\mathbb{R}^d$. Moreover, the function is nonsmooth because along any ray $z = ru$, the derivative jumps at $r = \tau$.

Finally, we show that $H_{\tau,\mu,q}$ is not star-convex with respect to $\overline{z} = 0$. Choose any unit vector u and set $z := 4\tau u$. Then $\|z\| > \tau$ and $\|z/4\| = \tau$, so

$$H_{\tau,\mu,q}(z/4) - \frac{1}{4} H_{\tau,\mu,q}(z) = q(u)\left(\tau + \mu\tau^2 - \frac{1}{4}\left(2\tau + 16\mu\tau^2\right)\right) = q(u)\tau\left(\frac{1}{2} - 3\mu\tau\right) > 0,$$

thanks to $\mu\tau < 1/8 < 1/6$. Hence, $H_{\tau,\mu,q}$ is not star-convex with the center $\overline{z} = 0$. □

B.5 Proof of Example 6

Proof. Let $\mathcal{L} : \mathbb{R}^{d_y \times (d_x+1)} \to \mathbb{R}^{d_y N}$ be defined as $\mathcal{L}(U) := \mathrm{vec}(U\widetilde{X})$. It follows from $\widetilde{A}^\star \widetilde{X} = Y$ that, for every $\widetilde{A} \in \mathbb{R}^{d_y \times (d_x+1)}$,

$$\mathrm{vec}(\widetilde{A}\widetilde{X} - Y) = \mathrm{vec}\big((\widetilde{A} - \widetilde{A}^\star)\widetilde{X}\big) = \mathcal{L}(\widetilde{A} - \widetilde{A}^\star),$$

i.e.,

$$h_{\mathrm{stitch}}(\widetilde{A}) = H_{\tau,\mu,q}\big(\mathcal{L}(\widetilde{A} - \widetilde{A}^\star)\big).$$

We first record the rank property of $\mathcal{L}$. Under the standard vectorization identity,

$$\mathcal{L}(U) = (\widetilde{X}^\top \otimes I_{d_y})\,\mathrm{vec}(U).$$

Since $\widetilde{X}$ is full-row-rank, the matrix $\widetilde{X}^\top$ has the full column rank, i.e., $\widetilde{X}^\top \otimes I_{d_y}$ has the full column rank, and

$$\sigma_{\mathbf{min}}(\mathcal{L}) = \sigma_{\mathbf{min}}(\widetilde{X}^\top \otimes I_{d_y}) = \sigma_{\mathbf{min}}(\widetilde{X}).$$

In particular, $\mathcal{L}$ is injective.

The local Lipschitz continuity of h_{stitch} follows from Proposition 5 because $H_{\tau,\mu,q}$ is locally Lipschitz and h_{stitch} is its composition with an affine map.

Next, we show that h_{stitch} is nonsmooth. Since $\mathcal{L}$ is injective, for every nonzero $U \in \mathbb{R}^{d_y \times (d_x+1)}$, we have $\mathcal{L}U \neq 0$. Fix such a U, and set $v := \mathcal{L}U$ and $u := \frac{v}{\|v\|}$. Continuing along the lines $\widetilde{A}^\star + tU$, we have

$$h_{\mathrm{stitch}}(\widetilde{A}^\star + tU) = H_{\tau,\mu,q}(tv).$$

For sufficiently small $t > 0$,

$$H_{\tau,\mu,q}(tv) = \big(t\|v\| + \mu t^2 \|v\|^2\big) q(u),$$

whereas for sufficiently small $t < 0$,

$$H_{\tau,\mu,q}(tv) = \big(|t|\|v\| + \mu t^2 \|v\|^2\big) q(-u).$$

Consequently, the one-dimensional restriction $t \mapsto h_{\mathrm{stitch}}(\widetilde{A}^\star + tU)$ has a kink at $t = 0$. Hence h_{stitch} is nonsmooth at $\widetilde{A}^\star$.

We now show that h_{stitch} is not star-convex with respect to $\widetilde{A}^\star$. Since $\mathcal{L}$ is injective, choose any $U \neq 0$, and set $u := \frac{\mathcal{L}U}{\|\mathcal{L}U\|}$ and $t := \frac{4\tau}{\|\mathcal{L}U\|}$. Then $t\mathcal{L}U = 4\tau u$. By Proposition 5, the function $H_{\tau,\mu,q}$ is not star-convex with respect to 0. More explicitly, using the computation in the proof of Proposition 5,

$$H_{\tau,\mu,q}(\tau u) > \frac{1}{4} H_{\tau,\mu,q}(4\tau u),$$

i.e.,

$$H_{\tau,\mu,q}\left(\frac{t}{4}\mathcal{L}U\right) > \frac{1}{4} H_{\tau,\mu,q}(t\mathcal{L}U).$$

Using the identity $h_{\mathrm{stitch}}(\widetilde{A}^\star + sU) = H_{\tau,\mu,q}(s\mathcal{L}U)$, we come to

$$h_{\mathrm{stitch}}\left(\widetilde{A}^\star + \frac{t}{4}U\right) > \frac{1}{4} h_{\mathrm{stitch}}(\widetilde{A}^\star + tU).$$

In addition, it can be deduced that

$$\widetilde{A}^\star + \frac{t}{4}U = \frac{3}{4}\widetilde{A}^\star + \frac{1}{4}(\widetilde{A}^\star + tU).$$

Since $h_{\mathrm{stitch}}(\widetilde{A}^\star) = 0$, the star-convexity with respect to $\widetilde{A}^\star$ implies

$$h_{\mathrm{stitch}}\left(\frac{3}{4}\widetilde{A}^\star + \frac{1}{4}(\widetilde{A}^\star + tU)\right) \leq \frac{3}{4} h_{\mathrm{stitch}}(\widetilde{A}^\star) + \frac{1}{4} h_{\mathrm{stitch}}(\widetilde{A}^\star + tU) = \frac{1}{4} h_{\mathrm{stitch}}(\widetilde{A}^\star + tU),$$

which contradicts the strict inequality above. Therefore, h_{stitch} is not star-convex with respect to $\widetilde{A}^\star$.

It remains to prove strong quasar-convexity. Since $H_{\tau,\mu,q}$ is $\left(\frac{1}{2}, 6\mu\right)$-strongly quasar-convex with respect to $0 \in \mathbf{arg\,min}\, H_{\tau,\mu,q}$ and since $\mathcal{L}$ has full column rank and $\mathcal{L}(0) = 0 \in \mathbf{arg\,min}\, H_{\tau,\mu,q}$, the linear-composition rule for strongly quasar-convex functions (see (Ahookhosh et al., 2026a, Proposition 21)) gives that the function $U \mapsto H_{\tau,\mu,q}(\mathcal{L}U)$ is $\left(\frac{1}{2}, 6\mu\,\sigma_{\mathbf{min}}(\mathcal{L})^2\right) = \left(\frac{1}{2}, 6\mu\,\sigma_{\mathbf{min}}(\widetilde{X})^2\right)$-strongly quasar-convex with respect to $U = 0$. Finally, the translation $U = \widetilde{A} - \widetilde{A}^\star$ preserves strong quasar-convexity, and therefore h_{stitch} is $\left(\frac{1}{2}, 6\mu\,\sigma_{\mathbf{min}}(\widetilde{X})^2\right)$-strongly quasar-convex with respect to $\widetilde{A}^\star$. □

B.6 Identifiable feature spans

The full-rank assumptions in the clean alignment examples can be expressed directly on the empirical feature span. If $\mathrm{rank}(Z_s) = r < d_s$, take a thin singular value decomposition

$$Z_s = U_r \Sigma_r V_r^\top.$$

Then

$$WZ_s = (WU_r)(\Sigma_r V_r^\top),$$

i.e., the alignment objective depends on W only through the identifiable coefficient WU_r. The reduced feature matrix $\widehat{Z}_s := U_r^\top Z_s = \Sigma_r V_r^\top \in \mathbb{R}^{r\times N}$ has full row rank. As such, the full-rank formulation is equivalent to working on the empirical feature span. The same reduction applies to the augmented matrix $\widetilde{X}$ in the affine stitching model. Directions orthogonal to the empirical span are not identifiable from the observed pairs and create flat directions in the unreduced parameterization.

C Proofs of Section 3 and further properties of inexact high-order oracle

The following remark clarifies the well-definedness of Definition 8 (**a**).

Remark 23. *Since Q_k is proper, lower semicontinuous, bounded below, and coercive, its infimum is attained. Therefore, for every $\varepsilon_k \geq 0$, the model-value gap condition*

$$Q_k(x^{k+1}) \leq \inf_{y\in\mathbb{R}^n} Q_k(y) + \varepsilon_k$$

is well defined. In particular, when $\varepsilon_k = 0$, this condition reduces to the exact high-order proximal step $x^{k+1} \in T_{\beta_k,p}(x^k)$. Thus, the model-value gap oracle is a direct relaxation of the exact proximal oracle used in HiPPA (Ahookhosh et al., 2026a).

By Definition 8 (**a**), for every $z \in \mathbb{R}^n$, it can be concluded that

$$Q_k(x^{k+1}) \le Q_k(z) + \varepsilon_k. \tag{41}$$

In particular, choosing $z = x^k$ yields the descent estimate

$$h(x^{k+1}) + \frac{1}{p\beta_k} \left\|x^{k+1} - x^k\right\|^p \le h(x^k) + \varepsilon_k. \tag{42}$$

Setting $z = \overline{x}$ leads to

$$h(x^{k+1}) - h^\star \le \frac{1}{p\beta_k} \left(\left\|x^k - \overline{x}\right\|^p - \left\|x^{k+1} - x^k\right\|^p \right) + \varepsilon_k. \tag{43}$$

Remark 24 (A model-value gap does not automatically imply a residual bound). *For nonsmooth proximal models, a small model-value gap does not generally imply a small Clarke residual. For example, for $Q(y) = |y|$, the value-based gap $Q(y) - \inf Q$ tends to zero as $y \to 0$, but $\mathbf{dist}(0, \partial^C Q(y)) = 1$ for $y \neq 0$. Thus, converting a model-value gap estimate into a residual estimate requires an additional error-bound or growth condition for the proximal model.*

We now give the bridge from model-value gap inexactness to metric inexactness. This bridge controls the distance to the exact proximal solution set under local model growth; it is distinct from, and does not imply, a same-point residual error bound.

Assumption 25 (Local metric growth of the proximal model). *There exist neighborhoods U, V of $\overline{x}$, constants $c_{\mathrm{mg}} > 0$ and $s \ge 1$, and an interval $I = [\beta_{\min}, \beta_{\max}]$, such that, for all $x \in U$, all $y \in V$, and all $\beta \in I$,*

$$Q_{x,\beta}(y) - \inf_{z \in \mathbb{R}^n} Q_{x,\beta}(z) \ge c_{\mathrm{mg}} \, \mathbf{dist}\big(y, T_{\beta,p}(x)\big)^s. \tag{44}$$

Proposition 26 (Model-value gap implies metric inexactness). *Suppose Assumption 25 holds. If $x^k \in U$, $x^{k+1} \in V$, $\beta_k \in I$, and x^{k+1} satisfies (11), then*

$$\mathbf{dist}\big(x^{k+1}, T_{\beta_k,p}(x^k)\big) \le c_{\mathrm{mg}}^{-1/s} \varepsilon_k^{1/s}. \tag{45}$$

Proof. By Assumption 25, applied with $x = x^k$, $y = x^{k+1}$, and $\beta = \beta_k$, we have

$$c_{\mathrm{mg}} \mathbf{dist}(x^{k+1}, T_{\beta_k,p}(x^k))^s \le Q_k(x^{k+1}) - \inf Q_k \le \varepsilon_k.$$

Taking the s-th root gives our desired result. □

C.1 Auxiliary one-step estimates

Here, we collect three elementary estimates that will be used repeatedly. The first one follows directly from the model-value gap condition and gives the basic descent mechanism of the method.

Lemma 27 (Model-value gap descent inequality). *Let $\overline{x} \in X^\star$, and let $\{x^k\}_{k \in \mathbb{N}_0}$ be generated by Algorithm 1. Then, for every $k \ge 0$,*

$$h(x^{k+1}) - h(\overline{x}) \le \frac{1}{p\beta_k} \left(\left\|x^k - \overline{x}\right\|^p - \left\|x^{k+1} - x^k\right\|^p \right) + \varepsilon_k. \tag{46}$$

Moreover, Moreover, for every $N \ge 1$,

$$\sum_{k=0}^{N-1} \frac{1}{p\beta_k} \left\|x^{k+1} - x^k\right\|^p \le h(x^0) - h(\overline{x}) + \sum_{k=0}^{N-1} \varepsilon_k. \tag{47}$$

Proof. The inequality (46) is exactly (43). Summing (42) from $k = 0$ to $N - 1$ yields (47). □

The next estimate is a perturbed Fejér-type inequality. It is obtained by applying the residual-to-gap inequality to the approximate optimality condition of the proximal subproblem. The signed form keeps the useful geometry of the proximal displacement, while the norm form gives a simpler bound.

Lemma 28 (**Perturbed one-step inequality**). *Assume that h is locally Lipschitz and (κ, γ)-strongly quasar-convex with respect to $\overline{x} \in X^\star$. Let $x^k, x^{k+1} \in \mathbf{dom}(h)$, and suppose that x^{k+1} admits a residual r^{k+1} satisfying (12). Define $d^k := x^{k+1} - x^k$. Then, for every $k \in \mathbb{N}_0$,*

$$\kappa\left(h(x^{k+1}) - h(\overline{x})\right) + \frac{\kappa\gamma}{2}\left\|x^{k+1} - \overline{x}\right\|^2 \leq \frac{1}{\beta_k}\left\|d^k\right\|^{p-2}\left\langle d^k, \overline{x} - x^{k+1}\right\rangle + \left\|r^{k+1}\right\|\left\|x^{k+1} - \overline{x}\right\|. \tag{48}$$

Moreover,

$$\kappa\left(h(x^{k+1}) - h(\overline{x})\right) + \frac{\kappa\gamma}{2}\left\|x^{k+1} - \overline{x}\right\|^2 \leq \frac{1}{\beta_k}\left\|d^k\right\|^{p-1}\left\|x^{k+1} - \overline{x}\right\| + \left\|r^{k+1}\right\|\left\|x^{k+1} - \overline{x}\right\|. \tag{49}$$

If $p = 2$, then the sharper identity

$$2\beta_k\kappa\left(h(x^{k+1}) - h(\overline{x})\right) + \beta_k\kappa\gamma\left\|x^{k+1} - \overline{x}\right\|^2 \leq \left\|x^k - \overline{x}\right\|^2 - \left\|x^{k+1} - \overline{x}\right\|^2 - \left\|d^k\right\|^2 + 2\beta_k\left\|r^{k+1}\right\|\left\|x^{k+1} - \overline{x}\right\|. \tag{50}$$

holds.

Proof. Applying Proposition 20 with $x = x^k$ and $y = x^{k+1}$ yields

$$\kappa\left(h(x^{k+1}) - h(\overline{x})\right) + \frac{\kappa\gamma}{2}\left\|x^{k+1} - \overline{x}\right\|^2 \leq \frac{1}{\beta_k}\left\|d^k\right\|^{p-2}\left\langle d^k, \overline{x} - x^{k+1}\right\rangle + \left\|r^{k+1}\right\|\left\|x^{k+1} - \overline{x}\right\|,$$

which gives (48). Estimate the transport term by Cauchy–Schwarz to obtain (49). If $p = 2$, then

$$2\left\langle d^k, \overline{x} - x^{k+1}\right\rangle = \left\|x^k - \overline{x}\right\|^2 - \left\|x^{k+1} - \overline{x}\right\|^2 - \left\|d^k\right\|^2,$$

and multiplying (48) by $2\beta_k$ yields (50). □

Proposition 29 (**Residual-and-descent accepted sequences**). *Let h be lower bounded and locally Lipschitz on a neighborhood of a bounded sequence $\{x^k\}_{k\in\mathbb{N}_0}$. Assume (9), let $\eta_k \to 0$, and suppose every step satisfies (21). Then*

$$\sum_{k=0}^{\infty}\left\|x^{k+1} - x^k\right\|^p < \infty,$$

and every cluster point is Clarke critical. If, in addition, h is κ-quasar-convex with respect to $\overline{x}$, then every cluster point is globally optimal. If the global minimizer is unique, the full sequence converges to it.

Proof. The descent part of (21) gives

$$h(x^{k+1}) + \frac{1}{p\beta_k}\left\|x^{k+1} - x^k\right\|^p \leq h(x^k).$$

Summing and using $\beta_k \leq \beta_{\mathbf{max}}$ proves step summability, hence $x^{k+1} - x^k \to 0$. The residual part gives $v^{k+1} \in \partial^C h(x^{k+1})$ such that

$$\left\|v^{k+1} + \beta_k^{-1} J_p(x^{k+1} - x^k)\right\| \leq \eta_k.$$

Because $\beta_k \geq \beta_{\mathbf{min}}$, it follows that $v^{k+1} \to 0$. Along any cluster subsequence, the outer semicontinuity of the Clarke subdifferential implies $0 \in \partial^C h(\widehat{x})$. Fact 17 gives global optimality when quasar geometry is maintained. Uniqueness and boundedness then yield full-sequence convergence. □

C.2 Restricted capped-loss residual certificate

Let $r_i = W z_{s,i} - z_{t,i}$. Away from cap kinks, define

$$\mathcal{A}_i(r_i) = \begin{cases} \{r_i/\|r_i\| + 2\mu r_i\}, & 0 < \|r_i\| < \tau, \\ \{2\mu r_i\}, & \|r_i\| > \tau, \\ \{u : \|u\| \le 1\}, & r_i = 0. \end{cases}$$

All summands are then Clarke regular. If $D = W - W^k$, the exact proximal residual appearing in (21) is

$$\min_{\xi_i \in \mathcal{A}_i(r_i)} \left\| \frac{1}{N} \sum_{i=1}^{N} \xi_i z_{s,i}^\top + \beta_k^{-1} \|D\|_F^{p-2} D \right\|_F. \tag{51}$$

The only free variables occur at exact zero residuals, where this is a convex minimum-norm problem over Euclidean unit balls; any feasible selection gives an upper bound and therefore a sufficient acceptance certificate. At a cap kink, the finite-sum Clarke sum rule need not be exact, so (51) is not used.

C.3 Proof of Theorem 9

Proof. (**a**) It follows from the descent estimate (42).
(**b**), Lemma 27 implies, for every $N \ge 1$,

$$\sum_{k=0}^{N-1} \frac{1}{p\beta_k} \left\|x^{k+1} - x^k\right\|^p \le h(x^0) - h(\overline{x}) + \sum_{k=0}^{N-1} \varepsilon_k.$$

Since $\overline{x} \in X^\star$, $h(\overline{x}) = h^\star$, and since $\sum_{k=0}^{\infty} \varepsilon_k < \infty$, the right-hand side is uniformly bounded in N. In light of $\beta_k \le \beta_{\mathbf{max}}$, it can be concluded that

$$\frac{1}{p\beta_{\mathbf{max}}} \sum_{k=0}^{N-1} \left\|x^{k+1} - x^k\right\|^p \le h(x^0) - h^\star + \sum_{k=0}^{\infty} \varepsilon_k.$$

Letting $N \to \infty$ yields

$$\sum_{k=0}^{\infty} \left\|x^{k+1} - x^k\right\|^p < \infty,$$

and consequently we come to $\left\|x^{k+1} - x^k\right\| \to 0$, giving our desired result.
(**c**) Assume that $\{x^k\}_{k \in \mathbb{N}_0}$ is bounded and let $x^{k_j} \to \widehat{x}$. Then $x^{k_j+1} \to \widehat{x}$. Fix $\lambda \in (0, 1]$ and set

$$z_j := (1-\lambda)x^{k_j+1} + \lambda \overline{x}.$$

Comparing the inexact proximal model at x^{k_j+1} with z_j and using quasar-convexity gives

$$\kappa\lambda\big(h(x^{k_j+1}) - h^\star\big) \le \frac{\left\|z_j - x^{k_j}\right\|^p - \left\|d^{k_j}\right\|^p}{p\beta_{k_j}} + \varepsilon_{k_j}.$$

Since

$$z_j - x^{k_j} = (1-\lambda)d^{k_j} + \lambda(\overline{x} - x^{k_j}),$$

letting $j \to \infty$ yields

$$\kappa\lambda(\ell - h^\star) \le \frac{\lambda^p}{p\beta_{\mathbf{min}}} \|\overline{x} - \widehat{x}\|^p.$$

After division by $\kappa\lambda$, let $\lambda \downarrow 0$. Since $p > 1$, we obtain $\ell = h^\star$. The lower semicontinuity then gives $h(\widehat{x}) = h^\star$, i.e., every cluster point is globally optimal. If $X^\star = \{\overline{x}\}$,the boundedness implies $x^k \to \overline{x}$.
(**d**) If h is (κ, γ)-strongly quasar-convex with $\gamma > 0$, the quadratic growth bound (25) gives a uniform bound on $\left\|x^k - \overline{x}\right\|$. The same growth bound makes the minimizer unique. The preceding bounded-sequence argument therefore gives $x^k \to \overline{x}$. □

C.4 Sufficient conditions for Assumption 10

The following result shows how the perturbed recursions follow from exact proximal-map estimates and metric inexactness. This separates the behavior of the exact high-order proximal step from the error introduced by the inexact subproblem solve.

Proposition 30 (Sharp clean growth and finite proximal identification). *For the clean capped feature-alignment model of Example 3, there exist $c > 0$ and a neighborhood of $W^\star$ such that*

$$h_{\mathrm{KD}}(W) - h_{\mathrm{KD}}(W^\star) \geq c \left\| W - W^\star \right\|_F, \qquad c = \frac{\sigma_{\min}(Z_s)}{N}.$$

For the clean anisotropic stitching model of Example 6, the analogous estimate holds with $c = \sigma_{\min}(\widetilde{X})$. In either case, after possibly shrinking the neighborhood, uniformly for $\beta \in [\beta_{\min}, \beta_{\max}]$ and every $p > 1$,

$$T_{\beta,p}(X) = \{x^\star\}$$

whenever the proximal center X is sufficiently close to the corresponding minimizer. Consequently, Assumption 25 holds locally with exponent $s = 1$.

Proof. For the capped model, choose the neighborhood so that every clean residual lies below the cap. With $D = W - W^\star$,

$$h_{\mathrm{KD}}(W) - h_{\mathrm{KD}}(W^\star) \geq \frac{1}{N} \sum_{i=1}^{N} \|Dz_{s,i}\| \geq \frac{1}{N} \left\| DZ_s \right\|_F \geq \frac{\sigma_{\min}(Z_s)}{N} \left\| D \right\|_F.$$

For the anisotropic model, $q \geq 1$ and $f_{\tau,\mu}(r) \geq r$ near the origin, while the affine design map obeys the corresponding smallest-singular-value bound.

Write either a sharp estimate as $h(y) - h^\star \geq c \left\| y - x^\star \right\|$. The convexity of $u \mapsto \|u - X\|^p / p$ at $x^\star$ yields

$$Q_{X,\beta}(y) - Q_{X,\beta}(x^\star) \geq \left(c - \beta^{-1} \left\| X - x^\star \right\|^{p-1} \right) \left\| y - x^\star \right\|.$$

Choose X so close to $x^\star$ that the coefficient is at least $c/2$ uniformly for $\beta \geq \beta_{\min}$. This gives strict local minimality and linear model growth. The quadratic growth already available for the clean strongly quasar-convex objectives excludes global proximal minimizers outside a sufficiently small neighborhood when X is close enough. Therefore, the exact proximal set is the singleton $\{x^\star\}$, and the same lower bound gives local metric growth with $s = 1$. □

Remark 31. *For these sharp clean examples, the exact proximal map is locally constant for every $p > 1$. Hence, they do not display nontrivial exact p-dependent local asymptotics; after identification, the inexact behavior is governed by the forcing error.*

Proposition 32 (A verification route for the perturbed recursions). *Let*

$$y^{k+1} \in \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left\{ h(y) + \frac{1}{p\beta_k} \left\| y - x^k \right\|^p \right\}$$

be an exact high-order proximal point, and suppose the computed point x^{k+1} satisfy a metric inexactness estimate

$$\left\| x^{k+1} - y^{k+1} \right\| \leq \Delta_k. \tag{52}$$

Then the following implications hold.

(a) *Suppose $p \in (1, 2]$ and the exact proximal step satisfies*

$$\left\| y^{k+1} - \overline{x} \right\|^2 \leq \eta \left\| x^k - \overline{x} \right\|^2,$$

*for some $\eta \in (0, 1)$. If $\Delta_k^2 \leq \overline{C} \varepsilon_k$, for some $\overline{C} > 0$, then Assumption 10 (**a**) holds for any $\eta' \in (\eta, 1)$, after possibly increasing the constant C.*

(b) *Suppose $p > 2$ and the exact proximal step satisfies*

$$\left\| y^{k+1} - \overline{x} \right\| \leq a \left\| x^k - \overline{x} \right\|^{p-1},$$

*for some $a > 0$. If $\Delta_k \leq \overline{C} \sqrt{\varepsilon_k}$, for some $\overline{C} > 0$, then Assumption 10 (**b**) holds with constants depending only on a and $\overline{C}$.*

Proof. (**a**) By (52) and Young's inequality, for every $\theta > 0$, we get

$$r_{k+1}^2 = \left\|x^{k+1} - \overline{x}\right\|^2 \leq (1+\theta)\left\|y^{k+1} - \overline{x}\right\|^2 + (1+\theta^{-1})\Delta_k^2.$$

Let us choose $\theta > 0$ small enough so that $(1+\theta)\eta < 1$, set $\eta' = (1+\theta)\eta$, and use $\Delta_k^2 \leq \overline{C}\varepsilon_k$. This consequently gives Assumption 10 (**a**).
(**b**) Invoking the triangle inequality leads to

$$r_{k+1} \leq \left\|y^{k+1} - \overline{x}\right\| + \Delta_k \leq a r_k^{p-1} + \overline{C}\sqrt{\varepsilon_k},$$

which is Assumption 10 (**b**). □

C.5 Proof of Theorem 11

Proof. We prove the two cases separately. Let us set $r_k := \|x^k - \overline{x}\|$.

Suppose $p \in (1, 2]$. By Assumption 10(**a**), for all sufficiently large k,

$$r_{k+1}^2 \leq \eta r_k^2 + C\varepsilon_k.$$

Let k_0 be large enough so that this recursion holds for all $k \geq k_0$, and so that

$$\varepsilon_k \leq E\rho^{2k}, \quad k \geq k_0.$$

Iterating the recursion for $k > k_0$ leads to

$$\begin{aligned} r_k^2 &\leq \eta^{k-k_0} r_{k_0}^2 + C\sum_{j=k_0}^{k-1} \eta^{k-1-j}\varepsilon_j \\ &\leq \eta^{k-k_0} r_{k_0}^2 + CE\sum_{j=k_0}^{k-1} \eta^{k-1-j}\rho^{2j}. \end{aligned}$$

Due to the inequality $\eta < \rho^2$, the first term satisfies

$$\eta^{k-k_0} r_{k_0}^2 \leq \rho^{2(k-k_0)} r_{k_0}^2 = \rho^{-2k_0} r_{k_0}^2 \rho^{2k}.$$

For the convolution term, we write

$$\sum_{j=k_0}^{k-1} \eta^{k-1-j}\rho^{2j} = \rho^{2k-2}\sum_{j=k_0}^{k-1}\left(\frac{\eta}{\rho^2}\right)^{k-1-j} \leq \rho^{2k-2}\frac{1}{1-\eta/\rho^2},$$

i.e.,

$$r_k^2 \leq \left(\rho^{-2k_0} r_{k_0}^2 + \frac{CE\rho^{-2}}{1-\eta/\rho^2}\right)\rho^{2k}.$$

Consequently, there exists $A > 0$ such that $r_k \leq A\rho^k$ for all sufficiently large k.

If h is locally Lipschitz near $\overline{x}$, then there exists $L > 0$ such that, for all sufficiently large k,

$$h(x^k) - h^\star \leq L\left\|x^k - \overline{x}\right\|,$$

i.e.,

$$h(x^k) - h^\star \leq C_h\rho^k$$

for a suitable constant $C_h > 0$.

Suppose $p > 2$. By In light of Assumption 10(**b**), it is clear that

$$r_{k+1} \leq a r_k^{p-1} + b\sqrt{\varepsilon_k},$$

for all sufficiently large k. Applying the forcing condition $\sqrt{\varepsilon_k} \leq M r_k^{p-1+\delta}$ results in

$$r_{k+1} \leq a r_k^{p-1} + bM r_k^{p-1+\delta} = \left(a + bM r_k^\delta\right) r_k^{p-1}.$$

Since $x^k \to \overline{x}$, we have $r_k \to 0$. Hence, the factor $a + bM r_k^\delta$ is bounded for all sufficiently large k, i.e., there exists $B > 0$ such that

$$r_{k+1} \leq B r_k^{p-1},$$

for all sufficiently large k. Dividing both sides by r_k results in

$$\frac{r_{k+1}}{r_k} \leq B r_k^{p-2} \to 0,$$

due to $p > 2$ and $r_k \to 0$, which consequently leads to a superlinear convergence.

Finally, if h is locally Lipschitz near $\overline{x}$, then for all sufficiently large k,

$$h(x^k) - h^\star \leq C_h^p \left\|x^k - \overline{x}\right\|,$$

after increasing C_h^p if necessary. □

Remark 33 (**Sharper value rates under local upper quadratic growth**). *If, in addition to the assumptions of Theorem 11, there exists $L_{\mathrm{uq}} > 0$ such that*

$$h(x) - h^\star \leq L_{\mathrm{uq}} \|x - \overline{x}\|^2$$

locally around $\overline{x}$, then the value estimates in Theorem 11 can be sharpened. In the case $p \in (1, 2]$, one gets

$$h(x^k) - h^\star \leq \widetilde{C}_h \rho^{2k},$$

for all sufficiently large k. In the case $p > 2$, one gets

$$h(x^k) - h^\star \leq L_{\mathrm{uq}} \left\|x^k - \overline{x}\right\|^2,$$

for all sufficiently large k.

C.6 Proof of Theorem 13

Proof. Set $r_k := \|x^k - \overline{x}\|$ and $g_k := h(x^k) - h(\overline{x})$. We consider the two regimes separately. First, let $1 < p \leq 2$. By Theorem 11, under the forcing condition $\varepsilon_k = \mathcal{O}(\rho^{2k})$, with $\rho \in (0, 1)$, there exist constants $C_1 > 0$ and $k_0 \in \mathbb{N}_0$ such that $r_k \leq C_1 \rho^k$, for all $k \geq k_0$. Thus, $r_k \leq \varepsilon$ is guaranteed whenever $C_1 \rho^k \leq \varepsilon$. Since $\rho \in (0, 1)$, this is equivalent to

$$k \geq \frac{\log(C_1/\varepsilon)}{\log(1/\rho)},$$

leading to $N_r(\varepsilon) = \mathcal{O}\big(\log(\varepsilon^{-1})\big)$.

Now, let $p > 2$. By Theorem 11, under the forcing condition $\varepsilon_k = \mathcal{O}\left(r_k^{2p-2+2\delta}\right)$, there exist constants $C > 0$ and $k_0 \in \mathbb{N}$ such that

$$r_{k+1} \leq C r_k^{p-1} \qquad \forall k \geq k_0.$$

Set $s := p - 1 > 1$. Since $r_k \to 0$, increasing k_0 if necessary, we may assume that

$$R_0 := C^{1/(s-1)} r_{k_0} < 1.$$

For $j \geq 0$, let us define

$$R_j := C^{1/(s-1)} r_{k_0+j},$$

i.e.,

$$R_{j+1} = C^{1/(s-1)} r_{k_0+j+1} \leq C^{1/(s-1)} C r_{k_0+j}^s = \left(C^{1/(s-1)} r_{k_0+j}\right)^s = R_j^s.$$

By induction, it holds that $R_j \leq R_0^{s^j}$, which ensures

$$r_{k_0+j} \leq C^{-1/(s-1)} R_0^{s^j}.$$

To guarantee $r_{k_0+j} \leq \varepsilon$, it is enough that

$$C^{-1/(s-1)} R_0^{s^j} \leq \varepsilon,$$

equivalently,

$$R_0^{s^j} \leq C^{1/(s-1)} \varepsilon.$$

Taking logarithms and using $R_0 \in (0,1)$, it suffices that

$$s^j \geq \frac{\log\left(C^{-1/(s-1)}\varepsilon^{-1}\right)}{\log(R_0^{-1})},$$

i.e.,

$$j \geq \frac{\log\log\left(C^{-1/(s-1)}\varepsilon^{-1}\right) - \log\log(R_0^{-1})}{\log s}.$$

Consequently, $N_r(\varepsilon) = \mathcal{O}\big(\log\log(\varepsilon^{-1})\big)$. This proves the distance-complexity estimates in (17).

Now, we move from distance bounds to function-value bounds. Assume that h is locally Lipschitz near $\overline{x}$. Then there exist a neighborhood U of $\overline{x}$ and a constant $L_{\mathrm{loc}} > 0$ such that

$$|h(x) - h(y)| \leq L_{\mathrm{loc}}\|x - y\|, \qquad x, y \in U.$$

Since $x^k \to \overline{x}$, for all sufficiently large k, both x^k and $\overline{x}$ belong to U. Therefore, taking $y = \overline{x}$, we get

$$g_k = h(x^k) - h(\overline{x}) \leq L_{\mathrm{loc}} r_k$$

for all sufficiently large k. Hence, $g_k \leq \varepsilon$ is guaranteed by requiring

$$r_k \leq \frac{\varepsilon}{L_{\mathrm{loc}}},$$

i.e., $N_g(\varepsilon)$ has the same order as $N_r(\varepsilon)$, with only a change in constants. This proves the stated function-value complexity under local Lipschitz continuity. This completes the proof. □

D Proofs for the corruption and recovery results

Proof of Proposition 14. Let h_E denote the capped objective under (18). Global optimality gives

$$h_E(\widehat{W}) \leq h_E(W^\star) = A_E.$$

Dropping the nonnegative capped term yields

$$\frac{\mu}{N}\left\|(\widehat{W} - W^\star)Z_s - E\right\|_F^2 \leq A_E.$$

The triangle inequality and

$$\frac{1}{\sqrt{N}}\|(W - W^\star)Z_s\|_F \geq \sqrt{\lambda_-}\,\|W - W^\star\|_F$$

prove (19). Applying the same triangle inequality before squaring gives (20). If a fraction ϱ of columns is replaced and their norms are uniformly bounded, then both δ_2 and $\sqrt{A_E/\mu}$ are $\mathcal{O}(\sqrt{\varrho})$. □

Proof of Theorem 15. Put $W = W^\star + U$ and shrink to a convex neighborhood in which no corrupted residual crosses its cap. On this neighborhood, the fixed-pattern branch is convex in U and its subdifferential at 0 is $\partial L_0(0) - B_E$. Hence $U = 0$ minimizes this local branch exactly whenever $B_E \in \partial L_0(0)$. This is equivalent to nonnegativity of every directional derivative $L_0'(0;U) - \langle B_E, U\rangle_F$, which gives the stated condition on α_E. □

Corollary 34 (Conic sufficient certificate). *Let $Z_0 \in \mathbb{R}^{|\mathcal{I}|\times d_s}$ stack the rows $z_{s,i}^\top$ for $i \in \mathcal{I}$ and assume Z_0 has full column rank. Define*

$$G := NB_E(Z_0^\top Z_0)^{-1}Z_0^\top = [g_i].$$

If $\|g_i\| \leq 1$ for every column, then $W^\star$ is a local minimizer in the fixed-pattern neighborhood. If $\mathbf{max}_i\,\|g_i\| \leq 1 - \delta$, then

$$\alpha_E \geq \frac{\delta\,\sigma_{\mathbf{min}}(Z_0)}{N}.$$

Proof. The subdifferential has the representation

$$\partial L_0(0) = \left\{ \frac{1}{N} G_0 Z_0 : \|(G_0)_i\| \le 1 \right\}.$$

The displayed G satisfies $GZ_0 = NB_E$, so the unit-ball condition gives $B_E \in \partial L_0(0)$. Under strict margins, we have

$$L_0'(0;U) - \langle B_E, U \rangle_F \ge \frac{\delta}{N} \sum_{i \in \mathcal{I}} \|U z_{s,i}\| \ge \frac{\delta \sigma_{\min}(Z_0)}{N} \|U\|_F .$$

Taking $\|U\|_F = 1$ proves the claim. □

Proposition 35 (Local corrupted proximal identification). *Suppose $\alpha_E > 0$ and remain inside the fixed-pattern neighborhood of Theorem 15. After shrinking that neighborhood if necessary,*

$$h_E(W) - h_E(W^\star) \ge \frac{\alpha_E}{2} \|W - W^\star\|_F .$$

For a proximal center X, if

$$\frac{\alpha_E}{2} - \beta^{-1} \|X - W^\star\|_F^{p-1} > 0, \tag{53}$$

then $W^\star$ is a strict local minimizer of $W \mapsto h_E(W) + (p\beta)^{-1} \|W - X\|_F^p$ on that same branch.

Proof. The strict directional margin and compactness of the unit sphere give the local sharp estimate displayed. With $D = W - W^\star$, convexity of $V \mapsto \|V - X\|_F^p / p$ at $W^\star$ yields

$$\begin{aligned} Q_{X,\beta}(W) - Q_{X,\beta}(W^\star) &\ge \frac{\alpha_E}{2} \|D\|_F + \beta^{-1} \left\langle \|W^\star - X\|_F^{p-2} (W^\star - X), D \right\rangle \\ &\ge \left(\frac{\alpha_E}{2} - \beta^{-1} \|X - W^\star\|_F^{p-1} \right) \|D\|_F . \end{aligned}$$

The coefficient is positive by (53). This is a local identification statement only; it does not imply global uniqueness for the nonconvex proximal subproblem. □

E Additional numerical details

E.1 Order-sensitivity diagnostics

The main text reports the representative orders $p \in \{2, 3, 4, 5, 6\}$. Table 3 gives the finer fixed-protocol sweep. For every $p > 2$ we use the same family

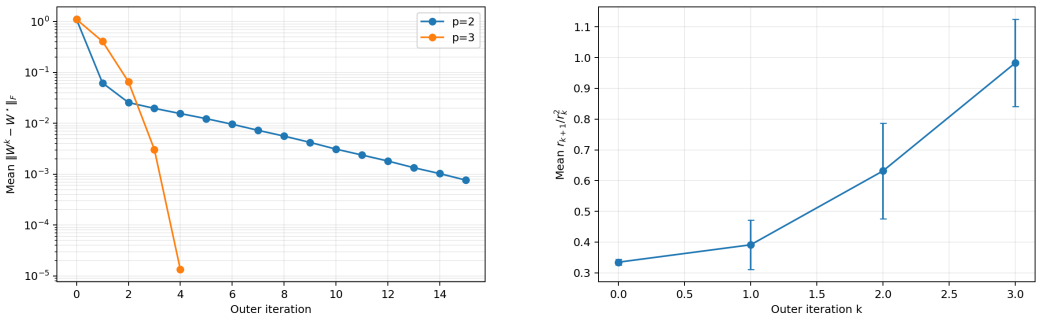
$$\sqrt{\varepsilon_k} = 0.05\, r_k^{p-1+0.25},$$

with the same data, five perturbation directions, local basin, lower solver, endpoint, and $\beta = 1/m$ as in Section 5. The endpoint count saturates at three outer steps from $p = 4$ onward. At the same time, lower-level work grows markedly for larger orders. For $p = 7$ and $p = 8$, the observed trajectories still reach the endpoint in three outer steps, but the prescribed model-value tolerance falls below what the double-precision lower solve certifies on all seeds. These rows are therefore reported only as finite-precision stress tests.

For $p = 3$, the mean values of r_{k+1}/r_k^2 over the four nonterminal transitions are 0.334, 0.391, 0.631, and 0.983, while the late-stage $p = 2$ ratios r_{k+1}/r_k are approximately 0.74–0.77. An instrumented replay confirms that every lower-level objective/gradient evaluation for $p \le 6$ remains inside the verified quadratic basin, with the largest observed radius equal to 0.643 of the conservative basin radius.

This behavior is not an artifact of using one numerical value of β for every order. Repeating the sensitivity check with the scale-covariant choice

$$\beta_p = \frac{r_0^{p-2}}{m}$$



(a) Mean distance to the exact affine head.

(b) $p = 3$ quadratic-rate diagnostic.

Figure 1: Local theorem-validation diagnostics inside the verified quadratic basin.

Table 3: Finer order-sensitivity sweep for the capped-quadratic local theorem-validation problem. "Certified" counts the seeds for which every accepted lower solve satisfies its prescribed model-value tolerance.

p	outer steps	lower-level evaluations	certified seeds
2	14.8	571.2	5/5
2.5	4.8	397.0	5/5
3	4.0	529.8	5/5
3.5	3.2	391.2	5/5
4	3.0	416.2	5/5
4.5	3.0	465.4	5/5
5	3.0	523.0	5/5
6	3.0	776.2	5/5
7	3.0	1013.0	4/5
8	3.0	1194.6	3/5

produces the same qualitative conclusion: the outer-step reduction saturates around $p = 4$, whereas lower-level effort grows for larger p. The role of p is therefore best viewed as a geometry- and implementation-dependent tradeoff, rather than a parameter to be made as large as possible.

The limit $p \to \infty$ also changes the character of the proximal model. For fixed β and displacement d,

$$\frac{\|d\|^p}{p\beta} \longrightarrow \begin{cases} 0, & \|d\| \le 1, \\ +\infty, & \|d\| > 1, \end{cases}$$

pointwise. With $\beta_p = s^{p-2}/m$, the corresponding limiting radius is s. Hence the large-p limit is a hard trust-region-type constraint, not an "infinitely accelerated" version of a finite-p proximal step. Moreover, preserving the order-$p-1$ inexact rate requires tolerances of the form $\sqrt{\varepsilon_k} = O(r_k^{p-1+\delta})$, which become increasingly demanding once $r_k < 1$. This explains both the observed diminishing returns and the loss of strict finite-precision certification in the $p = 7, 8$ stress tests.

E.2 Corrupted-feature protocol and certificate audit

For each corruption level and seed, endpoint attainment is evaluated over

$$\varepsilon_f \in \{5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}, \qquad \varepsilon_s \in \{5 \times 10^{-3}, 10^{-2}, 2.5 \times 10^{-2}, 5 \times 10^{-2}\},$$

using

$$\frac{[h(W) - h_{\mathrm{ref}}]_+}{\mathbf{max}\{1, |h_{\mathrm{ref}}|\}} \le \varepsilon_f, \qquad \mathbf{dist}(0, \partial^C h(W)) \le \varepsilon_s.$$

Here h_{ref} is the minimum finite objective recorded across the $p = 2$ and $p = 3$ trajectories for the same instance and is used only as an attained reference value.

Across all 1,200 accepted HiPPA steps in the reported runs, every accepted point lies on a smooth branch of the capped objective, satisfies the implemented proximal-residual certificate, and obeys

$$Q_k(x^{k+1}) \le h(x^k).$$

The complete per-instance traces and lower-level implementation are provided in the supplementary code.

F Further examples: Strong quasar-convexity beyond local Lipschitzness

Here, we present a simple mechanism for constructing functions that are nonsmooth, not necessarily locally Lipschitz, nonconvex, and not star-convex, yet strongly quasar-convex on bounded convex sets. The construction is based on a positively homogeneous residual model raised to a fractional power $\theta \in (0,1)$. We first prove a master lemma and then apply it to robust LAD regression, robust multi-task regression, and robust matrix sensing.

Theorem 36 (Master lemma). *Let $K \subset \mathbf{B}(\overline{x}, R)$ be a nonempty convex set with $\overline{x} \in K$, where $R > 0$. Let $\theta \in (0,1)$, and let $\psi : K \to [0,\infty)$ satisfy*

$$\psi(\lambda\overline{x} + (1-\lambda)x) = (1-\lambda)\psi(x), \qquad \forall x \in K,\ \forall \lambda \in [0,1], \tag{54}$$

and

$$\psi(x) \ge m_\psi \|x - \overline{x}\|, \qquad \forall x \in K, \tag{55}$$

for some $m_\psi > 0$. Define

$$h(x) := \psi(x)^\theta.$$

Then, for every $\kappa \in (0,\theta)$, the function h is (κ,γ)-strongly quasar-convex on K with respect to $\overline{x}$, where

$$\gamma = \frac{2(\theta-\kappa)}{\kappa}\, m_\psi^\theta R^{\theta-2}. \tag{56}$$

Moreover, if ψ is not identically zero on K, then h is not convex and not star-convex with respect to $\overline{x}$. In this case, h is also not locally Lipschitz at $\overline{x}$ relative to K.

Proof. Fix $x \in K$ and $\lambda \in [0,1]$. By (54),

$$h(\lambda\overline{x} + (1-\lambda)x) = \big((1-\lambda)\psi(x)\big)^\theta = (1-\lambda)^\theta h(x).$$

Since $\theta \in (0,1)$, the map $t \mapsto t^\theta$ is concave on $[0,1]$, and hence $(1-\lambda)^\theta \le 1 - \theta\lambda$, i.e.,

$$h(\lambda\overline{x} + (1-\lambda)x) \le (1-\theta\lambda)h(x) = (1-\kappa\lambda)h(x) - (\theta-\kappa)\lambda h(x). \tag{57}$$

From (55), we obtain

$$h(x) = \psi(x)^\theta \ge m_\psi^\theta \|x - \overline{x}\|^\theta.$$

It follows from $x \in K \subset \mathbf{B}(\overline{x}, R)$ and $\theta - 2 < 0$ that

$$\|x - \overline{x}\|^\theta = \|x - \overline{x}\|^2 \|x - \overline{x}\|^{\theta-2} \ge R^{\theta-2}\|x - \overline{x}\|^2,$$

i.e.,

$$h(x) \ge m_\psi^\theta R^{\theta-2}\|x - \overline{x}\|^2.$$

Substituting this estimate into (57), it can be concluded that

$$h(\lambda\overline{x} + (1-\lambda)x) \le (1-\kappa\lambda)h(x) - (\theta-\kappa)\lambda m_\psi^\theta R^{\theta-2}\|x - \overline{x}\|^2.$$

Due to $1 - \frac{\lambda}{2-\kappa} \le 1$, the choice (56) implies

$$(\theta-\kappa)\lambda m_\psi^\theta R^{\theta-2}\|x - \overline{x}\|^2 \ge \lambda\left(1 - \frac{\lambda}{2-\kappa}\right)\frac{\kappa\gamma}{2}\|x - \overline{x}\|^2,$$

i.e.,

$$h(\lambda\overline{x} + (1-\lambda)x) \le (1-\kappa\lambda)h(x) - \lambda\left(1 - \frac{\lambda}{2-\kappa}\right)\frac{\kappa\gamma}{2}\|x - \overline{x}\|^2.$$

Since $h(\overline{x}) = \psi(\overline{x})^\theta = 0$ by (54) with $\lambda = 1$, this is exactly the definition of (κ, γ)-strong quasar-convexity.

Now, suppose that ψ is not identically zero on K. Then there exists d such that $\overline{x} + d \in K$ and $\psi(\overline{x} + d) > 0$. It follows from the convexity of K, $\overline{x} + td \in K$ for every $t \in [0, 1]$, and by (54),

$$h(\overline{x} + td) = t^\theta h(\overline{x} + d).$$

It can be deduced from $\theta \in (0, 1)$ that $t^\theta > t$ for every $t \in (0, 1)$, i.e.,

$$h(\overline{x} + td) > t\, h(\overline{x} + d) = (1 - t)h(\overline{x}) + t\, h(\overline{x} + d),$$

due to $h(\overline{x}) = 0$. Thus, h is not convex. The same strict inequality also shows that the star-convex inequality fails, i.e., h is not star-convex with respect to $\overline{x}$.

Finally, it holds that

$$\frac{|h(\overline{x} + td) - h(\overline{x})|}{\|td\|} = \frac{t^\theta h(\overline{x} + d)}{t\|d\|} = \frac{h(\overline{x} + d)}{\|d\|} t^{\theta - 1} \to +\infty$$

as $t \downarrow 0$. Hence h is not locally Lipschitz at $\overline{x}$ relative to K. □

Remark 37. *All examples below are local in the radius R. Theorem 36 provides strong quasar-convexity on a bounded convex set $K \subset \mathbf{B}(\overline{x}, R)$, not necessarily on the whole ambient space.*

F.1 Nonconvex robust LAD regression

Proposition 38. *Let $A \in \mathbb{R}^{m \times n}$ have the full column rank, let $b = A\overline{x}$, and let $\theta \in (0, 1)$. Define*

$$h_1(x) := \|Ax - b\|_1^\theta,$$

on a nonempty convex set $K \subset \mathbf{B}(\overline{x}, R)$ with $\overline{x} \in K$. Then, for every $\kappa \in (0, \theta)$, the function h_1 is (κ, γ_1)-strongly quasar-convex on K with respect to $\overline{x}$, where

$$\gamma_1 = \frac{2(\theta - \kappa)}{\kappa}\, \sigma_{\mathbf{min}}(A)^\theta\, R^{\theta - 2}.$$

Moreover, h_1 is not locally Lipschitz at $\overline{x}$, is nonconvex, and is not star-convex with respect to $\overline{x}$.

Proof. Let us define

$$\psi_1(x) := \|A(x - \overline{x})\|_1,$$

i.e.,

$$\psi_1(\lambda\overline{x} + (1 - \lambda)x) = (1 - \lambda)\psi_1(x), \qquad \forall x \in K,\ \forall \lambda \in [0, 1].$$

In addition,

$$\psi_1(x) \geq \|A(x - \overline{x})\|_2 \geq \sigma_{\mathbf{min}}(A)\|x - \overline{x}\|_2.$$

Consequently, Theorem 36 applies to $m_\psi = \sigma_{\mathbf{min}}(A)$.

The nonsmoothness at $\overline{x}$ is immediate from $h_1(\overline{x} + td) = t^\theta h_1(\overline{x} + d)$, for every direction d such that $Ad \neq 0$. □

F.2 Nonconvex robust multi-task regression

Let $x_i \in \mathbb{R}^d$ and $W^* \in \mathbb{R}^{p \times d}$, and let us set $y_i = W^* x_i \in \mathbb{R}^p$. Here, we equip $\mathbb{R}^{p \times d}$ with the Frobenius norm $\|\cdot\|_F$.

Proposition 39. *Let $X = [x_1, \ldots, x_N] \in \mathbb{R}^{d \times N}$ have full row rank, let $\theta \in (0, 1)$, and define*

$$h_2(W) := \left(\frac{1}{N} \sum_{i=1}^N \|Wx_i - y_i\|_2 \right)^\theta,$$

on a nonempty convex set

$$K \subset \{W \in \mathbb{R}^{p \times d} : \|W - W^*\|_F \leq R\}, \qquad W^* \in K.$$

Then, for every $\kappa \in (0, \theta)$, the function h_2 is (κ, γ_2)-strongly quasar-convex on K with respect to W^*, where

$$\gamma_2 = \frac{2(\theta - \kappa)}{\kappa} \left(\frac{\sigma_{\mathbf{min}}(X)}{N} \right)^\theta R^{\theta - 2}.$$

Moreover, h_2 is not locally Lipschitz at $W^\star$, is nonconvex, and is not star-convex with respect to $W^\star$.

Proof. Let us define

$$\psi_2(W) := \frac{1}{N}\sum_{i=1}^{N} \|(W - W^*)x_i\|_2,$$

i.e.,

$$\psi_2(\lambda W^* + (1-\lambda)W) = (1-\lambda)\psi_2(W), \qquad \forall W \in K,\ \forall \lambda \in [0,1].$$

In addition,

$$\psi_2(W) \geq \frac{1}{N}\left(\sum_{i=1}^{N} \|(W - W^*)x_i\|_2^2\right)^{1/2} = \frac{1}{N}\|(W - W^*)X\|_F \geq \frac{\sigma_{\mathbf{min}}(X)}{N}\|W - W^*\|_F.$$

Consequently, Theorem 36 applies with

$$m_\psi = \frac{\sigma_{\mathbf{min}}(X)}{N}.$$

Again, the nonsmoothness at W^* follows from

$$h_2(W^* + tD) = t^\theta h_2(W^* + D),$$

for any matrix direction D such that $DX \neq 0$. □

F.3 Nonconvex robust matrix sensing

Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ be a linear operator, let $b = \mathcal{A}(X^*)$, and let us equip $\mathbb{R}^{m\times n}$ with the Frobenius norm.

Proposition 40. *Assume that $\mathcal{A}$ is injective, let $\theta \in (0,1)$, and define*

$$h_3(X) := \|\mathcal{A}(X) - b\|_1^\theta,$$

on a nonempty convex set

$$K \subset \{X \in \mathbb{R}^{m\times n} : \|X - X^*\|_F \leq R\}, \qquad X^* \in K.$$

Then, for every $\kappa \in (0,\theta)$, the function h_3 is (κ,γ_3)-strongly quasar-convex on K with respect to X^*, where

$$\gamma_3 = \frac{2(\theta-\kappa)}{\kappa}\, m_{\mathcal{A}}^\theta\, R^{\theta-2}, \qquad m_{\mathcal{A}} := \min_{\|U\|_F=1} \|\mathcal{A}(U)\|_1 > 0.$$

Moreover, h_3 is not locally Lipschitz at X^*, is nonconvex, and is not star-convex with respect to X^*.

Proof. Let us define

$$\psi_3(X) := \|\mathcal{A}(X - X^*)\|_1,$$

i.e.,

$$\psi_3(\lambda X^* + (1-\lambda)X) = (1-\lambda)\psi_3(X), \qquad \forall X \in K,\ \forall \lambda \in [0,1].$$

Since $\mathcal{A}$ is injective, the continuous map $U \mapsto \|\mathcal{A}(U)\|_1$ is strictly positive on the compact unit sphere $\{U : \|U\|_F = 1\}$, hence $m_{\mathcal{A}} > 0$ and

$$\psi_3(X) \geq m_{\mathcal{A}}\|X - X^*\|_F.$$

The result follows from Theorem 36 with $m_\psi = m_{\mathcal{A}}$. □