INITIALIZATION TRANSFER FOR NONCONVEX STATISTICAL ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonconvex estimation can give different solutions from different starting values. With limited target data, an estimate from a related source population can provide a useful start. Population differences can make this start inaccurate and can bias a fit that pools both datasets. In this article, we study initialization transfer, which uses the source estimate only to start fitting with target data. We show how source accuracy and population differences determine entry into a region where target fitting converges. For a fixed number of parameters, reliable entry and enough updates give the same large-sample uncertainty as the target-only local solution. In noiseless sparse signal recovery, a sufficiently close source estimate permits recovery from a target sample nearly linear in the number of nonzero coefficients. For noisy recovery from squared Gaussian measurements in fixed dimension, pooling can lose target coverage even when its source fraction tends to zero. Checking the completed fit and restarting when needed preserves nominal large-sample coverage without source accuracy conditions. Simulations and a handwritten-digit analysis show improved target fitting when target data or iterations are limited, and smaller errors than pooling at larger population differences.

1 INTRODUCTION

Many statistical estimators are computed by iterative updates of a nonconvex criterion, and different starting values can lead to different solutions (Netrapalli, Jain and Sanghavi, 2013; Candès, Li and Soltanolkotabi, 2015; Fan and Li, 2001; Zhang, 2010; Chi, Lu and Chen, 2019). When the target dataset is small, a suitable start can be hard to find. Researchers often start target fitting at an estimate from a larger, related source dataset, a practice known as fine-tuning in machine learning (Yosinski et al., 2014; Ganesh et al., 2023). In our handwritten-digit analysis, for example, a dictionary learned from source digit classes starts the fit for target classes, while only target images enter the updates. We study when a source estimate provides a reliable start and how using source data affects confidence coverage for the target population.

Local convergence theory guarantees accurate estimation from a suitable starting value (Balakrishnan, Wainwright and Yu, 2017; Loh, 2017; Tu et al., 2016), and starts constructed from target observations (Netrapalli, Jain and Sanghavi, 2013; Candès, Li and Soltanolkotabi, 2015; Tu et al., 2016; Chi, Lu and Chen, 2019; Wang, Lu and Liu, 2014) or chosen randomly (Sun, Qu and Wright, 2018; Chen et al., 2019) require enough target data to succeed. Inference with local optima allows data-dependent starts but loses coverage when the start misses the basin of the desired solution (Chen, 2023). For fine-tuning, Obst et al. (2022) express the fine-tuned linear regression estimator through the source and target least-squares fits, source initialization reduces the sample complexity of online stochastic gradient descent for single-index models (Jones-McCormick, Jagannath and Sen, 2025), and public pretraining can locate a basin in which private training then optimizes (Ganesh et al., 2023). Transfer methods that place source data in the estimating criterion include joint fitting and penalized parameter differences (Caruana, 1997; Li, Cai and Li, 2022; Tian and Feng, 2023), and Kim and Park (2026) start a target-only folded-concave correction at a pooled estimate. Supplementary Section A.1 discusses further related work. Our comparison treats a separately estimated source start, keeps the target criterion unchanged, and studies target confidence coverage relative to pooling.

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.

In this article, we study initialization transfer for nonconvex statistical estimation. It starts the target gradient updates $\theta^{t+1} = \theta^t - \alpha \nabla L_{\mathsf{T},n_\mathsf{T}}(\theta^t)$, where $L_{\mathsf{T},n_\mathsf{T}}$ is the average loss over the n_T target observations and $\alpha > 0$ is a step size, at an estimate $\widehat{\theta}_\mathsf{S}$ from independent source observations. Only target data enter the updates, which leaves the target estimating equation unchanged. We compare this procedure with pooling, which fits the source and target data jointly.

This article makes three contributions. First, we separate reaching the intended solution from estimating it accurately. For a fixed number of parameters, any start that enters the region of local target curvature, including a start that depends on the target sample, leads after enough target updates to the same large-sample distribution as the target-only local solution. A source start can change the probability of reaching this solution and the optimization error, and we bound this probability by the source estimation error and the population difference.

Second, we connect source accuracy to the sample size needed for sparse target refinement. Earlier work establishes local recovery with order $k \log(ep/k)$ Gaussian measurements for a signal with at most k nonzero coefficients among p (Soltanolkotabi, 2019; Jagatap and Hegde, 2019). We relate the required starting accuracy to source estimation error and the population difference. For signals with comparable nonzero magnitudes, sufficient conditions for diagonal spectral, truncated-power and generalized exponential spectral initialization remain quadratic in k, up to logarithmic factors (Cai, Li and Ma, 2016; Cai, Li and You, 2023; Xu, Zhang and Wang, 2026). A close source estimate can provide a start without these target initialization conditions. This is a comparison of sufficient conditions, not a lower bound for all target-only methods. In dense phase retrieval, target spectral initialization already succeeds under the sample condition of Corollary 2.

Third, we show that the two transfer modes differ for target inference at fixed dimension. For signal recovery from noisy squared Gaussian measurements, known as phase retrieval, with close but distinct source and target signals, the target coverage of pooling tends to an explicit value below the nominal level when the source sample size grows as the square root of the target sample size, and to zero for larger source samples, even when the source fraction tends to zero and the source noise has no finite variance. For initialization transfer, a check of the completed fit's norm, gradient and empirical Hessian, followed by a target-only restart when the check fails, gives nominal large-sample coverage for any source estimate. Simulations and a real data analysis of handwritten digits show that initialization transfer improves target fitting when target data or iterations are limited and outperforms pooling at larger population differences.

Sections 2–4 give the local theory and the comparison with pooling, Section 5 treats phase retrieval, and Section 6 presents the numerical studies. The Supplementary Materials give all proofs, further examples and numerical details.

2 Source Initialization and Target Refinement

We observe independent samples from a target population of interest and a related source population. Our goal is to estimate a target parameter that minimizes the target expected loss, using the source observations to provide a starting value. We use $\mathbb{E}$ and $\mathbb{P}$ for expectation and probability, and $\|\cdot\|$ and $\|\cdot\|_{\mathrm{op}}$ for Euclidean and operator norms. Superscript $\top$ denotes transpose, and ∇ and ∇^2 denote the parameter gradient and Hessian. Let p be the parameter dimension and I_p the identity. For symmetric matrices A and B, $A \preceq B$ means that $B - A$ is positive semidefinite. Let $\mathsf{N}(\mu, \Sigma)$ denote the normal distribution with mean μ and covariance Σ, and let $\Rightarrow$ and $\xrightarrow{\mathbb{P}}$ denote convergence in distribution and probability.

Let $Z_{a,1}, \ldots, Z_{a,n_a}$ be a sample of size n_a with independent observations from a distribution P_a, where $a = \mathsf{T}$ denotes the target and $a = \mathsf{S}$ the source. The two samples are independent. For a loss function ℓ_a, expectation $\mathbb{E}_a$ for a generic observation $Z \sim P_a$, and a parameter θ in the parameter space $\Theta \subset \mathbb{R}^p$, we define the population and sample criteria as

$$L_a(\theta) = \mathbb{E}_a\{\ell_a(\theta; Z)\}, \qquad L_{a,n_a}(\theta) = \frac{1}{n_a} \sum_{i=1}^{n_a} \ell_a(\theta; Z_{a,i}).$$

We define the true parameter θ_a^* as a minimizer of the population criterion L_a, and we denote the population difference by the distance $\delta = \|\theta_\mathsf{S}^* - \theta_\mathsf{T}^*\|$.

To estimate the target parameter, we compute a source estimator $\widehat{\theta}_{\mathsf{S}}$ from the source observations alone. Initialization transfer uses the starting value $u = \widehat{\theta}_{\mathsf{S}}$ and applies target refinement. With step size $\alpha > 0$ and iteration index $t = 0, 1, \ldots$, this refinement updates the parameter using only the target sample criterion:

$$\theta^0(u) = u, \qquad \theta^{t+1}(u) = \theta^t(u) - \alpha \nabla L_{\mathsf{T},n_{\mathsf{T}}}\{\theta^t(u)\}.$$

For distributional statements, we assign a fixed element of Θ to the local estimator or an iterate whenever it is undefined.

Another way to use source data is source-data transfer, which incorporates source observations into the criterion used after initialization. Our examples use pooling with a source weight $w \in (0, 1)$, which minimizes the combined criterion $(1 - w)L_{\mathsf{T},n_{\mathsf{T}}} + wL_{\mathsf{S},n_{\mathsf{S}}}$. If the population criteria L_{T} and L_{S} are differentiable at the interior minimizer θ^*_{T}, their gradients satisfy

$$\nabla L_{\mathsf{T}}(\theta^*_{\mathsf{T}}) = 0, \qquad \nabla\{(1 - w)L_{\mathsf{T}} + wL_{\mathsf{S}}\}(\theta^*_{\mathsf{T}}) = w\nabla L_{\mathsf{S}}(\theta^*_{\mathsf{T}}). \tag{1}$$

Initialization transfer leaves the target estimating equation unchanged, whereas pooling includes a source score. When the source score $\nabla L_{\mathsf{S}}(\theta^*_{\mathsf{T}})$ is nonzero, the target parameter does not solve the pooled population estimating equation.

3 Local Equivalence and Basin Entry

To study when target refinement succeeds, we define a region around the target parameter where the population loss has positive curvature. For a radius $a > 0$, let $\mathbb{B}_a = \{\theta : \|\theta - \theta^*_{\mathsf{T}}\| \le a\}$ denote this neighborhood. Condition 1 specifies the required positive curvature and stable Hessian within this region. This condition restricts only the target population, and it does not assume any sample events or properties of the source estimate.

Condition 1 (Local target geometry). *Let Θ be open. For constants $0 < r < R$, $0 < m \le L < \infty$, and $M < \infty$, $\mathbb{B}_R \subset \Theta$, L_{T} is twice continuously differentiable on $\mathbb{B}_R$, $\nabla L_{\mathsf{T}}(\theta^*_{\mathsf{T}}) = 0$, and, for every $\theta \in \mathbb{B}_R$,*

$$mI_p \preceq \nabla^2 L_{\mathsf{T}}(\theta) \preceq LI_p, \qquad \left\|\nabla^2 L_{\mathsf{T}}(\theta) - \nabla^2 L_{\mathsf{T}}(\theta^*_{\mathsf{T}})\right\|_{\mathrm{op}} \le M \left\|\theta - \theta^*_{\mathsf{T}}\right\|.$$

Condition 1 gives local population contraction. The target sample criterion must have comparable curvature. Assume it is almost surely twice continuously differentiable on $\mathbb{B}_R$. For $n = n_{\mathsf{T}}$, let deterministic sequences τ_n, Δ_n, η_n tend to zero and satisfy:

$$\mathbb{P}\left\{\|\nabla L_{\mathsf{T},n}(\theta^*_{\mathsf{T}})\| \le \tau_n,\ \sup_{\theta \in \mathbb{B}_R} \left\|\nabla^2 L_{\mathsf{T},n}(\theta) - \nabla^2 L_{\mathsf{T}}(\theta)\right\|_{\mathrm{op}} \le \Delta_n\right\} \ge 1 - \eta_n. \tag{2}$$

We require the Hessian error bound to satisfy $\Delta_n \le m/2$ and the radius to satisfy $r + 4\tau_n/m \le R$. We denote the population Hessian by $H = \nabla^2 L_{\mathsf{T}}(\theta^*_{\mathsf{T}})$ and the sample score by $S_n = \nabla L_{\mathsf{T},n}(\theta^*_{\mathsf{T}})$. To bound the contraction, we define the sample curvature bounds $m_n = m - \Delta_n$ and $L_n = L + \Delta_n$. For a step size $\alpha \in (0, 2/L_n)$, we let $\kappa_n = \max\{|1 - \alpha m_n|, |1 - \alpha L_n|\} < 1$ denote the contraction factor. Theorem 1 establishes that, with probability at least $1 - \eta_n$, every start in the inner ball $\mathbb{B}_r$ converges to the same target local estimator.

Theorem 1. *Assume Condition 1,* (2)*, $\Delta_n \le m/2$, and $r + 4\tau_n/m \le R$. With probability at least $1 - \eta_n$, $L_{\mathsf{T},n}$ has a unique stationary point $\widehat{\theta}_{\mathsf{T}}$ in $\mathbb{B}_R$, and, for every $u \in \mathbb{B}_r$ and $t \ge 0$,*

$$\theta^t(u) \in \mathbb{B}_R, \qquad \left\|\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\| \le \frac{2}{m}\|S_n\|, \qquad \left\|\theta^t(u) - \widehat{\theta}_{\mathsf{T}}\right\| \le \kappa_n^t \left\|u - \widehat{\theta}_{\mathsf{T}}\right\|.$$

Uniformly over $u \in \mathbb{B}_r$,

$$\left\|\theta^t(u) - \theta^*_{\mathsf{T}} + H^{-1}S_n\right\| \le C\left\{\Delta_n \|S_n\| + \|S_n\|^2 + \kappa_n^t\left(r + \frac{2}{m}\|S_n\|\right)\right\},$$

where C depends only on m and M.

Theorem 1 separates optimization error, which decreases geometrically, from terms determined by target sampling. Let t_n be nonnegative integer iteration counts and Z an $\mathbb{R}^p$-valued random vector. The next corollary gives the target local solution's limiting distribution when optimization error is negligible.

Corollary 1. *Let p be fixed. Assume the conditions of Theorem 1 except for the Hessian Lipschitz bound in Condition 1. If $n^{1/2}S_n \Rightarrow Z$ and $n^{1/2}\kappa_n^{t_n} r \to 0$, then, for every sequence of starts u_n with $\mathbb{P}\{u_n \in \mathbb{B}_r\} \to 1$,*

$$n^{1/2}\{\theta^{t_n}(u_n) - \theta_{\mathsf{T}}^*\} \Rightarrow -H^{-1}Z$$

and the same limit holds for $\widehat{\theta}_{\mathsf{T}}$.

Because the bound in Theorem 1 holds uniformly over the target ball, Corollary 1 applies to any random start that enters this ball with probability tending to one. This includes starts that depend on the target sample. For fixed p, the Supplementary Materials verify the concentration bound (2) using finite score variance and an integrable local Hessian bound.

To ensure source entry into $\mathbb{B}_r$, suppose a separate source analysis provides a deterministic error radius $a_{\mathsf{S}}(s)$, for $s \geq 1$, satisfying:

$$\mathbb{P}\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{S}}^*\right\| > a_{\mathsf{S}}(s)\} \leq e^{-s}. \tag{3}$$

Proposition 1 uses this source accuracy to give a sufficient condition for entry into the target ball $\mathbb{B}_r$. For its asymptotic conclusion, we define a deterministic sequence $s_n \geq 1$.

Proposition 1. *Assume the conditions of Theorem 1 and* (3)*. If $\delta + a_{\mathsf{S}}(s) \leq r$, then the conclusions of Theorem 1 hold for $u = \widehat{\theta}_{\mathsf{S}}$ with probability at least $1-\eta_n-e^{-s}$. For the distributional conclusion, assume instead the local, score, and iteration conditions of Corollary 1 and* (3)*. If $\delta + a_{\mathsf{S}}(s_n) \leq r$, and $\eta_n + e^{-s_n} \to 0$, then*

$$n^{1/2}\{\theta^{t_n}(\widehat{\theta}_{\mathsf{S}}) - \theta_{\mathsf{T}}^*\} \Rightarrow -H^{-1}Z.$$

Proposition 1 requires source error to fit within the margin $r - \delta$. Let $\tau_{\mathsf{S}} : (0, \infty) \to [0, 1]$ be a nonincreasing deviation bound. Theorem 2 quantifies entry when the source population parameter lies inside or outside $\mathbb{B}_r$.

Theorem 2. *Suppose that*

$$\mathbb{P}\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{S}}^*\right\| \geq t\} \leq \tau_{\mathsf{S}}(t), \qquad t > 0.$$

If $\delta < r$, then

$$\mathbb{P}\{\widehat{\theta}_{\mathsf{S}} \in \mathbb{B}_r\} \geq 1 - \tau_{\mathsf{S}}(r - \delta).$$

If $\delta > r$, then

$$\mathbb{P}\{\widehat{\theta}_{\mathsf{S}} \in \mathbb{B}_r\} \leq \tau_{\mathsf{S}}(\delta - r).$$

For fixed r and δ, Theorem 2 implies that a consistent source estimator enters $\mathbb{B}_r$ with probability tending to one when $\delta < r$ and to zero when $\delta > r$. The certified ball is a sufficient region and need not be the full attraction region. Supplementary Section D gives the exact entry probability of a Gaussian source estimate as in Supplementary Corollary 5 and a validation rule that bounds the probability of accepting a completed fit far from θ_{T}^* as in Supplementary Proposition 5.

4 Initialization and Source-Data Transfer

To compare the variance gain from pooling with the entry gain from initialization transfer, we study a local Gaussian model. We define $H^{1/2}$ as the symmetric positive definite square root of H, measure distances in the Hessian norm $\|x\|_H = (x^\top H x)^{1/2}$, and let $\delta_H = \|\theta_{\mathsf{S}}^* - \theta_{\mathsf{T}}^*\|_H$ denote the Hessian-scaled population difference. For fixed known scales $\sigma_{\mathsf{T}}, \sigma_{\mathsf{S}} > 0$, we assume the local estimators satisfy

$$H^{1/2}(\widehat{\theta}_{\mathsf{T}} - \theta_{\mathsf{T}}^*) \sim \mathsf{N}(0, \sigma_{\mathsf{T}}^2 n_{\mathsf{T}}^{-1} I_p), \qquad H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{S}}^*) \sim \mathsf{N}(0, \sigma_{\mathsf{S}}^2 n_{\mathsf{S}}^{-1} I_p),$$

independently. For $a \in \{\mathsf{T}, \mathsf{S}\}$, we write $v_a = p\sigma_a^2/n_a$ for the expected squared H-norm error about θ_a^*. For a fixed pooling weight $\lambda \in [0, 1]$, we combine the estimators as $\widehat{\theta}_\lambda = (1 - \lambda)\widehat{\theta}_{\mathsf{T}} + \lambda\widehat{\theta}_{\mathsf{S}}$, and we define a benchmark risk function

$$D_*(x) = \frac{v_{\mathsf{T}}(v_{\mathsf{S}} + x^2)}{v_{\mathsf{T}} + v_{\mathsf{S}} + x^2}.$$

To quantify optimality for a known bound $D \geq 0$, we define $\mathcal{M}(D)$ as the minimax expected squared H-norm error for estimating θ^*_{T}. This minimax risk evaluates estimators measurable in $(\widehat{\theta}_{\mathsf{T}}, \widehat{\theta}_{\mathsf{S}})$ over all mean pairs in $\mathbb{R}^p \times \mathbb{R}^p$ with a Hessian-scaled population difference $\delta_H \leq D$. Proposition 2 calculates the exact risk of every pooling weight and bounds the minimax benchmark.

Proposition 2. *For $\lambda \in [0,1]$,*

$$\mathcal{R}(\widehat{\theta}_\lambda) = \mathbb{E}\left\|\widehat{\theta}_\lambda - \theta^*_{\mathsf{T}}\right\|_H^2 = (1-\lambda)^2 v_{\mathsf{T}} + \lambda^2(v_{\mathsf{S}} + \delta_H^2).$$

The minimum over λ is $D_*(\delta_H)$. For every $D \geq 0$,

$$\tfrac{1}{8} D_*(D) \leq \mathcal{M}(D) \leq D_*(D).$$

Proposition 2 shows that pooling can reduce variance but introduces a squared-bias cost $\lambda^2\delta_H^2$, and that the weight $\lambda = v_{\mathsf{T}}/(v_{\mathsf{T}} + v_{\mathsf{S}} + D^2)$ attains the minimax risk over $\delta_H \leq D$ up to a factor of 8. Initialization transfer avoids this bias and changes only the probability of reaching the target local estimator. At small population differences pooling can be preferred. At intermediate differences a source start can enter the target region while pooling pays a material squared-bias cost, and for fixed $\delta > r$ the entry probability of a consistent source estimate tends to zero by Theorem 2. Supplementary Section F gives the exact risk crossover between the two modes as in Supplementary Proposition 9 and a validation rule that chooses between completed fits with risk close to that of the better one as in Supplementary Proposition 10.

5 Phase Retrieval

To verify every step of the theory from the observation model to the target error, we study phase retrieval. We represent the unknown signal by a vector. Each observation is a noisy squared projection of this signal, where the measurement weights are known but the sign is lost. For each study $a \in \{\mathsf{T}, \mathsf{S}\}$, we assume

$$Y_{a,i} = (X_{a,i}^\top \theta_a^*)^2 + \varepsilon_{a,i}, \qquad X_{a,i} \sim \mathsf{N}(0, I_p), \qquad \|\theta_a^*\| = 1,$$

where $X_{a,i}$ contains the measurement weights and $\varepsilon_{a,i}$ is centered noise independent of $X_{a,i}$. We bound the noise using the sub-Gaussian norm $\|\varepsilon_{a,i}\|_{\psi_2} \leq \sigma$, where $\|W\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}\exp(W^2/c^2) \leq 2\}$ for a real random variable W. Because reversing all signal signs leaves the observations unchanged, we measure distance using the sign-invariant metric $d_\pm(u,v) = \min\{\|u - v\|, \|u + v\|\}$. To estimate the target signal, we apply gradient descent to the sample criterion

$$F_{\mathsf{T},n_{\mathsf{T}}}(u) = \frac{1}{4n_{\mathsf{T}}} \sum_{i=1}^{n_{\mathsf{T}}} \{(X_{\mathsf{T},i}^\top u)^2 - Y_{\mathsf{T},i}\}^2.$$

To construct the source start, we compute the sample moment matrix $\widehat{M}_{\mathsf{S}} = n_{\mathsf{S}}^{-1} \sum_i Y_{\mathsf{S},i} X_{\mathsf{S},i} X_{\mathsf{S},i}^\top$ and extract its unit leading eigenvector $\widehat{v}_{\mathsf{S}}$. We also compute the sample mean $\overline{Y}_{\mathsf{S}} = n_{\mathsf{S}}^{-1} \sum_i Y_{\mathsf{S},i}$. We define the source estimator as $\widehat{\theta}_{\mathsf{S}} = \{\max(\overline{Y}_{\mathsf{S}}, 0)\}^{1/2} \widehat{v}_{\mathsf{S}}$. To state the source error bound, we define, for $n \geq 2$ and $s \geq 1$,

$$b_{n,p}(s) = \sqrt{\frac{p+s}{n}} + \frac{(s + \log n)(p+s)}{n}.$$

Under its source and target sample conditions, Corollary 2 establishes that enough gradient steps from this source start attain the target-only error rate with high probability.

Corollary 2. *Fix $s > \log 6$ and $n_{\mathsf{S}}, n_{\mathsf{T}} \geq 2$. There are constants $c_\sigma, C_\sigma < \infty$, depending only on σ, such that, if*

$$d_\pm(\theta^*_{\mathsf{S}}, \theta^*_{\mathsf{T}}) + C_\sigma b_{n_{\mathsf{S}},p}(s) \leq \tfrac{1}{8},$$

and $n_{\mathsf{T}} \geq C_\sigma(p + \log n_{\mathsf{T}})^2(\log n_{\mathsf{T}})^2$, then gradient descent on $F_{\mathsf{T},n_{\mathsf{T}}}$, initialized at $\widehat{\theta}_{\mathsf{S}}$ with step size $1/8$, satisfies, for every $t \geq 0$,

$$d_\pm(\theta^t, \theta^*_{\mathsf{T}}) \leq C_\sigma \sqrt{\frac{p + \log n_{\mathsf{T}}}{n_{\mathsf{T}}}} + \frac{3}{20}\left(\frac{15}{16}\right)^t$$

with probability at least $1 - 6e^{-s} - c_\sigma n_{\mathsf{T}}^{-1}$.

Corollary 2 gives the target-only error rate after enough updates. Its proof combines source spectral localization with uniform target gradient and Hessian bounds. The quadratic sample condition is sharp up to logarithmic factors for this uniform argument: fixed-ball Hessian control fails when $p \to \infty$ and $n_{\mathsf{T}}/p^2 \to 0$. Under the same sample condition, the target spectral start also enters the radius-$1/8$ ball. Supplementary Sections O.3 and J.3 give these calculations. Thus this dense model does not establish a lower target sample order from source initialization.

5.1 When the start limits the target sample size

A source estimate can provide the starting accuracy needed for target refinement when a target initialization guarantee requires more observations. In sparse phase retrieval, the target signal $\theta^*_{\mathsf{T}} \in \mathbb{R}^p$ has at most k nonzero entries for an integer $1 \le k \le p$, and we observe $Y_{\mathsf{T},i} = (X_{\mathsf{T},i}^\top \theta^*_{\mathsf{T}})^2$ with independent $X_{\mathsf{T},i} \sim \mathsf{N}(0, I_p)$ for $i = 1, \dots, n_{\mathsf{T}}$. We write $\|z\|_0$ for the number of nonzero entries of z and $\mathcal{H}_k(z)$ for the vector that keeps the k entries of z with the largest absolute values and sets the others to zero, resolving ties by a fixed rule. Starting from z_0, the target refinement computes

$$z_{t+1} = \mathcal{H}_k\Big[z_t - \frac{1}{n_{\mathsf{T}}} \sum_{i=1}^{n_{\mathsf{T}}} \{|X_{\mathsf{T},i}^\top z_t| - Y_{\mathsf{T},i}^{1/2}\} \operatorname{sgn}(X_{\mathsf{T},i}^\top z_t) X_{\mathsf{T},i}\Big], \tag{4}$$

which is projected gradient descent with unit step on the amplitude loss $(2n_{\mathsf{T}})^{-1} \sum_i \{Y_{\mathsf{T},i}^{1/2} - |X_{\mathsf{T},i}^\top z|\}^2$, with $\operatorname{sgn}(0) = 0$.

Corollary 3. *Let $\theta^*_{\mathsf{T}} \in \mathbb{R}^p \setminus \{0\}$ be fixed with $\|\theta^*_{\mathsf{T}}\|_0 \le k$, and let the target sample above be independent of the source sample. Let $\widehat{\theta}_{\mathsf{S}}$ be computed from the source sample with $\big\|\widehat{\theta}_{\mathsf{S}}\big\|_0 \le k$, and set $z_0 = \widehat{\theta}_{\mathsf{S}}$ in* (4)*. There are numerical constants $r_0, c, \gamma > 0$ such that, if $n_{\mathsf{T}} \ge c\, k \log(ep/k)$, then with probability at least*

$$1 - 9e^{-\gamma n_{\mathsf{T}}} - \mathbb{P}\{d_\pm(\widehat{\theta}_{\mathsf{S}}, \theta^*_{\mathsf{T}}) > r_0 \|\theta^*_{\mathsf{T}}\|\},$$

$d_\pm(z_t, \theta^*_{\mathsf{T}}) \le (2/3)^t r_0 \|\theta^*_{\mathsf{T}}\|$ for every $t \ge 0$.

Corollary 3 connects local sample requirements to source entry. Supplementary Section J.3 proves the update guarantee for a small absolute radius r_0. For a k-sparse source population signal, a source error bound a_{S} gives entry after hard thresholding when the population distance plus $2a_{\mathsf{S}}$ is at most $r_0\|\theta^*_{\mathsf{T}}\|$. Modern initialization bounds improve with energy concentration but remain quadratic for equal-magnitude nonzero coefficients (Cai, Li and You, 2023; Xu, Zhang and Wang, 2026).

5.2 Target inference and pooling

To study asymptotic inference, we fix p and write $n = n_{\mathsf{T}}$, $N = n + n_{\mathsf{S}}$, and $w_N = n_{\mathsf{S}}/N$. We replace the sub-Gaussian condition by a finite first absolute source-noise moment and a positive finite target-noise variance, and we denote a generic noise variable in study a by ε_a. The source-initialized target estimate $\widehat{u}_n$ and the pooled estimate $\widehat{u}_P$, whose start is the spectral estimate from both samples, use step size $1/8$ and the stopping rule (14). For a criterion with sample size m, this rule stops at the first iterate whose gradient norm is at most $\{m \log(m+1)\}^{-1}$, or at a cap of order $(\log m)^2$. For $0 < \beta < 1$, let $\chi^2_{p,1-\beta}$ be the $1-\beta$ quantile of the chi-square distribution with p degrees of freedom. For an estimate $\widehat{u}$, a covariance estimate $\widehat{V}$ and a sample size m, the sign-symmetric Wald region is $\{u : m \min_{s\in\{-1,1\}} (u - s\widehat{u})^\top \widehat{V}^{-1} (u - s\widehat{u}) \le \chi^2_{p,1-\beta}\}$. The target region $\mathcal{C}_n(\beta)$ uses $\widehat{u}_n$, size n, and covariance $\widehat{\nu}_n^2 (2I_p + 4\widehat{u}_n\widehat{u}_n^\top)^{-1}/2$, where $\widehat{\nu}_n^2 = n^{-1} \sum_i \{Y_{\mathsf{T},i} - (X_{\mathsf{T},i}^\top \widehat{u}_n)^2\}^2$. The pooled region $\mathcal{C}_P(\beta)$ uses $\widehat{u}_P$, size N, and the pooled sandwich covariance $\widehat{V}_P$ with scores centered within each study (Supplementary Section H). A region is empty if its estimate or covariance is undefined or its covariance is not positive definite.

To analyze pooling bias, we fix unit signals with $0 < d_\pm(\theta^*_{\mathsf{S}}, \theta^*_{\mathsf{T}}) < 1/8$ and align their signs. We define the inner product $c = (\theta^*_{\mathsf{S}})^\top \theta^*_{\mathsf{T}}$ and the target Hessian $H = 2I_p + 4\theta^*_{\mathsf{T}} (\theta^*_{\mathsf{T}})^\top$. We define

$$g^*_{\mathsf{S}} = 2\{\theta^*_{\mathsf{T}} - c\theta^*_{\mathsf{S}}\}, \qquad b = -H^{-1} g^*_{\mathsf{S}}, \qquad V_{\mathsf{T}} = \frac{\nu_{\mathsf{T}}^2}{2} H^{-1},$$

where ν_{T}^2 is the target-noise variance. The term g_{S}^* is the expected source score at the target signal, which matches $\nabla L_{\mathsf{S}}(\theta_{\mathsf{T}}^*)$ in (1) under finite source variance. The term b is the coefficient of the first-order pooled displacement, and V_{T} is the target asymptotic covariance. We define θ_w as the minimizer of the pooled population criterion with source weight w, after subtracting $Y_{a,i}^2/4$ from each observation loss. This subtraction leaves the gradients unchanged and makes the population criterion finite under the stated moments. We choose $\theta_w^\top \theta_{\mathsf{T}}^* > 0$ as described in Supplementary Section Q. We let $\widetilde{u}_P$ denote the member of $\{\widehat{u}_P, -\widehat{u}_P\}$ closest to θ_{T}^*, and we let $G_{p,\lambda}$ denote the noncentral chi-square distribution function with p degrees of freedom and noncentrality λ. Theorem 3 establishes the asymptotic limit of the pooled estimator and calculates its limiting target coverage probability.

Theorem 3. *Assume fixed p, the stated Gaussian designs and independent centered noise, and*

$$n_{\mathsf{S}}, n \to \infty, \qquad \frac{n_{\mathsf{S}}}{\sqrt{n}} \to a \in [0, \infty), \qquad \mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty, \qquad 0 < \nu_{\mathsf{T}}^2 < \infty.$$

Then

$$\sqrt{n}(\widetilde{u}_P - \theta_{\mathsf{T}}^*) \Rightarrow \mathsf{N}(ab, V_{\mathsf{T}}), \qquad \widehat{V}_P \xrightarrow{\mathbb{P}} V_{\mathsf{T}},$$

and

$$\mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_P(\beta)\} \longrightarrow G_{p, a^2 b^\top V_{\mathsf{T}}^{-1} b}(\chi_{p,1-\beta}^2).$$

Also, $\mathbb{P}\{\theta_{w_N} \in \mathcal{C}_P(\beta)\} \to 1 - \beta$.

Theorem 3 shows that pooling has limiting target coverage strictly below the nominal level whenever $a > 0$, although the source weight w_N tends to zero, and Supplementary Corollary 6 shows that target coverage tends to zero when $n_{\mathsf{S}}/\sqrt{n} \to \infty$. Local misspecification produces shifted limits in other settings (Hjort and Claeskens, 2003), and here the shift follows from the source sample size, the signal difference and the computed pooled fit. For initialization transfer, Supplementary Theorem 8 gives nominal target coverage for any fixed nonorthogonal source direction, whereas pooling with a nonvanishing source fraction has zero limiting target coverage for these signal pairs (Supplementary Proposition 22).

To allow inaccurate source starts, we check the completed fit's norm, gradient and empirical Hessian as in (9). We restart from the target spectral estimate if the check fails. Write $\widehat{u}_n^c$ for the accepted estimate and $\widetilde{u}_n^c$ for its sign closest to θ_{T}^*. Its region $\mathcal{C}_n^c(\beta)$ uses the target covariance above. If no fit passes, we assign estimate zero. We report $\mathbb{R}^p$ if no fit passes or the residual variance is zero or undefined. The next theorem permits any finite list of source starts and selection among accepted fits using target data.

Theorem 4. *Assume the Gaussian target phase model, fixed p, and*

$$\|\theta_{\mathsf{T}}^*\| = 1, \qquad \mathbb{E}\varepsilon_{\mathsf{T}} = 0, \qquad 0 < \nu_{\mathsf{T}}^2 = \mathbb{E}\varepsilon_{\mathsf{T}}^2 < \infty, \qquad \varepsilon_{\mathsf{T}} \text{ independent of } X_{\mathsf{T}}.$$

For every finite measurable list of source starts, the procedure of Supplementary Section K.2 accepts a fit with probability tending to one, and

$$\sqrt{n}(\widetilde{u}_n^c - \theta_{\mathsf{T}}^*) \Rightarrow \mathsf{N}(0, V_{\mathsf{T}}), \qquad \mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_n^c(\beta)\} \longrightarrow 1 - \beta.$$

Theorem 4 removes source accuracy conditions because every accepted fit approaches the same target local estimator after sign alignment. Supplementary Theorem 9 extends the procedure to unknown Gaussian design covariance and signal scale. Supplementary Section P gives simultaneous intervals for prediction intensities.

6 Numerical Studies

Our numerical studies ask whether a source start lowers the target sample size needed for recovery, whether pooling loses target coverage at the predicted source sample sizes, and whether a source start improves target fitting in finite samples and in real data.

We first study noiseless sparse phase retrieval with $p = 1000$, $k = 10$, target entries $\pm k^{-1/2}$ on a random support, and $n_{\mathsf{S}} = 3000$. Equal nonzero magnitudes give a case in which published initialization bounds remain quadratic in sparsity. The source signal shares the target support and has

Table 1: Success rates in noiseless sparse phase retrieval over 100 repetitions.

δ	n_{T}	Source init.	Target spectral	Target TPM	Target random	Pooled
0.05	150	1.00	0.00	0.00	0.00	0.00
0.05	300	1.00	0.07	0.15	0.00	0.00
0.05	600	1.00	0.69	0.75	0.00	0.00
0.20	150	1.00	0.01	0.00	0.00	0.00
0.20	300	1.00	0.04	0.17	0.00	0.00
0.20	600	1.00	0.58	0.78	0.00	0.00

sign-invariant distance $\delta \in \{0.05, 0.20\}$ from the target. Every fit uses (4). We compare source initialization with diagonal spectral, random and truncated-power target starts, and pooling. Truncated-power initialization, denoted by TPM, follows Cai, Li and You (2023), with five power sparsities selected by target training loss after refinement. The target sizes $n_{\mathsf{T}} \in \{150, 300, 600\}$ lie between $k \log(ep/k) \approx 56$ and $k^2 \log p \approx 691$. Success means relative sign-invariant error at most 10^{-6}. Each configuration has 100 repetitions.

Table 1 shows recovery by source initialization in every repetition, with relative error below 10^{-12}. Truncated-power starts improve on diagonal spectral starts at $n_{\mathsf{T}} = 300, 600$, but succeed in only 15%–17% of repetitions at 300 and 75%–78% at 600. The source start gives accurate recovery with fewer target observations in this flat-signal design. These experiments do not calibrate the constants in Corollary 3. Pooling has median target errors 0.042–0.048 at $\delta = 0.05$ and 0.169–0.192 at $\delta = 0.20$.

For noisy dense signals, we examine how source sample size affects pooling coverage. Figure 1 examines the 95% confidence regions of Theorem 3 with $p = 4$, a sign-invariant signal distance of 0.10, Gaussian noise with standard deviation 0.35, and $n_{\mathsf{S}} = \lceil a\sqrt{n_{\mathsf{T}}} \rceil$, where $\lceil x \rceil$ is the smallest integer at least x. We run 200 repetitions per configuration with the fixed step size and the gradient stopping rule. At $n_{\mathsf{T}} = 16{,}384$, pooled target coverage decreases from 0.940 at $a = 0.5$ to 0.060 at $a = 8$, while its coverage of the pooled population parameter is 0.945 and 0.960. The coverage of initialization transfer remains close to that of an oracle target start. With source noise having a finite mean and infinite variance, its target coverage is between 0.930 and 0.980 (Supplementary Table 16). When the source noise has no finite mean, the checked procedure of Theorem 4 raises target coverage from 0.345 to 0.945 at $n_{\mathsf{T}} = 4096$ (Supplementary Section V).

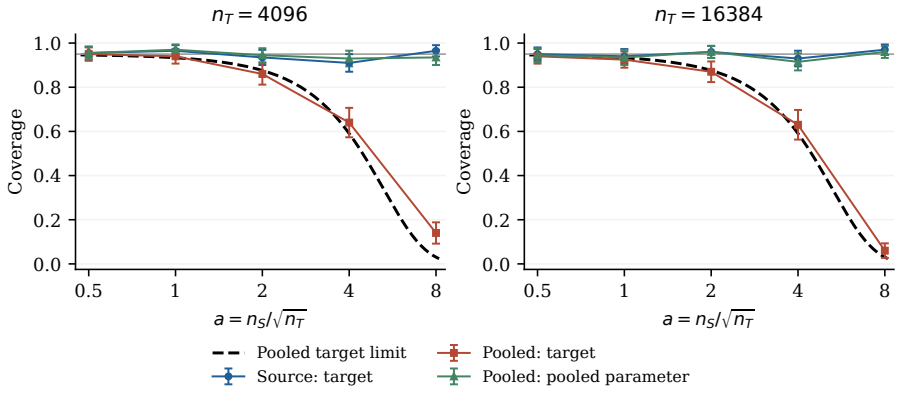

Figure 1: Target and pooled-parameter coverage across square-root source sample sizes.

To evaluate target fitting when the target sample is small relative to the dimension, we set $p = 24$, $n_{\mathsf{T}} = 55$, $n_{\mathsf{S}} = 320$, and Gaussian noise with standard deviation 0.35. The source signal is $(\theta^*_{\mathsf{T}} + \rho v)/\sqrt{1+\rho^2}$ for a unit vector v orthogonal to θ^*_{T}, with design shift $\rho \in \{0.35, 0.55, 0.75\}$, and we use 100 repetitions per configuration. The random comparator selects the best of eight starts by target training loss. Spectral starts use a stabilized scale, and refinement uses backtracking line search.

Table 2: Mean sign-invariant target error over 100 repetitions.

ρ	Pooled	Source init.	Random restart	Target spectral
0.35	0.295	0.250	0.373	0.924
0.55	0.443	0.252	0.430	0.875
0.75	0.563	0.277	0.371	0.913

Table 2 shows that initialization transfer has smaller mean error than pooling and both target-only starts at each of the three design shifts. Its paired win rates against pooling are 0.75, 0.96, and 0.95, and it has smaller error than target spectral starts in 85% to 88% of repetitions and than random restarts in 65% to 75%. Against eight random starts, it also reduces the frequency of errors above 0.35, from 27%–30% to 9%–17%. Choosing between the source-initialized and pooled fits by 100 target validation observations gives mean squared H-norm error 0.1465, compared with 0.1400 for the infeasible oracle choice (Supplementary Table 6). In matrix completion, each fit uses at most 30 updates, while the random baseline uses three such starts. Initialization transfer has the smallest mean error at both source rotation angles (Supplementary Table 5).

To test initialization transfer on real data, we analyze the scikit-learn handwritten-digits dataset (Pedregosa et al., 2011), with digits 0–4 as the source and digits 5–9 as the target. We fit a rank-16 nonnegative matrix factorization (Lee and Seung, 1999; 2000) to 30 target training images in each of 100 random splits and compute reconstruction root mean squared error (RMSE) on the 866 remaining target images, with each fitted dictionary fixed. The source start is a dictionary fitted to 400 source images. The comparators are pooling, one target random start, the target-only NNDSVDa start, which applies nonnegative double singular value decomposition with mean filling, and the best of eight target random starts selected by training loss. Each target or pooled fit is capped at 200 iterations, and the iteration counts exclude source fitting and test-image coefficient estimation.

Table 3: Target test reconstruction RMSE over 100 splits. Numbers in parentheses represent standard deviations, and the last column counts target-stage iterations across all starts.

Method	Mean	Median	90% quantile	Target iter.
Source initialization	0.1503 (0.0042)	0.1505	0.1550	85.48
Target random	0.1523 (0.0039)	0.1523	0.1570	200.00
Target NNDSVDa	0.1542 (0.0042)	0.1547	0.1593	200.00
Target 8 random starts	0.1535 (0.0040)	0.1536	0.1588	1599.37
Pooled dictionary	0.1592 (0.0038)	0.1591	0.1644	200.00

Table 3 shows the smallest mean, median and 90-percentile test RMSE for initialization transfer. Relative to one target random start, mean RMSE decreases by 0.0020, or 1.3%, with a paired standard error of 0.0004. It has lower error on 69% of splits, compared with 91% against NNDSVDa and 84% against eight random starts.

7 Discussion and Limitations

A source start can raise the probability of reaching the target solution and, as in sparse phase retrieval, reduce the target sample size needed to reach it, while the target data alone determine the first-order uncertainty after reliable entry and enough updates. Pooling can reduce variance, but its population bias can remove target coverage even when the source fraction tends to zero. When the source estimate may be inaccurate, we recommend checking the completed target fit and restarting from a target-only start when the check fails. Our inference results assume a fixed dimension, the sparse guarantee concerns noiseless measurements, and our empirical comparisons use fixed target optimization budgets that exclude source estimation. Extending the local equivalence to deep networks requires a characterization of their basins that our results do not provide.

AI USE STATEMENT

We used large language model tools extensively in this work. They wrote and reviewed the numerical code, searched the literature, checked bibliographic records, and edited the text. The authors take full responsibility for all content.

ETHICS STATEMENT

This work develops statistical theory and uses simulated data and a public handwritten-digit data set. We did not collect human-subject data or use personally identifiable information.

REPRODUCIBILITY STATEMENT

The Supplementary Materials contain every proof, the model-specific verification, complete simulation and data-analysis protocols, and additional diagnostics. The supplementary archive includes the scripts, seeds, stored raw results, summaries, and metadata used to construct Tables 1–3, Figure 1, and the supplementary tables and figures.

REFERENCES

Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: from population to sample-based analysis. *Annals of Statistics* **45**, 77–120.

Cai, T. T., Li, X. and Ma, Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *Annals of Statistics* **44**, 2221–2251.

Cai, J.-F., Li, J. and You, J. (2023). Provable sample-efficient sparse phase retrieval initialized by truncated power method. *Inverse Problems* **39**, 075008.

Candès, E. J., Li, X. and Soltanolkotabi, M. (2015). Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Transactions on Information Theory* **61**, 1985–2007.

Caruana, R. (1997). Multitask learning. *Machine Learning* **28**, 41–75.

Chen, Y., Chi, Y., Fan, J. and Ma, C. (2019). Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming* **176**, 5–37.

Chen, Y.-C. (2023). Statistical inference with local optima. *Journal of the American Statistical Association* **118**, 1940–1952.

Chi, Y., Lu, Y. M. and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: an overview. *IEEE Transactions on Signal Processing* **67**, 5239–5269.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Ganesh, A., Haghifam, M., Nasr, M., Oh, S., Steinke, T., Thakkar, O., Guha Thakurta, A. and Wang, L. (2023). Why is public pretraining necessary for private model training? *Proceedings of the 40th International Conference on Machine Learning*, 10611–10627.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Jagatap, G. and Hegde, C. (2019). Sample-efficient algorithms for recovering structured signals from magnitude-only measurements. *IEEE Transactions on Information Theory* **65**, 4434–4456.

Jones-McCormick, T., Jagannath, A. and Sen, S. (2025). Provable benefits of unsupervised pretraining and transfer learning via single-index models. *Proceedings of the 42nd International Conference on Machine Learning*, 28350–28376.

Kim, J. and Park, S. (2026). Transfer learning in high-dimensional Ising models. *arXiv:2607.03005*.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* **13**, 556–562.

Li, S., Cai, T. T. and Li, H. (2022). Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society, Series B* **84**, 149–173.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics* **45**, 866–896.

Netrapalli, P., Jain, P. and Sanghavi, S. (2013). Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems* **26**, 2796–2804.

Obst, D., Ghattas, B., Claudel, S., Cugliari, J., Goude, Y. and Oppenheim, G. (2022). Improved linear regression prediction by transfer learning. *Computational Statistics & Data Analysis* **174**, 107499.

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Soltanolkotabi, M. (2019). Structured signal recovery from quadratic measurements: breaking sample complexity barriers via nonconvex optimization. *IEEE Transactions on Information Theory* **65**, 2374–2400.

Sun, J., Qu, Q. and Wright, J. (2018). A geometric analysis of phase retrieval. *Foundations of Computational Mathematics* **18**, 1131–1198.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association* **118**, 2684–2697.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M. and Recht, B. (2016). Low-rank solutions of linear matrix equations via Procrustes flow. *Proceedings of the 33rd International Conference on Machine Learning*, 964–973.

Wang, Z., Lu, H. and Liu, H. (2014). Tighten after relax: minimax-optimal sparse PCA in polynomial time. *Advances in Neural Information Processing Systems* **27**, 3383–3391.

Xu, M., Zhang, Y. and Wang, J. (2026). Achieving optimal sample complexity for a broader class of signals in sparse phase retrieval. *IEEE Transactions on Signal Processing* **74**, 1735–1750.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems* **27**, 3320–3328.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

SUPPLEMENTARY MATERIALS

These Supplementary Materials give the full model, all formal conditions and results, complete proofs, model-specific verification, and the numerical protocols. The opening sections give the detailed statements and proofs of the main results. Auxiliary concentration bounds and model calculations follow. The numerical sections report the experiments, their implementation, and paired comparisons. Lemma 9 establishes the target sample condition through quadratic Gaussian features. Proposition 23 relates the implemented validation criterion to population prediction loss and parameter error.

A MODEL, ESTIMATORS, AND TRANSFER MODES

Initialization transfer and pooling differ precisely in the estimating criterion they use after selecting a starting value. We formalize this algorithmic distinction before developing the local theory.

To model the two studies, we assume $Z_{\mathsf{T},i}$ for $i = 1, \ldots, n_{\mathsf{T}}$ are independent observations from the target distribution P_{T}. We also assume $Z_{\mathsf{S},i}$ for $i = 1, \ldots, n_{\mathsf{S}}$ are independent observations from the source distribution P_{S}. We treat these two samples as mutually independent. Within the parameter space $\Theta \subset \mathbb{R}^p$, we define the population and sample criteria for each study $a \in \{\mathsf{T}, \mathsf{S}\}$ as

$$L_a(\theta) = \mathbb{E}_a\{\ell_a(\theta; Z)\}, \qquad L_{a,n_a}(\theta) = \frac{1}{n_a} \sum_{i=1}^{n_a} \ell_a(\theta; Z_{a,i}).$$

We define the target and source population parameters as the minimizers

$$\theta^*_{\mathsf{T}} \in \arg\min_{\theta \in \Theta} L_{\mathsf{T}}(\theta), \qquad \theta^*_{\mathsf{S}} \in \arg\min_{\theta \in \Theta} L_{\mathsf{S}}(\theta),$$

and we measure the population difference through the source-target shift and its norm

$$\Delta = \theta^*_{\mathsf{S}} - \theta^*_{\mathsf{T}}, \qquad \delta = \|\Delta\|.$$

We state the local target theory using the Euclidean metric. To handle nonidentifiability, such as missing signs in phase retrieval or unknown rotations in factor models, Appendix L explains the conditions needed to extend these guarantees to quotient metrics.

To refine a starting value using only the target sample, we run gradient descent with step size $\alpha > 0$:

$$\theta^{k+1} = \theta^k - \alpha \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta^k), \qquad k = 0, 1, \ldots.$$

We define the target local estimator $\widehat{\theta}_{\mathsf{T}}$ as the stationary point of $L_{\mathsf{T},n_{\mathsf{T}}}$ strictly within a local neighborhood of θ^*_{T}, provided this point exists and is unique. To define the estimator and iterates for every sample, we assign an arbitrary fixed element of Θ whenever the local estimator or a gradient iterate remains undefined. To incorporate external information, we compute the source estimator $\widehat{\theta}_{\mathsf{S}}$ using only the source sample. We then use this source estimate to initialize the target algorithm. For pooling, we include the source observations in a combined criterion.

Definition 1. *The source is used as an initializer if $\widehat{\theta}_{\mathsf{S}}$ is used only as θ^0 in the target algorithm and the subsequent criterion is $L_{\mathsf{T},n_{\mathsf{T}}}$. The source is used as data if observations from P_{S} enter the estimating criterion used after the starting value is chosen.*

Definition 1 establishes the core conceptual separation we use throughout this analysis. Under initialization transfer, the source data affect the starting value, which can change target basin entry and finite-iteration error. Under data transfer, the source observations enter the final estimating equation. While a combined procedure can implement both operations simultaneously, these two distinct transfer mechanisms can produce different statistical distributions.

A.1 FURTHER RELATED WORK

Yang et al. (2023) study inference after a comparability test determines whether trial and observational data are combined. Our completed-fit check addresses convergence to a target solution, while every accepted estimate uses the target criterion alone.

Earlier work links starting values to prediction accuracy (Kuzborskij and Lampert, 2018), relates statistical accuracy to update stability (Ho et al., 2025), and develops debiased inference for sparse phase retrieval, including simultaneous confidence intervals (Yao, 2022). Pretraining studies examine prediction improvements (Kumar et al., 2022; Wu et al., 2022; Shachaf, Brutzkus and Globerson, 2021; Ghane, Akhtiamov and Hassibi, 2024; Tahir, Ganguli and Rotskoff, 2025), bound these gains (Mousavi Kalan et al., 2020), study how source training affects downstream transfer (Hedges, 2026), or quantify how learned features affect target prediction (McNamara and Balcan, 2017; Lauditi, Bordelon and Pehlevan, 2026; Tinati and Tu, 2026). Empirical studies examine whether learned features improve predictions and affect the solution reached (Yosinski et al., 2014; Ash and Adams, 2020; Neyshabur, Sedghi and Zhang, 2020; Sadrtdinov et al., 2023), or show that reusing parameters after fitting noise can hinder further learning (Shin et al., 2024). These prediction results do not by themselves determine whether uncertainty statements remain valid for the target population after source initialization.

Other transfer methods use source information during target fitting by fitting tasks jointly (Caruana, 1997; Baxter, 2000), adjusting for population differences (Shimodaira, 2000; Sugiyama, Krauledat and Müller, 2007; Ben-David et al., 2010; Pan and Yang, 2010), or limiting parameter differences (Bastani, 2021; Li, Cai and Li, 2022; Tian and Feng, 2023). Theory characterizes when these methods improve target estimation (Hanneke and Kpotufe, 2019; Cai and Wei, 2021; Reeve, Cannings and Samworth, 2021). Recent studies also combine features (Xu et al., 2026), select sources using language information (Chen, Zhou and Al Kontar, 2026), use source estimates in the expectation-maximization algorithm for Gaussian mixtures (Tian et al., 2026), limit source use to points near the classification boundary (Fan, Gao and Klusowski, 2025), or study uncertainty after variable selection (Tam, My and Le Duy, 2025). Our pooling comparison does not exclude improvements from these other uses of source data. Meta-learning methods learn starting values for adaptation across tasks (Finn, Abbeel and Levine, 2017; Fallah, Mokhtari and Ozdaglar, 2020; Balcan et al., 2021), while other methods adapt through intermediate populations (Kumar, Ma and Liang, 2020), permute parameters to transfer learning trajectories (Chijiwa, 2024), or change early updates to influence the solution reached by interacting agents (Shcherbinin et al., 2026). In our setting, the source estimate comes from a separate population with its own sampling error, and the target updates remain unchanged.

B Local Target Theory and First-Order Equivalence

We begin our analysis with a single condition on the target criterion. We state this condition entirely at the target population level, introducing sampling variability only through two concentration bounds. We restrict this condition to a local region. We do not require global convexity, and we make no assumptions about the criterion outside a fixed ball around the target parameter. For a radius $a > 0$, we define this ball as

$$\mathbb{B}_a = \{\theta \in \mathbb{R}^p : \|\theta - \theta^*_{\mathsf{T}}\| \le a\}.$$

Condition 2 (Local target geometry)**.** *Let $\Theta \subset \mathbb{R}^p$ be open. For constants $0 < r < R$, $0 < m \le L < \infty$, and $M < \infty$, $\mathbb{B}_R \subset \Theta$, L_{T} is twice continuously differentiable on $\mathbb{B}_R$, $\nabla L_{\mathsf{T}}(\theta^*_{\mathsf{T}}) = 0$, and*

$$mI_p \preceq \nabla^2 L_{\mathsf{T}}(\theta) \preceq LI_p, \qquad \left\|\nabla^2 L_{\mathsf{T}}(\theta) - \nabla^2 L_{\mathsf{T}}(\theta^*_{\mathsf{T}})\right\|_{\mathrm{op}} \le M \|\theta - \theta^*_{\mathsf{T}}\|$$

for all $\theta \in \mathbb{B}_R$.

Condition 2 restricts only the population properties. It requires that the target risk has positive curvature and that its Hessian deviates from its value at the target parameter by at most M times the parameter distance. We analyze sample criteria that remain twice continuously differentiable on $\mathbb{B}_R$ almost surely. This regularity follows from Condition 3 and holds for the phase-retrieval criterion. We separate the sampling properties into a distinct probability bound. We introduce deterministic sequences $\tau_{n_{\mathsf{T}}}, \Delta_{n_{\mathsf{T}}}, \eta_{n_{\mathsf{T}}}$ that tend to zero and for which a model-specific concentration result establishes

$$\mathbb{P}\left\{\|\nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta^*_{\mathsf{T}})\| \le \tau_{n_{\mathsf{T}}}, \quad \sup_{\theta \in \mathbb{B}_R} \left\|\nabla^2 L_{\mathsf{T},n_{\mathsf{T}}}(\theta) - \nabla^2 L_{\mathsf{T}}(\theta)\right\|_{\mathrm{op}} \le \Delta_{n_{\mathsf{T}}}\right\} \ge 1 - \eta_{n_{\mathsf{T}}}. \tag{5}$$

Section M proves the concentration bound in (5) under distributional tail conditions, while Section O verifies it directly for phase retrieval. The source data do not enter either target probability calcu-

lation. Instead, the source estimate determines the starting value and can affect entry into the local region. To analyze the target iteration, we define

$$H = \nabla^2 L_{\mathsf{T}}(\theta^*_{\mathsf{T}}), \qquad S_{n_{\mathsf{T}}} = \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta^*_{\mathsf{T}}),$$
$$m_{n_{\mathsf{T}}} = m - \Delta_{n_{\mathsf{T}}}, \qquad L_{n_{\mathsf{T}}} = L + \Delta_{n_{\mathsf{T}}}.$$

For a step size $\alpha \in (0, 2/L_{n_{\mathsf{T}}})$, we initialize the gradient iteration from Section A at $u \in \mathbb{B}_r$ by setting $\theta^0(u) = u$. The subsequent iterations follow

$$\theta^{t+1}(u) = \theta^t(u) - \alpha \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta^t(u)), \qquad t = 0, 1, 2, \ldots,$$

and, when $\Delta_{n_{\mathsf{T}}} \le m/2$, we compute their contraction factor as

$$\kappa_{n_{\mathsf{T}}} = \max\{|1 - \alpha m_{n_{\mathsf{T}}}|,\ |1 - \alpha L_{n_{\mathsf{T}}}|\} < 1.$$

Under its stated conditions, Theorem 5 establishes the existence and uniqueness of the target local estimator with probability at least $1 - \eta_{n_{\mathsf{T}}}$. On this event, it also proves uniform contraction for all starts in the certified ball and gives a first-order expansion with a uniform remainder bound.

Theorem 5. *Assume Condition 2,* (5)*, $\Delta_{n_{\mathsf{T}}} \le m/2$, and $r + 4\tau_{n_{\mathsf{T}}}/m \le R$, and let $\alpha \in (0, 2/L_{n_{\mathsf{T}}})$. With probability at least $1 - \eta_{n_{\mathsf{T}}}$, $L_{\mathsf{T},n_{\mathsf{T}}}$ has a unique stationary point $\widehat{\theta}_{\mathsf{T}}$ in $\mathbb{B}_R$,*

$$\left\|\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\| \le \frac{2}{m} \left\|S_{n_{\mathsf{T}}}\right\|,$$

and, simultaneously for every $u \in \mathbb{B}_r$ and every $t \ge 0$,

$$\theta^t(u) \in \mathbb{B}_R, \qquad \left\|\theta^t(u) - \widehat{\theta}_{\mathsf{T}}\right\| \le \kappa_{n_{\mathsf{T}}}^t \left\|u - \widehat{\theta}_{\mathsf{T}}\right\|,$$
$$\left\|\theta^t(u) - \theta^*_{\mathsf{T}}\right\| \le \frac{2}{m} \left\|S_{n_{\mathsf{T}}}\right\| + \kappa_{n_{\mathsf{T}}}^t \left(r + \frac{2}{m} \left\|S_{n_{\mathsf{T}}}\right\|\right),$$

and

$$\left\|\theta^t(u) - \theta^*_{\mathsf{T}} + H^{-1} S_{n_{\mathsf{T}}}\right\| \le C \left\{\Delta_{n_{\mathsf{T}}} \left\|S_{n_{\mathsf{T}}}\right\| + \left\|S_{n_{\mathsf{T}}}\right\|^2 + \kappa_{n_{\mathsf{T}}}^t \left(r + \frac{2}{m} \left\|S_{n_{\mathsf{T}}}\right\|\right)\right\},$$

where $C < \infty$ depends only on m and M.

Theorem 5 uniformly bounds the remainder over all starts in $\mathbb{B}_r$. While the remainder at a finite iteration can depend on the specific initialization, our bound depends on the target sample, concentration bounds, iteration count, local constants, and starting radius. Once the optimization term is negligible at the target statistical scale, Corollary 4 gives the same target first-order limit under its stated conditions. The source estimate affects the probability of successfully reaching this certified region.

Remark 1 (General contractive maps). The proof of Theorem 5 uses gradient descent only through the contraction property of the map $T_{n_{\mathsf{T}}}(\theta) = \theta - \alpha \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta)$ around $\widehat{\theta}_{\mathsf{T}}$ on $\mathbb{B}_R$. If an iterative map $T_{n_{\mathsf{T}}} : \mathbb{B}_R \to \mathbb{B}_R$ satisfies, for an error bound $e_{n_{\mathsf{T}}} \ge 0$, $\left\|T_{n_{\mathsf{T}}}(\theta) - \widehat{\theta}_{\mathsf{T}}\right\| \le \kappa \left\|\theta - \widehat{\theta}_{\mathsf{T}}\right\| + e_{n_{\mathsf{T}}}$ with $0 \le \kappa < 1$, the same induction gives $\left\|\theta^t(u) - \widehat{\theta}_{\mathsf{T}}\right\| \le \kappa^t \left\|u - \widehat{\theta}_{\mathsf{T}}\right\| + e_{n_{\mathsf{T}}}/(1-\kappa)$. This extension gives the optimization bound only. The first-order conclusion also requires a model-specific expansion of the target-sample fixed point and $e_{n_{\mathsf{T}}}/(1-\kappa) = o(n_{\mathsf{T}}^{-1/2})$. Related EM results bound optimization error by a geometric term and a statistical remainder (Balakrishnan, Wainwright and Yu, 2017). Appendix R verifies local target equivalence directly for a fixed-dimensional Gaussian mixture.

Corollary 4 takes this uniform expansion to the limit to establish the asymptotic equivalence. To derive this limit under the remaining conditions of the corollary, population Hessian continuity suffices in place of the Lipschitz bound. In contrast, we use the Lipschitz bound for the explicit finite-sample remainder in Theorem 5.

Corollary 4. *Let p be fixed. Assume the conditions of Theorem 5 except for the Hessian Lipschitz bound in Condition 2. If*

$$n_{\mathsf{T}}^{1/2} S_{n_{\mathsf{T}}} \Rightarrow Z, \qquad n_{\mathsf{T}}^{1/2} \kappa_{n_{\mathsf{T}}}^{t_{n_{\mathsf{T}}}} r \to 0,$$

then, for every sequence of starts $u_{n_{\mathsf{T}}}$ with $\mathbb{P}\{u_{n_{\mathsf{T}}} \in \mathbb{B}_r\} \to 1$,

$$n_{\mathsf{T}}^{1/2} \{\theta^{t_{n_{\mathsf{T}}}}(u_{n_{\mathsf{T}}}) - \theta^*_{\mathsf{T}}\} \Rightarrow -H^{-1} Z.$$

If $Z \sim \mathsf{N}(0, \Sigma)$, the limit is $\mathsf{N}(0, H^{-1}\Sigma H^{-1})$, the same limit as the target local estimator $\widehat{\theta}_{\mathsf{T}}$.

Theorem 5 guarantees a uniform remainder bound across the entire certified ball. By this uniformity, Corollary 4 applies to every random start that enters this ball with probability tending to one, even if the start depends on the target sample. This final limit exactly matches that of the target local estimator. For fixed p, Proposition 12 derives these target conditions from a mean-zero score with finite variance and an integrable local Hessian bound, under its other stated conditions. We next bound the probability that the source start enters $\mathbb{B}_r$ using the source estimation error.

We assume that an independent source-estimator analysis provides a deterministic radius $a_{\mathsf{S}}(s)$ such that, for every $s \geq 1$,

$$\mathbb{P}\left\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| > a_{\mathsf{S}}(s)\right\} \leq e^{-s}. \tag{6}$$

Section N derives general sufficient conditions for this bound, while Proposition 18 proves a model-specific bound for the phase-retrieval spectral estimator.

Proposition 3 combines the target geometry with this source localization bound. It proves that sufficient source accuracy and small source-target separation jointly give entry into $\mathbb{B}_r$ with high probability. For the asymptotic limit, we define a deterministic sequence $s_{n_{\mathsf{T}}} \geq 1$.

Proposition 3. *Assume the conditions of Theorem 5 and* (6)*. Fix* $s_{\mathsf{S}} \geq 1$*. If*

$$\delta + a_{\mathsf{S}}(s_{\mathsf{S}}) \leq r,$$

then the conclusions of Theorem 5 hold for target gradient descent initialized at $\widehat{\theta}_{\mathsf{S}}$ *with probability at least* $1 - \eta_{n_{\mathsf{T}}} - e^{-s_{\mathsf{S}}}$*. For the distributional conclusion, assume instead the local, score, and iteration conditions of Corollary 4 and* (6)*. If* $\delta + a_{\mathsf{S}}(s_{n_{\mathsf{T}}}) \leq r$*, and* $\eta_{n_{\mathsf{T}}} + e^{-s_{n_{\mathsf{T}}}} \to 0$*, then*

$$n_{\mathsf{T}}^{1/2}\{\theta^{t_{n_{\mathsf{T}}}}(\widehat{\theta}_{\mathsf{S}}) - \theta^*_{\mathsf{T}}\} \Rightarrow -H^{-1}Z.$$

Proposition 3 provides a finite-sample entry and refinement guarantee under its stated conditions. It gives an unconditional first-order limit when the stated local, score, and iteration conditions hold and the failure probability tends to zero. We note that this entry inequality is only a sufficient condition. If this inequality fails, the optimization algorithm might still succeed, because the true attraction set can be strictly larger than our certified ball.

C Proofs for Local Target Theory

We prove uniform target contraction and the first-order expansion first because Proposition 3 relies on the same local target event.

C.1 Proof of Theorem 5

Proof. On the event in (5), for $\theta \in \mathbb{B}_R$,

$$m_{n_{\mathsf{T}}} I_p \preceq \nabla^2 L_{\mathsf{T},n_{\mathsf{T}}}(\theta) \preceq L_{n_{\mathsf{T}}} I_p.$$

Let $\bar{\theta}_{n_{\mathsf{T}}}$ be a minimizer of $L_{\mathsf{T},n_{\mathsf{T}}}$ over the compact set $\mathbb{B}_R$. For every $\theta \in \mathbb{B}_R$ satisfying $\|\theta - \theta^*_{\mathsf{T}}\| = R$,

$$\begin{aligned}\langle \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta), \theta - \theta^*_{\mathsf{T}}\rangle &= \langle S_{n_{\mathsf{T}}}, \theta - \theta^*_{\mathsf{T}}\rangle \\ &\quad + \int_0^1 (\theta - \theta^*_{\mathsf{T}})^\top \nabla^2 L_{\mathsf{T},n_{\mathsf{T}}}\{\theta^*_{\mathsf{T}} + s(\theta - \theta^*_{\mathsf{T}})\}(\theta - \theta^*_{\mathsf{T}})\, ds \\ &\geq -\tau_{n_{\mathsf{T}}} R + m_{n_{\mathsf{T}}} R^2.\end{aligned}$$

Because $m_{n_{\mathsf{T}}} \geq m/2$ and $\tau_{n_{\mathsf{T}}} \leq mR/4$, moving toward θ^*_{T} decreases $L_{\mathsf{T},n_{\mathsf{T}}}$ at each boundary point of $\mathbb{B}_R$. Thus $\bar{\theta}_{n_{\mathsf{T}}}$ is interior. Set $\widehat{\theta}_{\mathsf{T}} = \bar{\theta}_{n_{\mathsf{T}}}$; then

$$\nabla L_{\mathsf{T},n_{\mathsf{T}}}(\widehat{\theta}_{\mathsf{T}}) = 0.$$

Strong convexity on $\mathbb{B}_R$ gives uniqueness in $\mathbb{B}_R$ and

$$\begin{aligned}m_{n_{\mathsf{T}}} \left\|\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\|^2 &\leq \left\langle \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\widehat{\theta}_{\mathsf{T}}) - \nabla L_{\mathsf{T},n_{\mathsf{T}}}(\theta^*_{\mathsf{T}}), \widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\rangle \\ &= -\left\langle S_{n_{\mathsf{T}}}, \widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\rangle \leq \|S_{n_{\mathsf{T}}}\| \left\|\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\|.\end{aligned}$$

This gives

$$\left\|\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\| \le \frac{\|S_{n_\mathsf{T}}\|}{m_{n_\mathsf{T}}} \le \frac{2}{m}\|S_{n_\mathsf{T}}\|.$$

Fix $u \in \mathbb{B}_r$. For $\theta \in \mathbb{B}_R$, define

$$T_{n_\mathsf{T}}(\theta) = \theta - \alpha \nabla L_{\mathsf{T},n_\mathsf{T}}(\theta).$$

Because $\nabla L_{\mathsf{T},n_\mathsf{T}}(\widehat{\theta}_{\mathsf{T}}) = 0$,

$$T_{n_\mathsf{T}}(\theta) - T_{n_\mathsf{T}}(\widehat{\theta}_{\mathsf{T}}) = \left[I_p - \alpha \int_0^1 \nabla^2 L_{\mathsf{T},n_\mathsf{T}}\{\widehat{\theta}_{\mathsf{T}} + s(\theta - \widehat{\theta}_{\mathsf{T}})\}\, ds\right](\theta - \widehat{\theta}_{\mathsf{T}}),$$

because the segment from θ to $\widehat{\theta}_{\mathsf{T}}$ lies in $\mathbb{B}_R$. The averaged Hessian has eigenvalues in $[m_{n_\mathsf{T}}, L_{n_\mathsf{T}}]$, giving contraction factor κ_{n_T}:

$$\left\|T_{n_\mathsf{T}}(\theta) - T_{n_\mathsf{T}}(\widehat{\theta}_{\mathsf{T}})\right\| \le \kappa_{n_\mathsf{T}} \left\|\theta - \widehat{\theta}_{\mathsf{T}}\right\|.$$

Inductively, the iterates remain in $\mathbb{B}_R$ and approach $\widehat{\theta}_{\mathsf{T}}$. Put

$$b_{n_\mathsf{T}} = \left\|\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\right\|.$$

Then $b_{n_\mathsf{T}} \le 2\tau_{n_\mathsf{T}}/m$, and $u \in \mathbb{B}_r$ gives

$$\left\|u - \widehat{\theta}_{\mathsf{T}}\right\| \le r + b_{n_\mathsf{T}}.$$

Assume $\theta^t(u) \in \mathbb{B}_R$. The contraction bound gives

$$\left\|\theta^{t+1}(u) - \widehat{\theta}_{\mathsf{T}}\right\| \le \kappa_{n_\mathsf{T}} \left\|\theta^t(u) - \widehat{\theta}_{\mathsf{T}}\right\|.$$

This implies

$$\left\|\theta^{t+1}(u) - \theta^*_{\mathsf{T}}\right\| \le \left\|\theta^{t+1}(u) - \widehat{\theta}_{\mathsf{T}}\right\| + b_{n_\mathsf{T}} \le r + 2b_{n_\mathsf{T}} \le r + 4\tau_{n_\mathsf{T}}/m \le R.$$

Starting from $\theta^0(u) = u \in \mathbb{B}_r \subset \mathbb{B}_R$, induction gives, for $t \ge 0$,

$$\theta^t(u) \in \mathbb{B}_R, \qquad \left\|\theta^t(u) - \widehat{\theta}_{\mathsf{T}}\right\| \le \kappa_{n_\mathsf{T}}^t \left\|u - \widehat{\theta}_{\mathsf{T}}\right\|.$$

Because $\left\|u - \widehat{\theta}_{\mathsf{T}}\right\| \le r + 2\|S_{n_\mathsf{T}}\|/m$, the statistical error bound follows.

For the expansion, put

$$d_{n_\mathsf{T}} = \widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}.$$

Because $\nabla L_{\mathsf{T},n_\mathsf{T}}(\widehat{\theta}_{\mathsf{T}}) = 0$,

$$0 = S_{n_\mathsf{T}} + \int_0^1 \nabla^2 L_{\mathsf{T},n_\mathsf{T}}(\theta^*_{\mathsf{T}} + s d_{n_\mathsf{T}}) d_{n_\mathsf{T}}\, ds.$$

Subtracting $S_{n_\mathsf{T}} + H d_{n_\mathsf{T}}$ gives

$$H d_{n_\mathsf{T}} + S_{n_\mathsf{T}} = -\int_0^1 \{\nabla^2 L_{\mathsf{T},n_\mathsf{T}}(\theta^*_{\mathsf{T}} + s d_{n_\mathsf{T}}) - H\} d_{n_\mathsf{T}}\, ds.$$

The empirical Hessian bound and population Lipschitz bound give

$$\begin{aligned}\left\|d_{n_\mathsf{T}} + H^{-1} S_{n_\mathsf{T}}\right\| &\le \left\|H^{-1}\right\| \int_0^1 \left\|\nabla^2 L_{\mathsf{T},n_\mathsf{T}}(\theta^*_{\mathsf{T}} + s d_{n_\mathsf{T}}) - H\right\|_{\mathrm{op}} ds\, \|d_{n_\mathsf{T}}\| \\ &\le \frac{1}{m}\left(\Delta_{n_\mathsf{T}} + \frac{M}{2}\|d_{n_\mathsf{T}}\|\right)\|d_{n_\mathsf{T}}\|.\end{aligned}$$

Because $\|d_{n_\mathsf{T}}\| \le 2\|S_{n_\mathsf{T}}\|/m$, there is $C < \infty$, depending only on m and M, such that

$$\left\|d_{n_\mathsf{T}} + H^{-1} S_{n_\mathsf{T}}\right\| \le C\{\Delta_{n_\mathsf{T}}\|S_{n_\mathsf{T}}\| + \|S_{n_\mathsf{T}}\|^2\}.$$

Finally,

$$\begin{aligned}\left\|\theta^t(u) - \theta^*_{\mathsf{T}} + H^{-1} S_{n_\mathsf{T}}\right\| &\le \left\|\theta^t(u) - \widehat{\theta}_{\mathsf{T}}\right\| + \left\|d_{n_\mathsf{T}} + H^{-1} S_{n_\mathsf{T}}\right\| \\ &\le \kappa_{n_\mathsf{T}}^t \left(r + \frac{2}{m}\|S_{n_\mathsf{T}}\|\right) + C\{\Delta_{n_\mathsf{T}}\|S_{n_\mathsf{T}}\| + \|S_{n_\mathsf{T}}\|^2\}.\end{aligned}$$

Enlarging C completes the proof. □

C.2 Proof of Corollary 4

For a positive deterministic sequence a_n, we write $W_n = O_p(a_n)$, or $O_{\mathbb{P}}(a_n)$, if $\|W_n\|/a_n$ is bounded in probability. We write $W_n = o_p(a_n)$, or $o_{\mathbb{P}}(a_n)$, if this ratio tends to zero in probability.

Proof. Let E_{n_T} be the event in (5), and write $n = n_\mathsf{T}$. Existence and contraction in Theorem 5 use no Hessian Lipschitz bound. On this event, put $d_n = \widehat{\theta}_\mathsf{T} - \theta^*_\mathsf{T}$ and

$$A_n = \int_0^1 \nabla^2 L_{\mathsf{T},n}(\theta^*_\mathsf{T} + v d_n)\, dv.$$

Then $A_n d_n = -S_n$, $A_n \succeq (m/2) I_p$, and $\|d_n\| \le 2\|S_n\|/m = O_p(n^{-1/2})$. Continuity of the population Hessian at θ^*_T gives

$$\|A_n - H\|_{\mathrm{op}} \le \Delta_n + \sup_{\|\theta - \theta^*_\mathsf{T}\| \le 2\|S_n\|/m} \left\|\nabla^2 L_\mathsf{T}(\theta) - H\right\|_{\mathrm{op}} = o_p(1).$$

The inverse identity $H^{-1} - A_n^{-1} = H^{-1}(A_n - H)A_n^{-1}$ implies

$$\sqrt{n}\, \left\|d_n + H^{-1} S_n\right\| \le \frac{2}{m^2} \|A_n - H\|_{\mathrm{op}} \left\|\sqrt{n} S_n\right\| = o_p(1).$$

The contraction bound also gives

$$\sup_{u \in \mathbb{B}_r} \sqrt{n} \left\|\theta^{t_n}(u) - \widehat{\theta}_\mathsf{T}\right\| \le \sqrt{n}\, \kappa_n^{t_n} \left(r + \frac{2}{m}\|S_n\|\right) = o_p(1).$$

The union bound gives

$$\mathbb{P}\{E_{n_\mathsf{T}}^c \cup (u_{n_\mathsf{T}} \notin \mathbb{B}_r)\} \le \eta_{n_\mathsf{T}} + \mathbb{P}\{u_{n_\mathsf{T}} \notin \mathbb{B}_r\} \to 0,$$

and Slutsky's theorem gives both limits. The start need not be independent of the target sample. This completes the proof. □

C.3 Proof of Proposition 3

Proof. By (6),

$$\mathbb{P}\{\left\|\widehat{\theta}_\mathsf{S} - \theta^*_\mathsf{S}\right\| \le a_\mathsf{S}(s_\mathsf{S})\} \ge 1 - e^{-s_\mathsf{S}}.$$

On this event,

$$\left\|\widehat{\theta}_\mathsf{S} - \theta^*_\mathsf{T}\right\| \le \left\|\widehat{\theta}_\mathsf{S} - \theta^*_\mathsf{S}\right\| + \|\theta^*_\mathsf{S} - \theta^*_\mathsf{T}\| \le a_\mathsf{S}(s_\mathsf{S}) + \delta \le r.$$

Theorem 5 applies uniformly on $\mathbb{B}_r$ with probability at least $1 - \eta_{n_\mathsf{T}}$; a union bound gives its conclusion for $\widehat{\theta}_\mathsf{S}$. For the limit, the same argument with s_{n_T} gives $\mathbb{P}\{\widehat{\theta}_\mathsf{S} \notin \mathbb{B}_r\} \le e^{-s_{n_\mathsf{T}}} \to 0$. The proof of Corollary 4 then gives

$$n_\mathsf{T}^{1/2}\{\theta^{t_{n_\mathsf{T}}}(\widehat{\theta}_\mathsf{S}) - \theta^*_\mathsf{T} + H^{-1} S_{n_\mathsf{T}}\} = o_p(1).$$

This completes the proof. □

D Basin Probability and Certificates

Entry into the target ball depends on the source error, the source-target distance, and the ball radius. A source tail bound gives bounds on these entry probabilities. Alternatively, comparing the output with a target reference provides a one-sided validation guarantee after fitting under the stated separation and concentration conditions.

We write $\mathbb{P}_\mathsf{S}$ for probability over the source sample. To calculate finite-sample basin entry, we assume a nonincreasing function $\tau_\mathsf{S} : (0, \infty) \to [0, 1]$ bounds the source tail probability:

$$\mathbb{P}_\mathsf{S}\left\{\left\|\widehat{\theta}_\mathsf{S} - \theta^*_\mathsf{S}\right\| \ge t\right\} \le \tau_\mathsf{S}(t), \qquad t > 0. \tag{7}$$

This bound restricts only the source estimator around its own population parameter. It makes no direct assumption about target-basin membership. We define the sufficient basin event as

$$S_r = \{\widehat{\theta}_\mathsf{S} \in \mathbb{B}_r\}.$$

Theorem 6 gives a lower bound when $\delta < r$ and an upper bound when $\delta > r$.

Theorem 6. *Assume* (7). *If $\delta < r$, then*

$$\mathbb{P}(S_r) \geq 1 - \tau_{\mathsf{S}}(r - \delta).$$

If $\delta > r$, then

$$\mathbb{P}(S_r) \leq \tau_{\mathsf{S}}(\delta - r).$$

If, in addition, for constants $C_{\mathsf{S}}, c_{\mathsf{S}}, \sigma_{\mathsf{S}} > 0$,

$$\tau_{\mathsf{S}}(t) \leq C_{\mathsf{S}} \exp\{-c_{\mathsf{S}} n_{\mathsf{S}} t^2/\sigma_{\mathsf{S}}^2\}, \qquad t > 0,$$

then, for $\delta < r$,

$$\mathbb{P}(S_r) \geq 1 - C_{\mathsf{S}} \exp\{-c_{\mathsf{S}} n_{\mathsf{S}} (r - \delta)^2/\sigma_{\mathsf{S}}^2\},$$

and, for $\delta > r$,

$$\mathbb{P}(S_r) \leq C_{\mathsf{S}} \exp\{-c_{\mathsf{S}} n_{\mathsf{S}} (\delta - r)^2/\sigma_{\mathsf{S}}^2\}.$$

Theorem 6 bounds the probability of a purely sufficient basin event. This formal tail calculation replaces informal threshold heuristics. We emphasize that this result does not classify every point outside $\mathbb{B}_r$ as a failure, because the true attraction set can be strictly larger.

To compare this bound with a Gaussian model, we use the local metric $\|u\|_H = (u^\top H u)^{1/2}$ and define the separation $\delta_H = \|\theta_{\mathsf{S}}^* - \theta_{\mathsf{T}}^*\|_H$. We let $G_{p,\lambda}$ denote the distribution function of a noncentral chi-square variable with p degrees of freedom and noncentrality parameter λ. Corollary 5 evaluates the exact entry probability in this metric.

Corollary 5. *Assume, for some $\sigma_{\mathsf{S}} > 0$, that*

$$H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{S}}^*) \sim \mathsf{N}(0, \sigma_{\mathsf{S}}^2 n_{\mathsf{S}}^{-1} I_p)$$

holds. Then

$$\mathbb{P}\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{T}}^*\right\|_H < r\} = G_{p, n_{\mathsf{S}} \delta_H^2/\sigma_{\mathsf{S}}^2}\left(\frac{n_{\mathsf{S}} r^2}{\sigma_{\mathsf{S}}^2}\right).$$

Corollary 5 provides an exact benchmark for a local Gaussian model. For more general models, we recommend using Theorem 6 alongside a formally verified source tail bound.

We can assess basin entry before executing the transfer using bounds on the unknown radius, shift, and source estimation error. For any $\alpha_{\mathsf{S}} \in (0, 1)$, we define a source error bound $q_{\mathsf{S}}(\alpha_{\mathsf{S}})$ to satisfy

$$\mathbb{P}_{\mathsf{S}}\left\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{S}}^*\right\| \leq q_{\mathsf{S}}(\alpha_{\mathsf{S}})\right\} \geq 1 - \alpha_{\mathsf{S}}.$$

Because the exact parameters remain unknown, our pre-transfer certificate uses conservative bounds for r and δ. We define data-dependent quantities L_r and U_δ that jointly satisfy

$$\mathbb{P}\{L_r \leq r,\ \delta \leq U_\delta\} \geq 1 - \alpha_0.$$

For an event A, the indicator $\mathbf{1}\{A\}$, also written 1_A, equals one on A and zero on its complement. For a chosen $\alpha_{\mathsf{S}} \in (0, 1)$, we define the certificate indicator

$$C_{\mathrm{pre}} = \mathbf{1}\{U_\delta + q_{\mathsf{S}}(\alpha_{\mathsf{S}}) < L_r\}.$$

Proposition 4 bounds the probability of false certification.

Proposition 4.

$$\mathbb{P}\{C_{\mathrm{pre}} = 1,\ \widehat{\theta}_{\mathsf{S}} \notin \mathbb{B}_r\} \leq \alpha_0 + \alpha_{\mathsf{S}}.$$

Proposition 4 provides an intentionally one-sided guarantee. If $C_{\mathrm{pre}} = 0$, the available data fail to certify the sufficient basin event. However, this failure does not guarantee that the source initializer will miss the target basin.

Alternatively, we can use validation data to check the output after the optimization finishes. Simple convergence diagnostics need not distinguish the target local estimator from a stable local minimizer located far away. To resolve this, our rejection rule compares the candidate output against a computable target reference using independent validation data. We define a bounded set $\Theta_0 \subset \Theta$ containing every candidate output, and we define the set of inaccurate candidates

$$\mathcal{A}_r = \{\theta \in \Theta_0 : \|\theta - \theta_{\mathsf{T}}^*\| \geq r\}.$$

This set depends on the unknown target parameter and is used only in the analysis. We compute the candidate $\bar{\theta} \in \Theta_0$ and the reference $\widehat{\theta}_{\mathrm{ref}} \in \Theta_0$ strictly without the validation sample. We evaluate their expected validation risk as $R_v(\theta) = \mathbb{E}_v V_n(\theta)$, taking the expectation only over the validation sample conditionally on the fitted outputs. Proposition 5 gives sufficient separation and concentration conditions for controlling false acceptance.

Proposition 5. *Suppose that, with probability at least* $1 - \alpha_0$ *over the fitting data,*

$$\inf_{\theta \in \mathcal{A}_r} R_v(\theta) > R_v(\widehat{\theta}_{\mathrm{ref}}) + 4\gamma_n,$$

and that, conditionally on the fitting data, with probability at least $1 - \alpha_v$*,*

$$|V_n(\widehat{\theta}_{\mathrm{ref}}) - R_v(\widehat{\theta}_{\mathrm{ref}})| \le \gamma_n, \qquad |V_n(\bar{\theta}) - R_v(\bar{\theta})| \le \gamma_n.$$

Accept the candidate $\bar{\theta} \in \Theta_0$ *only if*

$$V_n(\bar{\theta}) \le V_n(\widehat{\theta}_{\mathrm{ref}}) + 2\gamma_n.$$

Then

$$\mathbb{P}\{\bar{\theta} \text{ is accepted}, \ \|\bar{\theta} - \theta^*_{\mathsf{T}}\| \ge r\} \le \alpha_0 + \alpha_v.$$

Proposition 5 deliberately avoids assuming that the target local estimator is already available for comparison. Instead, it relies on a computable reference to provide a strictly one-sided guarantee. Under the stated conditions, the probability of accepting a candidate at distance at least r from the target parameter is at most $\alpha_0 + \alpha_v$. Acceptance does not prove that the starting point belonged to the certified ball.

The validation concentration conditions in Proposition 5 restrict only the completed candidate and reference outputs. Let $Z^v_{\mathsf{T},1}, \dots, Z^v_{\mathsf{T},n_v}$ be n_v independent target validation observations, independent of the fitting data. Let Z^v_{T} denote a generic validation observation. We consider the sample and population validation risks

$$V_n(\theta) = \frac{1}{n_v} \sum_{i=1}^{n_v} \ell_{\mathsf{T}}(\theta; Z^v_{\mathsf{T},i}), \qquad R_v(\theta) = \mathbb{E}_v\{\ell_{\mathsf{T}}(\theta; Z^v_{\mathsf{T}})\}.$$

For a real random variable W and $q \in \{1, 2\}$, we let $\|W\|_{\psi_q}$ denote the infimum of all $c > 0$ satisfying $\mathbb{E}\exp(|W|^q/c^q) \le 2$. We assume the candidates $\theta_1, \dots, \theta_K \in \Theta_0$ are computed strictly without the validation sample. Proposition 6 directly derives an explicit tolerance from a population tail bound.

Proposition 6. *Assume that, for a constant* $K_v < \infty$*,*

$$\sup_{\theta \in \Theta_0} \|\ell_{\mathsf{T}}(\theta; Z^v_{\mathsf{T}}) - R_v(\theta)\|_{\psi_1} \le K_v$$

holds. For every $s \ge 1$*, conditionally on the fitted outputs, with probability at least* $1 - e^{-s}$*,*

$$\max_{1 \le j \le K} |V_n(\theta_j) - R_v(\theta_j)| \le C K_v \left\{ \sqrt{\frac{s + \log(2K)}{n_v}} + \frac{s + \log(2K)}{n_v} \right\},$$

where $C < \infty$ *is a numerical constant.*

We apply Proposition 6 with $K = 2$ specifically to the candidate and reference in Proposition 5. Under the stated population tail bound, empirical concentration is needed only at these two fitted outputs. The population separation condition must still be verified.

E Proofs for Basin Probability and Certificates

Source deviation probabilities directly determine the basin entry bound. Under the stated separation condition, candidatewise validation concentration controls the probability of accepting a candidate far from the target parameter.

E.1 Proof of Theorem 6

Proof. If $\delta < r$, then

$$\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| < r - \delta\} \subset \{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{T}}\right\| < r\}.$$

The triangle inequality gives

$$\mathbb{P}(S_r) \geq 1 - \mathbb{P}\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| \geq r - \delta\} \geq 1 - \tau_{\mathsf{S}}(r - \delta).$$

If $\delta > r$, then

$$\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{T}}\right\| \leq r\} \subset \{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| \geq \delta - r\}.$$

This implies

$$\mathbb{P}(S_r) \leq \mathbb{P}\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| \geq \delta - r\} \leq \tau_{\mathsf{S}}(\delta - r).$$

The exponential bounds follow by substituting the exponential form of τ_{S}. □

E.2 Proof of Corollary 5

Proof. Let

$$W = \frac{n_{\mathsf{S}}^{1/2}}{\sigma_{\mathsf{S}}} H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{T}}).$$

Then

$$W \sim \mathsf{N}\left(\frac{n_{\mathsf{S}}^{1/2}}{\sigma_{\mathsf{S}}} H^{1/2}(\theta^*_{\mathsf{S}} - \theta^*_{\mathsf{T}}), I_p\right).$$

The event $\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{T}}\right\|_H < r$ is

$$\|W\|^2 < \frac{n_{\mathsf{S}} r^2}{\sigma_{\mathsf{S}}^2}.$$

The squared norm of W has a noncentral chi-square distribution with noncentrality $n_{\mathsf{S}}\delta_H^2/\sigma_{\mathsf{S}}^2$, which proves the formula. □

E.3 Proof of Proposition 4

Proof. On the event

$$\{L_r \leq r,\ \delta \leq U_\delta,\ \left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| \leq q_{\mathsf{S}}(\alpha_{\mathsf{S}})\},$$

$C_{\mathrm{pre}} = 1$ implies

$$\left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{T}}\right\| \leq \left\|\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}\right\| + \delta \leq q_{\mathsf{S}}(\alpha_{\mathsf{S}}) + U_\delta < L_r \leq r.$$

It follows that $\widehat{\theta}_{\mathsf{S}} \in \mathbb{B}_r$. The complement of the event above has probability at most $\alpha_0 + \alpha_{\mathsf{S}}$, which gives the result. □

E.4 Proof of Proposition 5

Proof. On the fitting-data separation event and the conditional validation concentration event in the proposition, if $\left\|\bar{\theta} - \theta^*_{\mathsf{T}}\right\| \geq r$, then $\bar{\theta} \in \mathcal{A}_r$, and

$$V_n(\bar{\theta}) \geq R_v(\bar{\theta}) - \gamma_n > R_v(\widehat{\theta}_{\mathrm{ref}}) + 3\gamma_n \geq V_n(\widehat{\theta}_{\mathrm{ref}}) + 2\gamma_n.$$

This contradicts the acceptance inequality. The probability that either the fitting-data separation or conditional validation event fails is at most $\alpha_0 + \alpha_v$, which proves the claim. □

E.5 Proof of Proposition 6

Proof. Condition on the fitted outputs. For a fixed j, Bernstein's inequality for centered sub-exponential averages (Vershynin, 2018, Theorem 2.8.1) gives, for every $x \geq 1$,

$$\mathbb{P}\left[|V_n(\theta_j) - R_v(\theta_j)| > CK_v\left\{\sqrt{\frac{x}{n_v}} + \frac{x}{n_v}\right\} \,\middle|\, \theta_1, \ldots, \theta_K\right] \leq 2e^{-x}.$$

Set $x = s + \log(2K)$. A union bound over $j = 1, \ldots, K$ gives conditional failure probability at most $2Ke^{-x} = e^{-s}$, which proves the result. □

F Data Transfer and Initialization Transfer

Statistical risk reduction and basin access rely on different aspects of the source sample. Once an optimization run enters the target basin, pooling source data can reduce the local quadratic risk, but population differences can introduce a source-target bias. In contrast, a source initializer can affect basin entry and finite-iteration error. Under the conditions of Corollary 4, reliable entry and enough updates give the refined estimator the same first-order distribution as the target local estimator.

To formally compare these effects, we define H as the positive definite target curvature. We define the local metric $\|\cdot\|_H$ as in Corollary 5, and we measure the local separation as $\delta_H = \|\theta^*_{\mathsf{S}} - \theta^*_{\mathsf{T}}\|_H$. For any candidate estimator $\widetilde{\theta}$ of the target parameter θ^*_{T}, we define its local target risk as

$$\mathcal{R}(\widetilde{\theta}) = \mathbb{E}\left\|\widetilde{\theta} - \theta^*_{\mathsf{T}}\right\|_H^2.$$

We analyze these trade-offs using a local second-moment comparison model. We write Var for the covariance matrix of a random vector. For specified constants $\sigma_{\mathsf{T}}, \sigma_{\mathsf{S}} > 0$ and sample sizes $n_{\mathsf{T}}, n_{\mathsf{S}}$, we assume

$$\mathbb{E}\{H^{1/2}(\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}})\} = 0, \qquad \mathrm{Var}\{H^{1/2}(\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}})\} = \frac{\sigma^2_{\mathsf{T}}}{n_{\mathsf{T}}} I_p,$$

$$\mathbb{E}\{H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}})\} = 0, \qquad \mathrm{Var}\{H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}})\} = \frac{\sigma^2_{\mathsf{S}}}{n_{\mathsf{S}}} I_p.$$

We further assume the two estimation errors maintain zero cross-covariance:

$$\mathbb{E}\left[H^{1/2}(\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}})\{H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}})\}^\top\right] = 0.$$

This local model intentionally does not claim that source data solve the global nonconvex optimization problem. Instead, it quantifies the risk of combining the two local estimators. To simplify notation, we define the baseline variances

$$v_{\mathsf{T}} = \frac{p\sigma^2_{\mathsf{T}}}{n_{\mathsf{T}}}, \qquad v_{\mathsf{S}} = \frac{p\sigma^2_{\mathsf{S}}}{n_{\mathsf{S}}}.$$

We next construct an oracle-gated algorithmic benchmark. We fix a basin radius $r > 0$ and assume a separately fitted source-data procedure successfully reaches its local estimator with probability $q \in [0, 1]$. We also assign an oracle penalty risk $R_0 > v_{\mathsf{T}}$ to any failed run. In this benchmark, a source initializer inside the H-ball of radius r returns $\widehat{\theta}_{\mathsf{T}}$. For successful runs, the oracle assigns the target local risk to initialization and the optimized local pooling risk to source-data transfer. While this construction serves strictly as a theoretical comparison benchmark, Proposition 5 provides a feasible one-sided rejection rule under its separation and concentration conditions.

For a fixed interpolation weight $\lambda \in [0, 1]$, we define the pooled local source-data estimator as

$$\widehat{\theta}_\lambda = (1 - \lambda)\widehat{\theta}_{\mathsf{T}} + \lambda\widehat{\theta}_{\mathsf{S}}.$$

Proposition 7 calculates the local risk for any fixed pooling rule, including the oracle-optimized weight.

Proposition 7. *Under the local second-moment comparison model,*

$$\mathcal{R}(\widehat{\theta}_\lambda) = (1-\lambda)^2 v_{\mathsf{T}} + \lambda^2 (v_{\mathsf{S}} + \delta_H^2).$$

The minimizing weight and minimum local source-data risk are

$$\lambda^* = \frac{v_{\mathsf{T}}}{v_{\mathsf{T}} + v_{\mathsf{S}} + \delta_H^2}, \qquad D_* = \frac{v_{\mathsf{T}}(v_{\mathsf{S}} + \delta_H^2)}{v_{\mathsf{T}} + v_{\mathsf{S}} + \delta_H^2}.$$

For a fixed nonzero weight λ, source-data transfer has larger local risk than the target local estimator exactly when

$$v_{\mathsf{S}} + \delta_H^2 > \left(\frac{2}{\lambda} - 1\right) v_{\mathsf{T}}.$$

This final inequality characterizes negative data transfer at finite sample sizes for any fixed nonzero pooling weight. However, if we optimize the weight using the true population shift, we guarantee $D_* \le v_{\mathsf{T}}$. In this oracle setting, the source data cannot increase the local quadratic risk because the admissible weights include zero.

While Proposition 7 evaluates only simple convex combinations, one might wonder if a more complex use of the two estimators could perform significantly better. Proposition 8 bounds the possible improvement in the Gaussian version of our benchmark. It proves that no measurable function of the two estimators can improve upon the worst-case pooling risk by more than a universal constant factor. This bound holds uniformly over all shifts bounded by a known level. In this Gaussian benchmark, we assume $H, \sigma_{\mathsf{T}}, \sigma_{\mathsf{S}}$ remain fixed and known. For a known upper bound $D \ge 0$, we define the minimax risk and the worst-case optimized pooling risk as

$$\mathcal{M}(D) = \inf_{\widetilde{\theta}} \sup_{\delta_H \le D} \mathbb{E} \left\| \widetilde{\theta} - \theta_{\mathsf{T}}^* \right\|_H^2, \qquad D_*(D) = \frac{v_{\mathsf{T}}(v_{\mathsf{S}} + D^2)}{v_{\mathsf{T}} + v_{\mathsf{S}} + D^2}.$$

We take the infimum over all measurable estimators $\widetilde{\theta} = \widetilde{\theta}(\widehat{\theta}_{\mathsf{T}}, \widehat{\theta}_{\mathsf{S}})$, and we take the supremum over all parameter pairs $(\theta_{\mathsf{T}}^*, \theta_{\mathsf{S}}^*) \in \mathbb{R}^p \times \mathbb{R}^p$ satisfying $\delta_H = \|\theta_{\mathsf{S}}^* - \theta_{\mathsf{T}}^*\|_H \le D$.

Proposition 8. *Assume that $H^{1/2}(\widehat{\theta}_{\mathsf{T}} - \theta_{\mathsf{T}}^*)$ and $H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{S}}^*)$ are independent with distributions $\mathsf{N}(0, n_{\mathsf{T}}^{-1}\sigma_{\mathsf{T}}^2 I_p)$ and $\mathsf{N}(0, n_{\mathsf{S}}^{-1}\sigma_{\mathsf{S}}^2 I_p)$. Then*

$$\tfrac{1}{8}\, D_*(D) \le \mathcal{M}(D) \le D_*(D).$$

Proposition 8 proves that pooling with the weight $\lambda = v_{\mathsf{T}}/(v_{\mathsf{T}} + v_{\mathsf{S}} + D^2)$ achieves minimax rate optimality over the bounded-shift class $\{\delta_H \le D\}$ in local quadratic risk, up to a universal factor of eight. This result compares worst-case risks over the shift-bounded class and does not assert pointwise optimality at a fixed δ_H. It rigorously justifies using $D_*(D)$ as a rate-optimal source-data benchmark whenever only the bound $\delta_H \le D$ is known.

To compare transfer modes, we fix all other features of the source-estimator distribution and define the probability of entry into the H-ball as

$$p(n_{\mathsf{S}}, \delta_H) = \mathbb{P}\{\left\| \widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{T}}^* \right\|_H < r\}.$$

We then define the oracle-gated benchmark risks for initialization transfer, data transfer, and their combination as

$$\mathcal{R}_{\text{init}} = p(n_{\mathsf{S}}, \delta_H) v_{\mathsf{T}} + \{1 - p(n_{\mathsf{S}}, \delta_H)\} R_0,$$
$$\mathcal{R}_{\text{data}} = q D_* + (1-q) R_0,$$
$$\mathcal{R}_{\text{both}} = p(n_{\mathsf{S}}, \delta_H) D_* + \{1 - p(n_{\mathsf{S}}, \delta_H)\} R_0.$$

The final expression $\mathcal{R}_{\text{both}}$ represents a purely algebraic benchmark that artificially combines target-basin entry with the oracle local source-data risk. We caution that it is not a proven convergence guarantee for any actual pooled criterion.

Proposition 9 derives the exact theoretical crossovers among these benchmark risks.

Proposition 9. *For the benchmark risks above,*

$$\mathcal{R}_{\text{init}} < \mathcal{R}_{\text{data}} \quad \Longleftrightarrow \quad p(n_{\mathsf{S}}, \delta_H)(R_0 - v_{\mathsf{T}}) > q(R_0 - D_*).$$

In addition,

$$\mathcal{R}_{\text{both}} \leq \mathcal{R}_{\text{init}},$$

with strict inequality when $p(n_{\mathsf{S}}, \delta_H) > 0$ and $D_* < v_{\mathsf{T}}$, and

$$\mathcal{R}_{\text{both}} \leq \mathcal{R}_{\text{data}}$$

whenever $p(n_{\mathsf{S}}, \delta_H) \geq q$.

Proposition 9 specifies the exact crossover conditions within our benchmark model. It formally demonstrates why universal strict dominance of one transfer mode over the other cannot hold. The combined benchmark improves on initialization alone when $p(n_{\mathsf{S}}, \delta_H) > 0$ and $D_* < v_{\mathsf{T}}$. It is also at most the data-transfer benchmark when $p(n_{\mathsf{S}}, \delta_H) \geq q$. To implement a feasible combined procedure, one would require a completely separate basin analysis for the new pooled criterion.

Crucially, the source-target shift enters the two transfer modes through fundamentally different mechanisms. Under the local second-moment comparison model with a fixed n_{S}, the optimized source-data value satisfies

$$v_{\mathsf{T}} - D_* = \frac{v_{\mathsf{T}}^2}{v_{\mathsf{T}} + v_{\mathsf{S}} + \delta_H^2}.$$

In contrast, for fixed v_{T} and R_0, the initialization benchmark risk is determined by the entry probability

$$p(n_{\mathsf{S}}, \delta_H) = \mathbb{P}\{\left\|\widehat{\theta}_{\mathsf{S}} - \theta_{\mathsf{T}}^*\right\|_H < r\}.$$

The optimized local pooling gain decreases continuously with δ_H^2. The change in entry probability depends on the source-error distribution and its position relative to the H-ball.

This benchmark comparison does not imply that an initialization failure incurs only a small cost. A failed start can easily converge to a stable empirical local minimum located far from the true target parameter. To control acceptance of such failures, the analyst can use Proposition 5 under its separation and concentration conditions. Bounding the risk of the resulting procedure requires further analysis of the backup estimator and of accepted outputs when validation fails. We do not establish such a risk bound.

Because the exact crossover condition in Proposition 9 depends on the unknown shift and entry probabilities, an analyst cannot directly evaluate it from the fitted data. Our final result gives a validation rule for choosing between the completed fits. Under the stated approximation condition, this rule controls expected risk relative to the better candidate on each fitted sample. The analyst does not need to know the crossover condition in advance.

We define B_0 as the supremum of $\|\theta - \theta_{\mathsf{T}}^*\|_H^2$ over the bounded candidate set Θ_0. Given two candidate outputs $\widehat{\theta}_a, \widehat{\theta}_b \in \Theta_0$, we define the selected estimator $\widehat{\theta}_{\text{sel}}$ as the output that achieves the smaller validation score V_n.

Proposition 10. *Let $\widehat{\theta}_a, \widehat{\theta}_b \in \Theta_0$ be the outputs of two transfer procedures and let V_n be a validation criterion. Assume that, for constants $c_0 \in \mathbb{R}$, $c_1 > 0$, a tolerance γ_n, and a level α_v,*

$$\mathbb{P}\left\{\max_{c \in \{a,b\}} \left| V_n(\widehat{\theta}_c) - c_0 - c_1 \left\|\widehat{\theta}_c - \theta_{\mathsf{T}}^*\right\|_H^2 \right| \leq \gamma_n \right\} \geq 1 - \alpha_v.$$

Then

$$\mathbb{E}\left\|\widehat{\theta}_{\text{sel}} - \theta_{\mathsf{T}}^*\right\|_H^2 \leq \mathbb{E}\min\left\{\left\|\widehat{\theta}_a - \theta_{\mathsf{T}}^*\right\|_H^2, \left\|\widehat{\theta}_b - \theta_{\mathsf{T}}^*\right\|_H^2\right\} + \frac{2\gamma_n}{c_1} + \alpha_v B_0.$$

Proposition 10 bounds the selected estimator's expected squared H-norm error by the expected error of the better candidate on each fitted sample, plus validation terms. The additional risk is bounded by $2\gamma_n/c_1 + \alpha_v B_0$. If the expected validation risk satisfies $R_v(\theta) = c_0 + c_1 \|\theta - \theta_{\mathsf{T}}^*\|_H^2$ everywhere on Θ_0, then Proposition 6 formally verifies this approximation event using its population tail bound and independent validation data. This risk identity holds exactly for squared prediction loss in a linear regression model with finite error variance, conditionally centered errors, and predictor second moment $c_1 H$. It remains distinct from the quartic risk found in phase retrieval.

G Proofs for Risk Comparisons

Our Gaussian comparison formally separates the local variance reduction that pooling can provide from the change in basin-entry probability due to initialization.

G.1 Proof of Proposition 7

Proof. Write

$$U_{\mathsf{T}} = H^{1/2}(\widehat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}), \qquad U_{\mathsf{S}} = H^{1/2}(\widehat{\theta}_{\mathsf{S}} - \theta^*_{\mathsf{S}}), \qquad d = H^{1/2}(\theta^*_{\mathsf{S}} - \theta^*_{\mathsf{T}}).$$

Then

$$H^{1/2}(\widehat{\theta}_\lambda - \theta^*_{\mathsf{T}}) = (1-\lambda)U_{\mathsf{T}} + \lambda U_{\mathsf{S}} + \lambda d.$$

Centering and zero cross-covariance remove the cross terms, including those with d, giving

$$\mathcal{R}(\widehat{\theta}_\lambda) = (1-\lambda)^2 \mathbb{E}\left\|U_{\mathsf{T}}\right\|^2 + \lambda^2 \mathbb{E}\left\|U_{\mathsf{S}}\right\|^2 + \lambda^2 \left\|d\right\|^2 = (1-\lambda)^2 v_{\mathsf{T}} + \lambda^2(v_{\mathsf{S}} + \delta_H^2).$$

This quadratic is minimized at

$$\lambda^* = \frac{v_{\mathsf{T}}}{v_{\mathsf{T}} + v_{\mathsf{S}} + \delta_H^2},$$

and substitution gives D_*. For fixed $\lambda > 0$,

$$\mathcal{R}(\widehat{\theta}_\lambda) > v_{\mathsf{T}}$$

is equivalent to

$$(1-\lambda)^2 v_{\mathsf{T}} + \lambda^2(v_{\mathsf{S}} + \delta_H^2) > v_{\mathsf{T}},$$

which simplifies to the stated inequality. □

G.2 Proof of Proposition 8

Proof. Put $x = H^{1/2}\widehat{\theta}_{\mathsf{T}}$, $y = H^{1/2}\widehat{\theta}_{\mathsf{S}}$, $\mu = H^{1/2}\theta^*_{\mathsf{T}}$, and $\nu = H^{1/2}\theta^*_{\mathsf{S}}$. Then $x \sim \mathsf{N}(\mu, s_{\mathsf{T}}^2 I_p)$ and $y \sim \mathsf{N}(\nu, s_{\mathsf{S}}^2 I_p)$ are independent with $s_{\mathsf{T}}^2 = \sigma_{\mathsf{T}}^2/n_{\mathsf{T}}$ and $s_{\mathsf{S}}^2 = \sigma_{\mathsf{S}}^2/n_{\mathsf{S}}$, the constraint is $\|\nu - \mu\| \le D$, the loss is $\|\widehat{\mu} - \mu\|^2$, and $v_{\mathsf{T}} = p s_{\mathsf{T}}^2$, $v_{\mathsf{S}} = p s_{\mathsf{S}}^2$.

For $\lambda = v_{\mathsf{T}}/(v_{\mathsf{T}} + v_{\mathsf{S}} + D^2)$, Proposition 7 gives, for $\|\nu - \mu\| \le D$,

$$\mathbb{E}\left\|(1-\lambda)x + \lambda y - \mu\right\|^2 = (1-\lambda)^2 v_{\mathsf{T}} + \lambda^2(v_{\mathsf{S}} + \|\nu - \mu\|^2) \le D_*(D).$$

Use two submodels for the lower bound; the second uses testing (Tsybakov, 2009, Chapter 2). When $\nu = \mu$, the precision-weighted average of x and y is sufficient for μ, with coordinate variance $(s_{\mathsf{T}}^{-2} + s_{\mathsf{S}}^{-2})^{-1}$. The Gaussian location minimax theorem (Lehmann and Casella, 1998, Chapter 5, Example 4.6) gives

$$\mathcal{M}(D) \ge p\left(\frac{1}{s_{\mathsf{T}}^2} + \frac{1}{s_{\mathsf{S}}^2}\right)^{-1} = \frac{v_{\mathsf{T}} v_{\mathsf{S}}}{v_{\mathsf{T}} + v_{\mathsf{S}}}.$$

For $\nu = 0$ and $\|\mu\| \le D$, y is pure noise. Assign independent uniform priors on $\mu_j \in \{-a, a\}$, where $a = \min\{D/\sqrt{p},\ s_{\mathsf{T}}\}$. Then $\|\mu\| = a\sqrt{p} \le D$. For coordinate j, only x_j contains information about μ_j. Estimating μ_j reduces to Gaussian testing: $\mu_j = a$ and $\widehat{\mu}_j \le 0$ imply $(\widehat{\mu}_j - a)^2 \ge a^2$, with the symmetric bound for $\mu_j = -a$. The Neyman–Pearson bound gives

$$\frac{1}{2}\mathbb{E}_a(\widehat{\mu}_j - a)^2 + \frac{1}{2}\mathbb{E}_{-a}(\widehat{\mu}_j + a)^2 \ge a^2 \Phi\left(-\frac{a}{s_{\mathsf{T}}}\right) \ge a^2 \Phi(-1),$$

where Φ is the standard normal distribution function and $\mathbb{E}_{\pm a}$ averages over the prior on the remaining coordinates. Summing over coordinates and using $\Phi(-1) \ge 0.15$,

$$\mathcal{M}(D) \ge 0.15\, p a^2 = 0.15 \min(D^2, v_{\mathsf{T}}).$$

To combine the bounds, write $A = v_{\mathsf{T}} v_{\mathsf{S}}/(v_{\mathsf{T}} + v_{\mathsf{S}})$ and $B = \min(v_{\mathsf{T}}, D^2)$. Because $v_{\mathsf{T}} D^2/(v_{\mathsf{T}} + v_{\mathsf{S}} + D^2) \le v_{\mathsf{T}} D^2/(v_{\mathsf{T}} + D^2) \le B$,

$$D_*(D) = \frac{v_{\mathsf{T}} v_{\mathsf{S}}}{v_{\mathsf{T}} + v_{\mathsf{S}} + D^2} + \frac{v_{\mathsf{T}} D^2}{v_{\mathsf{T}} + v_{\mathsf{S}} + D^2} \le A + B.$$

The two lower bounds give $A \le \mathcal{M}(D)$ and $B \le \mathcal{M}(D)/0.15$, which implies

$$D_*(D) \le A + B \le \mathcal{M}(D)\left(1 + \frac{1}{0.15}\right) \le 8\,\mathcal{M}(D).$$

This completes the proof. □

G.3 Proof of Proposition 9

Proof. By definition,

$$\mathcal{R}_{\text{init}} < \mathcal{R}_{\text{data}}$$

is equivalent to

$$p(n_{\mathsf{S}}, \delta_H) v_{\mathsf{T}} + \{1 - p(n_{\mathsf{S}}, \delta_H)\} R_0 < q D_* + (1-q) R_0.$$

Moving terms gives

$$p(n_{\mathsf{S}}, \delta_H)(R_0 - v_{\mathsf{T}}) > q(R_0 - D_*).$$

The formula for the crossover follows by substituting the expression for D_*. Next,

$$\mathcal{R}_{\text{init}} - \mathcal{R}_{\text{both}} = p(n_{\mathsf{S}}, \delta_H)(v_{\mathsf{T}} - D_*) \ge 0,$$

with strict inequality when $p(n_{\mathsf{S}}, \delta_H) > 0$ and $D_* < v_{\mathsf{T}}$. Finally,

$$\mathcal{R}_{\text{data}} - \mathcal{R}_{\text{both}} = \{p(n_{\mathsf{S}}, \delta_H) - q\}(R_0 - D_*),$$

which is nonnegative whenever $p(n_{\mathsf{S}}, \delta_H) \ge q$. □

G.4 Proof of Proposition 10

Proof. On the event in the statement, for each $c \in \{a, b\}$,

$$c_1 \left\| \widehat{\theta}_c - \theta^*_{\mathsf{T}} \right\|_H^2 \ge V_n(\widehat{\theta}_c) - c_0 - \gamma_n, \qquad c_1 \left\| \widehat{\theta}_c - \theta^*_{\mathsf{T}} \right\|_H^2 \le V_n(\widehat{\theta}_c) - c_0 + \gamma_n.$$

Because $V_n(\widehat{\theta}_{\text{sel}}) = \min\{V_n(\widehat{\theta}_a), V_n(\widehat{\theta}_b)\}$,

$$c_1 \left\| \widehat{\theta}_{\text{sel}} - \theta^*_{\mathsf{T}} \right\|_H^2 \le \min_{c \in \{a,b\}} V_n(\widehat{\theta}_c) - c_0 + \gamma_n \le c_1 \min_{c \in \{a,b\}} \left\| \widehat{\theta}_c - \theta^*_{\mathsf{T}} \right\|_H^2 + 2\gamma_n.$$

On the complement of the event, $\left\| \widehat{\theta}_{\text{sel}} - \theta^*_{\mathsf{T}} \right\|_H^2 \le B_0$. Take expectations and use nonnegativity of the minimum. The weaker comparison follows from $\mathbb{E}\min\{A, B\} \le \min\{\mathbb{E}A, \mathbb{E}B\}$. This completes the proof. □

H Source Sample Size and Target Coverage under Pooling

A source fraction tending to zero does not automatically remove pooling bias from target inference. Instead, the relevant condition compares the source sample size with the square root of the target sample size. We derive this boundary for the raw pooled criterion. In contrast, source initialization avoids this relative-rate condition when both sample sizes diverge and the conditions of Proposition 19 hold.

We use the model, signal, and moment conditions of Theorem 3 and first define its pooled criterion and covariance. We write $n = n_{\mathsf{T}}$.

We choose signs such that $c = (\theta^*_{\mathsf{S}})^\top \theta^*_{\mathsf{T}} > 0$, and we assume both signals have norm one. For $0 < w < 1$, we define

$$A_w = (1-w)\theta^*_{\mathsf{T}}(\theta^*_{\mathsf{T}})^\top + w\theta^*_{\mathsf{S}}(\theta^*_{\mathsf{S}})^\top, \qquad \lambda_\pm(w) = \frac{1 \pm \sqrt{1 - 4w(1-w)(1-c^2)}}{2}.$$

We define v_w as the unit leading eigenvector of A_w, maintaining $v_w^\top \theta^*_{\mathsf{T}} > 0$, and we construct

$$\theta_w = \left\{ \frac{1 + 2\lambda_+(w)}{3} \right\}^{1/2} v_w.$$

We express the parameter-dependent part of the pooled population loss as

$$Q_w(u) = \frac{3}{4}\|u\|^4 - \frac{1}{2}\|u\|^2 - u^\top A_w u.$$

For each study, we obtain this expression by subtracting $Y^2/4$ from the observation loss before taking expectations. This subtraction preserves every sample minimizer and gradient. It enables population comparison under finite first absolute noise moments, even when the expected squared loss diverges.

We define the total size $N = n_\mathsf{T} + n_\mathsf{S}$ and the source fraction $w_N = n_\mathsf{S}/N$. We define $F_{\mathsf{S},n_\mathsf{S}}$ by the same quartic loss as $F_{\mathsf{T},n_\mathsf{T}}$, using the source observations. We pool the two empirical criteria as

$$F_{P,N} = (1 - w_N)F_{\mathsf{T},n_\mathsf{T}} + w_N F_{\mathsf{S},n_\mathsf{S}}.$$

We compute the initial estimate by multiplying a unit leading eigenvector of the pooled response-weighted second-moment matrix by the square root of the positive part of the pooled response mean. We then apply gradient descent to $F_{P,N}$ with step $1/8$ and the stopping rule (14) with N replacing n; we denote this output by $\widehat{u}_P$.

For the score of observation i in study a, we define

$$g_{a,i}(u) = \{(X_{a,i}^\top u)^2 - Y_{a,i}\}(X_{a,i}^\top u)X_{a,i}, \qquad \overline{g}_a(u) = \frac{1}{n_a}\sum_{i=1}^{n_a} g_{a,i}(u).$$

We define the empirical Hessian $\widehat{H}_P$ of the pooled criterion evaluated at $\widehat{u}_P$ and the covariance estimates

$$\widehat{\Sigma}_P = \sum_{a\in\{\mathsf{T},\mathsf{S}\}} \frac{n_a}{N}\frac{1}{n_a}\sum_{i=1}^{n_a}\{g_{a,i}(\widehat{u}_P) - \overline{g}_a(\widehat{u}_P)\}\{g_{a,i}(\widehat{u}_P) - \overline{g}_a(\widehat{u}_P)\}^\top,$$
$$\widehat{V}_P = \widehat{H}_P^{-1}\widehat{\Sigma}_P\widehat{H}_P^{-1}.$$

We center the scores separately within each study. We fix their sample sizes by design; our model does not treat the study label as a random variable.

We let g_a denote a generic score from study a.

The following proof formally verifies covariance consistency under the theorem's conditions using only a finite first absolute source-noise moment.

Theorem 3 guarantees nominal limiting target coverage when $n_\mathsf{S} = o(\sqrt{n})$. However, a positive square-root-order source sample introduces a nonzero mean into the target limiting distribution, even as its fraction of the pooled sample tends to zero. In this regime, the limiting normal distribution has the target covariance V_T. This analysis requires only a first absolute source-noise moment because source sampling variation becomes negligible at the target scale, while the source population score contributes to the limiting mean when $a > 0$.

Proof of Theorem 3. Because $n_\mathsf{S}/\sqrt{n} \to a$, $N/n \to 1$, $w_N \to 0$, and $\sqrt{n}w_N \to a$. Gaussian moments and the finite first absolute source-noise moment give uniform convergence of the source empirical Hessian on a bounded parameter ball and convergence of its score at θ^*_T. The target score satisfies the central limit theorem. Thus

$$\sqrt{n}\,\nabla F_{P,N}(\theta^*_\mathsf{T}) = (1 - w_N)\sqrt{n}\,\nabla F_{\mathsf{T},n}(\theta^*_\mathsf{T}) + \sqrt{n}w_N\,\overline{g}_\mathsf{S}(\theta^*_\mathsf{T}) \Rightarrow \mathsf{N}(ag^*_\mathsf{S}, \nu^2_\mathsf{T}H/2).$$

The pooled empirical Hessian converges uniformly to the target Hessian on a sufficiently small fixed ball. The pooled response-weighted second moment and response mean converge to their target values. Hence the raw pooled spectral start enters that ball with probability tending to one. The stationary-point and contraction proof of Theorem 5 applies with population comparison $Q_\mathsf{T}(u) = 3\|u\|^4/4 - \|u\|^2/2 - (u^\top\theta^*_\mathsf{T})^2$. The same gradient stopping bound gives

$$\sqrt{n}(\widetilde{u}_P - \theta^*_\mathsf{T}) = -H^{-1}\sqrt{n}\,\nabla F_{P,N}(\theta^*_\mathsf{T}) + o_\mathbb{P}(1),$$

which proves the shifted normal limit.

We verify covariance consistency without a second source-noise moment. On a fixed bounded ball, the source score norm is bounded by a constant times

$$J_i = \|X_{\mathsf{S},i}\|^4 + |\varepsilon_{\mathsf{S},i}| \, \|X_{\mathsf{S},i}\|^2 , \qquad \mathbb{E}J_i < \infty.$$

For independent copies of an integrable nonnegative variable, $\max_{i\le m} J_i/m \to 0$ in probability: for every $\epsilon > 0$,

$$\mathbb{P}\{\max_{i\le m} J_i > \epsilon m\} \le m\mathbb{P}\{J_1 > \epsilon m\} \le \epsilon^{-1}\mathbb{E}\{J_1 \, 1_{\{J_1>\epsilon m\}}\} \to 0.$$

Almost-sure convergence of the sample averages gives $m^{-1}\sum_i J_i = O_{\mathbb{P}}(1)$. With $m = n_{\mathsf{S}}$,

$$w_N \frac{1}{m}\sum_{i=1}^{m} J_i^2 \le \frac{m^2}{N}\left(\frac{\max_{i\le m} J_i}{m}\right)\left(\frac{1}{m}\sum_{i=1}^{m} J_i\right) = o_{\mathbb{P}}(1).$$

The within-source centered covariance is positive semidefinite and is bounded above by the uncentered second moment. Its weighted contribution to $\widehat{\Sigma}_P$ converges to zero. For the target scores, the finite second target-noise moment and Gaussian moments give uniform consistency of their empirical covariance near θ^*_{T}, as in the proof of Proposition 21. The weighted target contribution converges to $\nu^2_{\mathsf{T}} H/2$. Uniform Hessian consistency now gives $\widehat{V}_P \to V_{\mathsf{T}}$.

The explicit population minimizer is differentiable at $w = 0$. Alternatively, expansion of its zero population score at θ^*_{T} gives

$$0 = H(\theta_w - \theta^*_{\mathsf{T}}) + w g^*_{\mathsf{S}} + O\{\|\theta_w - \theta^*_{\mathsf{T}}\|^2 + w\,\|\theta_w - \theta^*_{\mathsf{T}}\|\}.$$

The eigenvalue and eigenvector formulas first give $\|\theta_w - \theta^*_{\mathsf{T}}\| = O(w)$. Hence

$$\theta_w - \theta^*_{\mathsf{T}} = wb + O(w^2). \tag{8}$$

For $w = w_N = O(n^{-1/2})$, the centered limit at θ_{w_N} is $\mathsf{N}(0, V_{\mathsf{T}})$, proving its coverage conclusion.

Because $N/n \to 1$, the aligned target Wald statistic converges to the noncentral chi-square distribution with parameter $a^2 b^\top V_{\mathsf{T}}^{-1} b$. The other sign is separated by a fixed distance, and continuity of this distribution gives the stated coverage. This completes the proof. □

To obtain a direct expression, we decompose $\theta^*_{\mathsf{S}} = c\theta^*_{\mathsf{T}} + (1-c^2)^{1/2} v$, where $v^\top \theta^*_{\mathsf{T}} = 0$ and $\|v\| = 1$. This yields

$$b = -\frac{1-c^2}{3}\theta^*_{\mathsf{T}} + c(1-c^2)^{1/2} v, \qquad a^2 b^\top V_{\mathsf{T}}^{-1} b = \frac{4a^2(1-c^2)}{\nu^2_{\mathsf{T}}}\left\{c^2 + \frac{1-c^2}{3}\right\}.$$

The noncentrality remains strictly positive for $a > 0$ and distinct signals. This formula quantifies the persistent coverage loss directly from the source sample-size coefficient, the signal separation, and the target noise. The strict reduction in coverage follows explicitly from the Poisson-mixture representation of the noncentral chi-square distribution: conditional on a positive Poisson count, its degrees of freedom strictly exceed p, making its probability below the fixed central quantile strictly smaller. Under the same signal and moment conditions, Proposition 19 preserves nominal limiting target coverage for source initialization as both sample sizes diverge, with no relative growth condition.

If $n_{\mathsf{S}}/\sqrt{n} \to \infty$ and $n_{\mathsf{S}}/n \to 0$, we also assume that $\mathbb{E}\varepsilon^2_{\mathsf{S}} < \infty$. Extending the centered polynomial-score calculation in the proof of Proposition 21 to $w_N \to 0$ gives $\sqrt{n}(\widehat{u}_P - \theta_{w_N}) = O_{\mathbb{P}}(1)$ and $\widehat{V}_P \to V_{\mathsf{T}}$ in probability. By (8),

$$\sqrt{n}\, d_{\pm}(\theta_{w_N}, \theta^*_{\mathsf{T}}) = \sqrt{n}\, w_N\{\|b\| + O(w_N)\} \longrightarrow \infty.$$

The confidence-region radii remain $O_{\mathbb{P}}(n^{-1/2})$, which makes target coverage tend to zero. We conclude that, under finite noise variances, the source sample-size regimes below, at, and above square-root order yield nominal limiting coverage, coverage below nominal, and zero limiting coverage for the raw pooled estimator. This distinction applies specifically to inference for the fixed target signal; the underlying pooled population parameter remains consistently estimable.

H.1 Larger source samples under a first-moment condition

A finite source variance is strictly unnecessary to prove the zero-coverage conclusion. We establish this by comparing the observed confidence radius directly with the population shift, even when the source sample matches or exceeds the target sample. We apply the estimates and regions defined in Theorem 3.

Corollary 6. *Assume fixed p, the stated Gaussian designs and independent centered noise, fixed unit signals, and*

$$0 < d_{\pm}(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*) < \frac{1}{8}, \qquad \mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty, \qquad 0 < \nu_{\mathsf{T}}^2 < \infty.$$

If $n, n_{\mathsf{S}} \to \infty$ and $n_{\mathsf{S}}/\sqrt{n} \to \infty$, then

$$\mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_n(\beta)\} \longrightarrow 1 - \beta, \qquad \mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_P(\beta)\} \longrightarrow 0.$$

Corollary 6 accommodates an infinite source variance and removes any upper bound on the source sample size. The pooled empirical covariance does not need to converge. The confidence radius divided by the distance of the nearest center from the target tends to zero in probability. This conclusion concerns target coverage and does not assert normal inference for the pooled population parameter.

Proof of Corollary 6. Put $m = n_{\mathsf{S}}$, $N = n + m$, $w = m/N$, and $\delta = d_{\pm}(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*)$. Align the two signals. Extend θ_w continuously to $w = 0, 1$, with $\theta_0 = \theta_{\mathsf{T}}^*$ and $\theta_1 = \theta_{\mathsf{S}}^*$. The eigengap of A_w is bounded below by $c > 0$ on $[0, 1]$. The formulas in Proposition 20 give

$$\max_{a \in \{\mathsf{T}, \mathsf{S}\}} \|\theta_w - \theta_a^*\| \le \delta + \delta^2/3, \qquad \|v_w - \theta_w\| \le \delta^2/3 < 1/192.$$

As in the proof of Proposition 21, the ball of radius $1/32$ about θ_w is contained in the ball of radius $1/6$ about each signal. Its population pooled Hessian has eigenvalues in $[1, 37/4]$, uniformly over $w \in [0, 1]$.

For a bounded set of parameter vectors, every entry of the sample score and Hessian is a polynomial in that vector with finitely many random coefficients. The centered target score coefficients have finite second moments. The source coefficients and both Hessian coefficient arrays are integrable. Independence of noise and design and Gaussian moments justify these statements. The coefficient-wise central limit theorem and strong convergence of sample means give, on any fixed bounded set K,

$$\sup_{u \in K} \|\overline{g}_{\mathsf{T}}(u) - \mathbb{E} g_{\mathsf{T}}(u)\| = O_{\mathbb{P}}(n^{-1/2}),$$
$$\sup_{u \in K} \|\overline{g}_{\mathsf{S}}(u) - \mathbb{E} g_{\mathsf{S}}(u)\| = o_{\mathbb{P}}(1),$$
$$\sup_{u \in K} \left\|\nabla^2 F_{P,N}(u) - \nabla^2 Q_w(u)\right\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Choose K containing all the local balls just described. Because the weighted population score at θ_w is zero,

$$\|\nabla F_{P,N}(\theta_w)\| = O_{\mathbb{P}}(\sqrt{n}/N) + o_{\mathbb{P}}(m/N) = o_{\mathbb{P}}(w),$$

where the last equality uses $\sqrt{n}/m \to 0$.

The member of the pooled sign pair closest to θ_w is also closest to θ_{T}^* on these local balls. The pooled response mean converges to one, and its weighted second-moment matrix differs from $I_p + 2A_w$ by $o_{\mathbb{P}}(1)$. The uniform eigengap and spectral perturbation imply that the sign-aligned pooled initializer is $v_w + o_{\mathbb{P}}(1)$. It enters the radius $1/64$ ball about θ_w with probability tending to one. Uniform Hessian convergence and the local deterministic proof of Theorem 5 give a unique stationary point $\widehat{\theta}_w$ in the radius $1/32$ ball and

$$\left\|\widehat{\theta}_w - \theta_w\right\| \le 2\,\|\nabla F_{P,N}(\theta_w)\| = o_{\mathbb{P}}(w).$$

Here the empirical Hessian eigenvalues lie in $[1/2, 39/4]$ with probability tending to one. The stationary point is within $1/256$ of θ_w with probability tending to one. The initializer is then within

$5/256$ of that point. The resulting ball about the stationary point lies inside the radius $1/32$ ball about θ_w. On this ball the step $1/8$ contracts by at most $15/16$, and all iterates stay in it. The stopping argument in Proposition 19, with sample size N, gives

$$\left\|\widetilde{u}_P - \widehat{\theta}_w\right\| \le \frac{2}{N\log(N+1)} = o(w), \qquad \|\widetilde{u}_P - \theta_w\| = o_{\mathbb{P}}(w).$$

These conclusions are uniform in the possible deterministic sequences $w = m/N$; convergence of w is unnecessary. In particular, $\left\|\widehat{H}_P^{-1}\right\|_{\mathrm{op}} \le 2$ with probability tending to one.

For $w > 0$, Proposition 20 gives $d_{\pm}(\theta_w, \theta_{\mathsf{T}}^*) > 0$. Equation (8) and $b \neq 0$ imply

$$\lim_{w \downarrow 0} \frac{d_{\pm}(\theta_w, \theta_{\mathsf{T}}^*)}{w} = \|b\| > 0.$$

The ratio extends to a positive continuous function on $[0,1]$. There is a constant $c_0 > 0$, depending only on the fixed signals, such that

$$d_{\pm}(\theta_w, \theta_{\mathsf{T}}^*) \ge c_0 w, \qquad \mathbb{P}\{d_{\pm}(\widehat{u}_P, \theta_{\mathsf{T}}^*) \ge c_0 w/2\} \longrightarrow 1.$$

This constant is used only in the proof, not in the procedure.

It remains to bound the observed region. For each study set $J_{a,i} = \|X_{a,i}\|^4 + |\varepsilon_{a,i}|\,\|X_{a,i}\|^2$. On a fixed bounded parameter set, $\|g_{a,i}(u)\| \le C J_{a,i}$. The target satisfies $\mathbb{E}J_{\mathsf{T},i}^2 < \infty$, while the source satisfies $\mathbb{E}J_{\mathsf{S},i} < \infty$. Hence,

$$\sum_{i=1}^{n} J_{\mathsf{T},i}^2 = O_{\mathbb{P}}(n), \qquad \frac{1}{m^2}\sum_{i=1}^{m} J_{\mathsf{S},i}^2 \le \frac{\max_{i \le m} J_{\mathsf{S},i}}{m}\,\frac{1}{m}\sum_{i=1}^{m} J_{\mathsf{S},i} = o_{\mathbb{P}}(1).$$

For any $\epsilon > 0$, integrability gives

$$\mathbb{P}\{\max_{i \le m} J_{\mathsf{S},i} > \epsilon m\} \le \epsilon^{-1}\mathbb{E}\{J_{\mathsf{S},1} 1_{\{J_{\mathsf{S},1} > \epsilon m\}}\} \longrightarrow 0;$$

the other factor converges to $\mathbb{E}J_{\mathsf{S},1}$.

For a symmetric matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote its largest and smallest eigenvalues. Within-study centering subtracts a positive semidefinite matrix from the uncentered second moment. On the event just proved,

$$\lambda_{\max}(\widehat{V}_P) \le \frac{C}{N}\left\{\sum_{i=1}^{n} J_{\mathsf{T},i}^2 + \sum_{i=1}^{m} J_{\mathsf{S},i}^2\right\}.$$

Put $r_N = \{\chi^2_{p,1-\beta}\lambda_{\max}(\widehat{V}_P)/N\}^{1/2}$ when the pooled region is nonempty, and set $r_N = 0$ when the region is empty. Each ellipsoid then lies in the Euclidean ball of radius r_N about its center. Because $N^2 w^2 = m^2$,

$$\frac{r_N^2}{w^2} \le C\left\{O_{\mathbb{P}}(n/m^2) + \frac{1}{m^2}\sum_{i=1}^{m} J_{\mathsf{S},i}^2\right\} = o_{\mathbb{P}}(1).$$

The exceptional event for bounded iterates and invertible Hessian has probability tending to zero. On its complement, coverage of the target requires $d_{\pm}(\widehat{u}_P, \theta_{\mathsf{T}}^*) \le r_N$. Hence

$$\mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_P(\beta)\} \le \mathbb{P}\{d_{\pm}(\widehat{u}_P, \theta_{\mathsf{T}}^*) < c_0 w/2\} + \mathbb{P}\{r_N \ge c_0 w/2\} + o(1) \longrightarrow 0.$$

Proposition 19 proves the target-only coverage statement for every divergent source sample size under the same moments. This completes the proof. □

I Model Examples

For phase retrieval, we formally verify the target geometry, empirical concentration, and source localization, and we state a fully finite-sample transfer guarantee. For a fixed-dimensional Gaussian mixture, we verify local target equivalence and explicitly characterize its asymptotic distribution. We use low-rank matrix sensing and folded-concave sparse regression to illustrate how source localization and shift enter model-specific conditions. We do not assert additional finite-sample guarantees for these latter two examples. Appendix R provides the supporting mathematical calculations.

I.1 PHASE RETRIEVAL

We assume the target and source observations satisfy

$$Y_{a,i} = (X_{a,i}^\top \theta_a^*)^2 + \varepsilon_{a,i}, \qquad a \in \{\mathsf{T}, \mathsf{S}\},$$

where $X_{a,i} \sim \mathsf{N}(0, I_p)$. We define the target empirical objective as

$$F_{\mathsf{T},n_\mathsf{T}}(u) = \frac{1}{4n_\mathsf{T}} \sum_{i=1}^{n_\mathsf{T}} \{(X_{\mathsf{T},i}^\top u)^2 - Y_{\mathsf{T},i}\}^2.$$

We denote the population target criterion by $F_\mathsf{T} = \mathbb{E}_\mathsf{T} F_{\mathsf{T},n_\mathsf{T}}$. Because θ and $-\theta$ generate identical distributions, the parameter remains identified only up to sign. We evaluate all distances using the signed metric

$$d_\pm(u, \theta) = \min\{\|u - \theta\|, \|u + \theta\|\}.$$

We define the source-target shift as $d_\pm(\theta_\mathsf{S}^*, \theta_\mathsf{T}^*)$. The source estimator can be any estimator computed strictly from the source sample, such as the spectral estimator applied in Section T.

This phase-retrieval model admits a complete verification of our general theory. For $\|\theta_\mathsf{T}^*\| = 1$, Appendix O computes the population Hessian in closed form and proves the explicit local bounds

$$I_p \preceq \nabla^2 F(u) \preceq \tfrac{37}{4} I_p \qquad \text{whenever } \|u - \theta_\mathsf{T}^*\| \le 1/6,$$

where F denotes the noiseless population objective. We verify that its Hessian Lipschitz constant remains bounded by 21 on this same ball. Appendix O also proves concentration bounds for the empirical gradient and Hessian of the quartic loss. These bounds rigorously establish the target concentration condition in (5). Theorem 7 establishes the target-side guarantee for source-initialized gradient descent. Subsequently, Corollary 7 applies this guarantee to a specific source estimator.

For $n \ge 2$ and $s \ge 1$, we define the source spectral localization rate

$$b_{n,p}(s) = \sqrt{\frac{p+s}{n}} + \frac{(s + \log n)(p+s)}{n}.$$

Theorem 7. *Assume $Y_{\mathsf{T},i} = (X_{\mathsf{T},i}^\top \theta_\mathsf{T}^*)^2 + \varepsilon_{\mathsf{T},i}$ with $X_{\mathsf{T},i} \sim \mathsf{N}(0, I_p)$, $\|\theta_\mathsf{T}^*\| = 1$, and $\varepsilon_{\mathsf{T},i}$ independent of $X_{\mathsf{T},i}$ with $\mathbb{E}\varepsilon_{\mathsf{T},i} = 0$ and $\|\varepsilon_{\mathsf{T},i}\|_{\psi_2} \le \sigma$. Assume that, for some $a_\mathsf{S} \ge 0$ and $\epsilon_\mathsf{S} \in [0, 1)$,*

$$\mathbb{P}\{d_\pm(\widehat{\theta}_\mathsf{S}, \theta_\mathsf{S}^*) > a_\mathsf{S}\} \le \epsilon_\mathsf{S}, \qquad d_\pm(\theta_\mathsf{S}^*, \theta_\mathsf{T}^*) + a_\mathsf{S} \le 1/8.$$

Let θ^t denote gradient descent on $F_{\mathsf{T},n_\mathsf{T}}$ with step size $\alpha = 1/8$ and $\theta^0 = \widehat{\theta}_\mathsf{S}$. There are constants $c_\sigma, C_\sigma < \infty$ depending only on σ such that, if $n_\mathsf{T} \ge 2$ and $n_\mathsf{T} \ge C_\sigma(p + \log n_\mathsf{T})^2(\log n_\mathsf{T})^2$, then with probability at least $1 - \epsilon_\mathsf{S} - c_\sigma n_\mathsf{T}^{-1}$, for every $t \ge 0$,

$$d_\pm(\theta^t, \theta_\mathsf{T}^*) \le C_\sigma \sqrt{\frac{p + \log n_\mathsf{T}}{n_\mathsf{T}}} + \frac{3}{20}\left(\frac{15}{16}\right)^t.$$

Theorem 7 cleanly separates the roles of the two samples. Source localization formally controls entry into the signed target ball. Separately, the target sample strictly controls optimization contraction and local estimation error. We ensure no source observation enters the target criterion. Corollary 7 verifies the required source assumption using a specific spectral estimator.

Corollary 7. *Assume the target model in Theorem 7. Let the source sample follow the same observation model with a unit-norm signal θ_S^*, independent noise with ψ_2-norm at most σ, and the spectral estimator in Proposition 18. Fix $s_\mathsf{S} > \log 6$ and $n_\mathsf{S}, n_\mathsf{T} \ge 2$. If*

$$d_\pm(\theta_\mathsf{S}^*, \theta_\mathsf{T}^*) + C_\sigma b_{n_\mathsf{S},p}(s_\mathsf{S}) \le \frac{1}{8},$$

and $n_\mathsf{T} \ge C_\sigma(p + \log n_\mathsf{T})^2(\log n_\mathsf{T})^2$, then, with probability at least $1 - 6e^{-s_\mathsf{S}} - c_\sigma n_\mathsf{T}^{-1}$,

$$d_\pm(\theta^t, \theta_\mathsf{T}^*) \le C_\sigma \sqrt{\frac{p + \log n_\mathsf{T}}{n_\mathsf{T}}} + \frac{3}{20}\left(\frac{15}{16}\right)^t, \qquad t \ge 0.$$

Corollary 7 successfully links the two data-generating processes, the spectral source estimator, and target gradient descent into a single finite-sample guarantee. Theorem 7 proves that after $t \geq 8 \log n_{\mathsf{T}}$ iterations, the source-initialized iterate successfully attains the error rate $\{(p + \log n_{\mathsf{T}})/n_{\mathsf{T}}\}^{1/2}$. This matches the error bound for the target local estimator. If the entry probability tends to one for a fixed dimension p, Corollary 4 guarantees the sign-aligned iterate also achieves the target first-order distribution with $H = 4\theta^*_{\mathsf{T}}(\theta^*_{\mathsf{T}})^\top + 2I_p$ and $\Sigma = \mathbb{E}(\varepsilon^2_{\mathsf{T}})\{I_p + 2\theta^*_{\mathsf{T}}(\theta^*_{\mathsf{T}})^\top\}$. The source properties influence this guarantee strictly through the localization radius a_{S}, the failure probability ϵ_{S}, and the population shift. We verify the target sample condition using the quadratic-feature concentration bound in Lemma 9.

For noiseless Gaussian phase retrieval, existing analyses give recovery guarantees with a sample size proportional to p up to logarithmic factors (Candès, Li and Soltanolkotabi, 2015; Sun, Qu and Wright, 2018). These previous analyses rely on local regularity or geometric bounds. These techniques differ significantly from our use of uniform Hessian control on a ball of fixed radius. Section O.3 rigorously proves that uniform Hessian control requires at least quadratic sample growth in p.

On the source side, spectral initialization has been studied under Gaussian designs (Netrapalli, Jain and Sanghavi, 2013; Candès, Li and Soltanolkotabi, 2015; Cai, Li and Ma, 2016). Proposition 18 provides a self-contained finite-sample bound for the estimator formed from the leading empirical eigenvector and the empirical response mean. For a fixed distance $d_{\pm}(\theta^*_{\mathsf{S}}, \theta^*_{\mathsf{T}}) < 1/8$, its entry condition holds once the source localization bound is sufficiently small. Corollary 8 establishes the fixed-dimensional target normal limit for this same source estimator and target iteration under its weaker moment and signal conditions. Both sample sizes must diverge and target refinement must use enough updates, but no relative sample-size condition is needed.

This phase-retrieval example precisely illustrates why pooling source data fundamentally differs from source initialization. Let $w_{\mathsf{T}} > 0$ and $w_{\mathsf{S}} > 0$ be the target and source weights, with $w_{\mathsf{T}} + w_{\mathsf{S}} = 1$. For unit-norm signals, the population pooled quartic criterion depends on $w_{\mathsf{T}}\theta^*_{\mathsf{T}}(\theta^*_{\mathsf{T}})^\top + w_{\mathsf{S}}\theta^*_{\mathsf{S}}(\theta^*_{\mathsf{S}})^\top$. Whenever $d_{\pm}(\theta^*_{\mathsf{S}}, \theta^*_{\mathsf{T}}) > 0$, this pooled minimizer generally does not equal either underlying signal. In sharp contrast, the source-initialized target refit uses source data exclusively to select a starting point. It then optimizes the target criterion entirely alone.

I.2 Low-Rank Matrix Sensing

We let d_1 and d_2 denote the numbers of rows and columns. We define the target parameter as a rank-r matrix with factors $U^*_{\mathsf{T}} \in \mathbb{R}^{d_1 \times r}$ and $V^*_{\mathsf{T}} \in \mathbb{R}^{d_2 \times r}$:

$$M^*_{\mathsf{T}} = U^*_{\mathsf{T}}(V^*_{\mathsf{T}})^\top \in \mathbb{R}^{d_1 \times d_2}.$$

We observe sensing matrices $A_{\mathsf{T},i} \in \mathbb{R}^{d_1 \times d_2}$ and responses $Y_{\mathsf{T},i}$. We write tr for the trace of a square matrix. With matrix inner product $\langle A, M\rangle = \mathrm{tr}(A^\top M)$ and errors $\varepsilon_{\mathsf{T},i}$, we model the target observations as

$$Y_{\mathsf{T},i} = \langle A_{\mathsf{T},i}, M^*_{\mathsf{T}}\rangle + \varepsilon_{\mathsf{T},i}.$$

We use $\|\cdot\|_F$ for the Frobenius norm and specify the nonconvex factor criterion as

$$F_{\mathsf{T},n_{\mathsf{T}}}(U, V) = \frac{1}{2n_{\mathsf{T}}}\sum_{i=1}^{n_{\mathsf{T}}}\{\langle A_{\mathsf{T},i}, UV^\top\rangle - Y_{\mathsf{T},i}\}^2 + \frac{1}{8}\left\|U^\top U - V^\top V\right\|_F^2.$$

We include the balancing penalty to successfully remove a flat scaling direction in the local chart. The balanced factor pair is identifiable only up to a common right orthogonal transformation. We evaluate the quotient distance as

$$d_{\mathrm{rot}}\{(U, V), (U', V')\} = \min_{Q^\top Q = I_r}\left[\|U - U'Q\|_F^2 + \|V - V'Q\|_F^2\right]^{1/2}.$$

The source sample provides an initial factor pair $(\widehat{U}_{\mathsf{S}}, \widehat{V}_{\mathsf{S}})$. Initialization transfer consists of running target gradient descent starting directly from this factor pair. We measure successful basin entry using the rotation-invariant metric d_{rot}.

Existing low-rank analyses successfully bound the distance to the target factor class under specific sampling and initialization conditions (Tu et al., 2016; Sun and Luo, 2016; Chi, Lu and Chen, 2019).

A source localization bound controls the initial distance to the target factor class. Applying these results also requires their other sampling and initialization conditions. Establishing exact equality with an oracle-start limit requires a completely separate analysis of the target iteration in aligned coordinates. Deriving a first-order distribution requires an explicit local expansion of the target estimator. Neither of these strong conclusions follows from a simple statistical-error distance bound alone. Let M_{S}^* denote the source matrix parameter. For pooled source and target samples, the local risk calculation must explicitly incorporate the matrix shift

$$\Delta_M = M_{\mathsf{S}}^* - M_{\mathsf{T}}^*.$$

This matrix shift serves as the exact matrix analogue of the vector shift analyzed in our local second-moment comparison model.

I.3 Two-Component Gaussian Mixtures

We consider a balanced two-component Gaussian mixture with known covariance,

$$Z_{a,i} \sim \frac{1}{2}\mathsf{N}(\mu_a^*, I_p) + \frac{1}{2}\mathsf{N}(-\mu_a^*, I_p), \qquad a \in \{\mathsf{T}, \mathsf{S}\}.$$

For $\mu_{\mathsf{T}}^* \neq 0$, the population negative log-likelihood is nonconvex. The mean vector is identifiable only up to sign. We again use the signed metric $d_{\pm}$ to evaluate distances. We write the EM update for the target sample as

$$M_{\mathsf{T},n_{\mathsf{T}}}(\mu) = \frac{1}{n_{\mathsf{T}}} \sum_{i=1}^{n_{\mathsf{T}}} Z_{\mathsf{T},i} \tanh(Z_{\mathsf{T},i}^\top \mu).$$

We define the target local estimator as a fixed point of $M_{\mathsf{T},n_{\mathsf{T}}}$ within a neighborhood of μ_{T}^*. We use the source estimator $\widehat{\mu}_{\mathsf{S}}$ directly as the initial EM starting point for the target sample.

Existing local EM analyses give convergence regions and finite-sample error bounds (Balakrishnan, Wainwright and Yu, 2017). Global analyses characterize population iterates and sample-based consistency (Xu, Hsu and Maleki, 2016). For this specific balanced mixture with a fixed dimension p and a nonzero μ_{T}^*, Appendix R rigorously verifies local sample contraction with probability tending to one and derives a first-order expansion. Source accuracy and population differences affect the probability that $\widehat{\mu}_{\mathsf{S}}$ successfully enters this local region. Under local contraction, the target update map and its unique local fixed point depend only on target data. Finite iterates still depend on the source start. They have the target first-order limit when entry probability tends to one and optimization error is $o_{\mathbb{P}}(n_{\mathsf{T}}^{-1/2})$. Pooling source observations into a joint target mixture likelihood defines a fundamentally different estimator. When $d_{\pm}(\mu_{\mathsf{S}}^*, \mu_{\mathsf{T}}^*) > 0$, pooling can shift the population optimum away from the target mean.

I.4 Nonconvex Sparse Regression

We specify the target linear model with response Y_{T}, covariate vector $X_{\mathsf{T}} \in \mathbb{R}^p$, coefficient $\beta_{\mathsf{T}}^* \in \mathbb{R}^p$, and error ε_{T} as

$$Y_{\mathsf{T}} = X_{\mathsf{T}}^\top \beta_{\mathsf{T}}^* + \varepsilon_{\mathsf{T}}.$$

For the sample criterion, $Y_{\mathsf{T}} \in \mathbb{R}^{n_{\mathsf{T}}}$ and $X_{\mathsf{T}} \in \mathbb{R}^{n_{\mathsf{T}} \times p}$ denote the response vector and design matrix. We write P_λ for a folded-concave penalty with tuning parameter $\lambda > 0$. We define the target estimator to minimize the objective

$$\frac{1}{2n_{\mathsf{T}}} \|Y_{\mathsf{T}} - X_{\mathsf{T}}\beta\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|).$$

We apply a folded-concave penalty, such as the smoothly clipped absolute deviation penalty or the minimax concave penalty (Fan and Li, 2001; Zhang, 2010). This penalty can make the objective nonconvex. Under restricted strong convexity and the stated regularity, sample-size, and tuning conditions, previous work bounds estimation error for stationary points in a constrained parameter space (Loh and Wainwright, 2015). Loh (2017) formally establishes local oracle equivalence for robust M-estimators under additional beta-min, sample-size, and tuning conditions. We can use a computed source estimator $\widehat{\beta}_{\mathsf{S}}$ of the source coefficient β_{S}^* directly as the starting point for a local

coordinate or proximal algorithm applied to the target criterion. A source start in the target oracle region can support a local guarantee once convergence of the chosen algorithm in that region is established. To claim first-order equivalence, one must derive a fully model-specific expansion of the local target estimator. We do not derive that explicit expansion for sparse regression in this manuscript.

The source-data benchmark differs substantially. For conditionally centered source errors, pooling adds a population score term at the target coefficient involving $\beta^*_{\mathsf{S}} - \beta^*_{\mathsf{T}}$ and the source predictor second moments. If the true supports differ, this shift can also affect the scores on inactive coordinates and target variable selection. Our abstract risk comparison applies in the target oracle coordinates if its centering and covariance conditions hold there. Support recovery and entry need separate verification. Appendix R describes the support and local accuracy conditions to be checked.

J Proofs for Phase-Retrieval Transfer

Under the source-target condition in Theorem 7, our source localization bound places a sign-aligned initializer within the target ball with probability at least $1-\epsilon_{\mathsf{S}}$. The target curvature and concentration bounds then jointly establish geometric contraction and the ultimate target-sample error. We provide proofs for the auxiliary calculations used below in Appendix O.

J.1 Proof of Theorem 7

Proof. Let $n = n_{\mathsf{T}}$, $\theta^* = \theta^*_{\mathsf{T}}$, and $F_n = F_{\mathsf{T},n_{\mathsf{T}}}$. Because F_n is even, starts $-\theta^0$ and θ^0 give opposite iterates with equal sign-invariant distance to θ^*. Let $\widetilde{\theta}^0 \in \{\widehat{\theta}_{\mathsf{S}}, -\widehat{\theta}_{\mathsf{S}}\}$ satisfy $\left\|\widetilde{\theta}^0 - \theta^*\right\| = d_{\pm}(\widehat{\theta}_{\mathsf{S}}, \theta^*)$, and let $\widetilde{\theta}^t$ be the iterates started at $\widetilde{\theta}^0$; then $d_{\pm}(\theta^t, \theta^*) \le \left\|\widetilde{\theta}^t - \theta^*\right\|$ for every t.

The triangle inequality for $d_{\pm}$ and the source event in Theorem 7 give

$$d_{\pm}(\widehat{\theta}_{\mathsf{S}}, \theta^*) \le d_{\pm}(\widehat{\theta}_{\mathsf{S}}, \theta^*_{\mathsf{S}}) + d_{\pm}(\theta^*_{\mathsf{S}}, \theta^*) \le \frac{1}{8}$$

with probability at least $1 - \epsilon_{\mathsf{S}}$; on this event $\widetilde{\theta}^0 \in \mathbb{B}_{1/8}$.

We verify Condition 2 and (5) with

$$r = \frac{1}{8}, \qquad R = \frac{1}{6}, \qquad m = 1, \qquad L = \frac{37}{4}, \qquad M = 21.$$

Propositions 15 and 17 give the population bounds and $\nabla F(\theta^*) = 0$. In Lemma 9, take $s = 2\log n_{\mathsf{T}}$ and use its bounds as $\tau_{n_{\mathsf{T}}}$ and $\Delta_{n_{\mathsf{T}}}$, with joint probability at least $1 - 10n_{\mathsf{T}}^{-2}$. The condition $n_{\mathsf{T}} \ge C_\sigma(p + \log n_{\mathsf{T}})^2(\log n_{\mathsf{T}})^2$, for large C_σ, gives $p \le n_{\mathsf{T}}$ and bounds ℓ and b by $C\log n_{\mathsf{T}}$. For $q = p + \log n_{\mathsf{T}}$, the score bound is at most $C_\sigma\sqrt{q/n_{\mathsf{T}}}\{1 + q\log n_{\mathsf{T}}/\sqrt{n_{\mathsf{T}}}\}$, which is bounded by $C_\sigma\sqrt{q/n_{\mathsf{T}}}$. Its Hessian bound is at most $C_\sigma\{q\sqrt{\log n_{\mathsf{T}}/n_{\mathsf{T}}} + q^2\log n_{\mathsf{T}}/n_{\mathsf{T}}\}$. Increasing the sample-size constant gives

$$\Delta_{n_{\mathsf{T}}} \le \frac{1}{2} = \frac{m}{2}, \qquad \tau_{n_{\mathsf{T}}} \le \frac{1}{160}, \qquad r + \frac{4\tau_{n_{\mathsf{T}}}}{m} \le \frac{1}{8} + \frac{1}{40} < R,$$

which verifies Condition 2 in full.

Theorem 5 applies with step size $\alpha = 1/8$. Because $1/2 \le m_{n_{\mathsf{T}}} \le 1$ and $37/4 \le L_{n_{\mathsf{T}}} \le 39/4$, we have $\alpha < 2/L_{n_{\mathsf{T}}}$, $\alpha m_{n_{\mathsf{T}}} \ge 1/16$, and $37/32 \le \alpha L_{n_{\mathsf{T}}} \le 39/32$. Hence

$$\kappa_{n_{\mathsf{T}}} = \max\{|1 - \alpha m_{n_{\mathsf{T}}}|,\ |1 - \alpha L_{n_{\mathsf{T}}}|\} \le \max\left\{1 - \frac{1}{16},\ \frac{7}{32}\right\} = \frac{15}{16}.$$

On the intersection of the source event and the target stochastic event, Theorem 5 gives, for every $t \ge 0$,

$$\left\|\widetilde{\theta}^t - \theta^*\right\| \le \frac{2}{m}\left\|S_{n_{\mathsf{T}}}\right\| + \kappa^t_{n_{\mathsf{T}}}\left(r + \frac{2}{m}\left\|S_{n_{\mathsf{T}}}\right\|\right) \le 2\tau_{n_{\mathsf{T}}} + \left(\frac{15}{16}\right)^t\left(\frac{1}{8} + \frac{1}{80}\right).$$

The second factor is at most $3/20$, and the bound on $\tau_{n_{\mathsf{T}}}$ above gives $2\tau_{n_{\mathsf{T}}} \le C_\sigma\{(p + \log n_{\mathsf{T}})/n_{\mathsf{T}}\}^{1/2}$. A union bound gives exceptional probability at most $\epsilon_{\mathsf{S}} + 10n_{\mathsf{T}}^{-2} \le \epsilon_{\mathsf{S}} + c_\sigma n_{\mathsf{T}}^{-1}$. This completes the proof. □

J.2 Proof of Corollary 7

Proof. The source-target inequality implies $C_\sigma b_{n_S,p}(s_S) \le 1/8 < 1/4$. Proposition 18 gives

$$\mathbb{P}\{d_\pm(\widehat{\theta}_S, \theta^*_S) > C_\sigma b_{n_S,p}(s_S)\} \le 6e^{-s_S}.$$

The source-target inequality in the corollary verifies the localization premise of Theorem 7 with $a_S = C_\sigma b_{n_S,p}(s_S)$ and $\epsilon_S = 6e^{-s_S} < 1$. Applying that theorem completes the proof. □

On the target event established in Theorem 7, sign symmetry gives a second local minimum near $-\theta^*$. We use the explicit quotient metric defined in Appendix L to account for this nonidentifiability. We apply our local theory strictly in one chart at a time, immediately after selecting a specific sign representative.

J.3 Sparse phase retrieval

Local recovery from a nearby sparse start with order $k\log(ep/k)$ Gaussian measurements is established in Soltanolkotabi (2019); Jagatap and Hegde (2019). We give a direct proof for (4) and then relate its starting region to source estimation error. The proof gives a small absolute radius; it does not assign a numerical radius to the experiments.

Write $n = n_T$, $x = \theta^*_T \neq 0$, and let A have rows $X^\top_{T,i}$. For row and column index sets I, J, let $A_{I,J}$ denote the corresponding submatrix, write A_J when all rows are used, and write $A_{I,\cdot}$ when all columns are used. Set $q = \min(3k, p)$ and $a = 1/12$. Choose a fixed $0 < \epsilon < 1/2$ such that

$$h(\epsilon) + \frac{\epsilon}{2}\log 2 \le \frac{a}{32}, \qquad h(u) = -u\log u - (1-u)\log(1-u), \quad 0 < u < 1.$$

Such a choice exists because the left side tends to zero with ϵ. Put

$$D = q\log(9ep/q), \qquad \tau = \epsilon/4, \qquad r_0 = \tau\sqrt{\frac{\epsilon}{2(1+a)}}.$$

The following bounds control the linear update on sparse coordinates and the contribution of observations whose fitted projection has the wrong sign.

Lemma 1. *If $n \ge 64D/a^2$, then, with probability at least $1 - 2e^{-na^2/64} - e^{-na/16}$,*

$$\max_{|J|\le q}\left\|\frac{A_J^\top A_J}{n} - I_{|J|}\right\|_{\mathrm{op}} \le a, \qquad \max_{\substack{|I|\le \epsilon n\\ |J|\le q}}\left\|\frac{A_{I,J}^\top A_{I,J}}{n}\right\|_{\mathrm{op}} \le a.$$

Lemma 1 is uniform over the selected rows and sparse coordinates. Thus the row set may depend on the current iterate.

Proof. For each set J of size q, choose a $1/4$-net of its unit sphere with at most 9^q points. There are at most $\binom{p}{q}9^q \le e^D$ such points. For a symmetric matrix B, approximation of a unit eigenvector by this net gives $\|B\|_{\mathrm{op}} \le 2\max_v |v^\top Bv|$.

If U has a chi-square distribution with n degrees of freedom, then $\mathbb{E}e^{\lambda(U-n)} = e^{-n\lambda}(1-2\lambda)^{-n/2}$. Its logarithm is at most $2n\lambda^2$ for $|\lambda| \le 1/4$. Using $\lambda = \pm b/4$ gives $\mathbb{P}\{|U/n - 1| > b\} \le 2e^{-nb^2/8}$ for $0 < b < 1$. Apply this bound with $b = a/2$ at each net point. The failure probability is at most

$$2e^{D - na^2/32} \le 2e^{-na^2/64}.$$

This proves the first bound. Smaller supports are contained in sets of size q.

For fixed I, J, v with $|I| \le \epsilon n$ and $\|v\| = 1$, $\|A_{I,J}v\|^2$ is chi-square with $|I|$ degrees of freedom. Its moment generating function at $1/4$ gives

$$\mathbb{P}\{\|A_{I,J}v\|^2 > an/2\} \le \exp\{-an/8 + (\epsilon n/2)\log 2\}.$$

The number of sets I with $|I| \le \epsilon n$ is at most $e^{nh(\epsilon)}$. Indeed, for integers $j \le \epsilon n$, $\epsilon^j(1-\epsilon)^{n-j} \ge e^{-nh(\epsilon)}$, and the binomial expansion sums these weighted counts to at most one. A union bound over rows, supports and net points gives failure probability at most

$$\exp\{D + nh(\epsilon) + (\epsilon n/2)\log 2 - an/8\} \le e^{-an/16},$$

because $D \le na/32$. The net inequality proves the second bound. This completes the proof. □

Proof of Corollary 3. For fixed x, let B count the observations with $|X_{\mathsf{T},i}^\top x| \le \tau\|x\|$. Each indicator has probability at most $\tau = \epsilon/4$. Markov's inequality applied to 2^B gives

$$\mathbb{P}\{B > \epsilon n/2\} \le \exp\{-\epsilon n(\log 2/2 - 1/4)\} \le e^{-\epsilon n/12}.$$

Assume $B \le \epsilon n/2$ and the two matrix bounds of Lemma 1. For every k-sparse z with $h = z - x$ and $\|h\| \le r_0\|x\|$, define $I(z) = \{i : \operatorname{sgn}(X_{\mathsf{T},i}^\top z) \ne \operatorname{sgn}(X_{\mathsf{T},i}^\top x)\}$. A sign change, including a zero fitted projection, implies $|X_{\mathsf{T},i}^\top x| \le |X_{\mathsf{T},i}^\top h|$. Hence,

$$|I(z)| \le \frac{\epsilon n}{2} + \frac{\|Ah\|^2}{\tau^2\|x\|^2} \le \frac{\epsilon n}{2} + \frac{n(1+a)r_0^2}{\tau^2} = \epsilon n.$$

This bound holds simultaneously for all such z.

Let b be the vector before hard thresholding in (4), and let $z^+ = \mathcal{H}_k(b)$. Set $r_i = |X_{\mathsf{T},i}^\top x|\{\operatorname{sgn}(X_{\mathsf{T},i}^\top z) - \operatorname{sgn}(X_{\mathsf{T},i}^\top x)\}$. Then

$$b - x = (I_p - A^\top A/n)h + A^\top r/n, \qquad r_i = 0 \text{ for } i \notin I(z), \qquad |r_i| \le 2|X_{\mathsf{T},i}^\top h|.$$

Let S be the union of the supports of x, z^+, and let T also include the support of z. Thus $|S| \le 2k$ and $|T| \le 3k$. The first matrix bound gives $\|\{(I_p - A^\top A/n)h\}_S\| \le a\|h\|$. For every unit vector v supported on S, the second gives

$$|v^\top A^\top r/n| \le \frac{2}{n}\|A_{I(z),\cdot}v\|\,\|A_{I(z),\cdot}h\| \le 2a\|h\|.$$

Hard thresholding is a nearest Euclidean projection onto the k-sparse vectors. Because x is feasible,

$$\|z^+ - x\|^2 \le 2(z^+ - x)^\top(b - x) \le 2\|z^+ - x\|\,\|(b-x)_S\|, \qquad \|z^+ - x\| \le 6a\|z - x\| = \tfrac{1}{2}\|z - x\|.$$

Induction keeps the iterates in the same ball. The matrix events and the band count are unchanged when x is replaced by $-x$. Hence, simultaneously for all sparse starts with $d_\pm(z_0, x) \le r_0\|x\|$,

$$d_\pm(z_t, x) \le 2^{-t} d_\pm(z_0, x) \le (2/3)^t r_0\|x\|, \qquad t \ge 0.$$

Because $q \le 3k$, $D \le C_0 k\log(ep/k)$ for an absolute C_0. Increasing the sample constant proves the required target event with failure probability at most $4e^{-\gamma n}$, where $\gamma = \min(a^2/64, a/16, \epsilon/12) > 0$. A union bound with the source-entry event proves the stated weaker bound with coefficient nine. This completes the proof. □

For the entry statement, suppose the source population signal is k-sparse and, for deterministic $a_{\mathsf{S}} \ge 0$ and $0 \le \epsilon_{\mathsf{S}} \le 1$, $\mathbb{P}\{d_\pm(\widetilde{\theta}_{\mathsf{S}}, \theta_{\mathsf{S}}^*) > a_{\mathsf{S}}\} \le \epsilon_{\mathsf{S}}$. Set $\widehat{\theta}_{\mathsf{S}} = \mathcal{H}_k(\widetilde{\theta}_{\mathsf{S}})$ and assume $2a_{\mathsf{S}} + d_\pm(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*) \le r_0\|\theta_{\mathsf{T}}^*\|$. For any k-sparse x and either sign $s \in \{-1, 1\}$, the nearest-projection property gives $\|\mathcal{H}_k(z) - sx\| \le 2\|z - sx\|$. Minimizing over the sign proves $d_\pm\{\mathcal{H}_k(z), x\} \le 2d_\pm(z, x)$, regardless of the rule for ties. The triangle inequality then gives, with probability at least $1 - \epsilon_{\mathsf{S}}$,

$$d_\pm(\widehat{\theta}_{\mathsf{S}}, \theta_{\mathsf{T}}^*) \le 2a_{\mathsf{S}} + d_\pm(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*) \le r_0\|\theta_{\mathsf{T}}^*\|.$$

The comparison with target-only initialization concerns sufficient conditions for specified algorithms. The diagonal spectral methods in Cai, Li and Ma (2016); Jagatap and Hegde (2019) have quadratic sparsity conditions for signals with comparable nonzero magnitudes. Later bounds depend on energy concentration: the stable sparsity in Cai, Li and You (2023) and the partial-energy ratios in Xu, Zhang and Wang (2026) permit nearly linear sample sizes for other signal classes. For equal-magnitude nonzero coefficients, these bounds remain quadratic up to logarithmic factors. This comparison gives no lower bound for all target-only methods.

For dense phase retrieval, write K for the target sample constant in Corollary 2. Set $s = \log n_{\mathsf{T}}$ in Proposition 18. Its source localization formula applied to the target sample gives

$$b_{n_{\mathsf{T}},p}(\log n_{\mathsf{T}}) \le \frac{1}{\sqrt{K}\sqrt{p + \log n_{\mathsf{T}}}\log n_{\mathsf{T}}} + \frac{2}{K(p + \log n_{\mathsf{T}})\log n_{\mathsf{T}}}.$$

Let C_σ be the fixed spectral constant of Proposition 18. Choose $K \ge \max\{2, (16C_\sigma)^2, 32C_\sigma\}$ and at least the target sample constant already required. Then $n_{\mathsf{T}} \ge 3$, $s \ge 1$, and $C_\sigma b_{n_{\mathsf{T}},p}(\log n_{\mathsf{T}}) \le 1/8$. The target spectral start enters the ball with probability at least $1 - 6/n_{\mathsf{T}}$.

K Further Guarantees for Phase-Retrieval Inference

A source direction completely outside the local target ball can still yield valid target inference. Our target checks correctly accommodate inaccurate source estimates, while our observed normalization readily adapts to unknown Gaussian design covariances and signal scales. For target observations, we define the shorthand $(X_i, Y_i, \varepsilon_i) = (X_{\mathsf{T},i}, Y_{\mathsf{T},i}, \varepsilon_{\mathsf{T},i})$.

K.1 Target Inference from a Nonorthogonal Source

The fixed-dimensional inference result does not require the source signal to lie inside the local target ball. Under the stated moment conditions, a fixed nonorthogonal source direction suffices as both sample sizes diverge. The population updates first iteratively shift this direction toward the true target. The empirical updates then enter the local region where our specific first-order expansion applies with probability tending to one. We define $\widetilde{u}_n$ as the precise sign of $\widehat{u}_n$ located closest to θ^*_{T}, and we define $H = 2I_p + 4\theta^*_{\mathsf{T}}(\theta^*_{\mathsf{T}})^\top$.

Theorem 8. *Assume the stated Gaussian phase model with fixed unit signals, fixed p, independent centered noise, and*

$$n_{\mathsf{S}}, n \to \infty, \qquad |(\theta^*_{\mathsf{S}})^\top \theta^*_{\mathsf{T}}| > 0, \qquad \mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty, \qquad 0 < \nu^2_{\mathsf{T}} = \mathbb{E}\varepsilon^2_{\mathsf{T}} < \infty.$$

For the raw source spectral start and the stated target stopping rule,

$$\sqrt{n}(\widetilde{u}_n - \theta^*_{\mathsf{T}}) \Rightarrow \mathsf{N}\left(0, \frac{\nu^2_{\mathsf{T}}}{2} H^{-1}\right), \qquad \widehat{\nu}^2_n \xrightarrow{\mathbb{P}} \nu^2_{\mathsf{T}}, \qquad \mathbb{P}\{\theta^*_{\mathsf{T}} \in \mathcal{C}_n(\beta)\} \to 1 - \beta.$$

Theorem 8 accommodates a fixed source-target angle arbitrarily close to a right angle. It requires absolutely no relative sample-size condition or source-noise variance assumptions. Because this remains a pointwise fixed-dimensional result, it does not guarantee uniform coverage for angles approaching a right angle as the sample size grows. Our subsequent proof demonstrates that the population iterations required to reach a fixed target neighborhood grow at most logarithmically in the inverse source-target alignment. The proof rigorously excludes premature algorithmic stopping before the empirical path enters that specified neighborhood. An entirely independent uniform unit start provides the exact same limiting coverage in a fixed dimension. The initial source information can affect early iterates, while the limiting target covariance remains the same under the theorem's conditions.

Chen et al. (2019, Section 2.1) analyzed the signal and orthogonal components of population phase updates. We explicitly derive the calculation for our fixed step and theoretically justify the observed stopping rule under finite noise moments.

K.1.1 Entry of the population path

We denote $\theta = \theta^*_{\mathsf{T}}$, fixing $\|\theta\| = 1$, and we define the map

$$G(u) = u - \frac{1}{8}\{(3\|u\|^2 - 1)u - 2(u^\top\theta)\theta\}, \qquad z^{t+1} = G(z^t).$$

This map remains entirely unchanged by any centered independent additive noise. The subsequent lemma explicitly bounds the number of iterations necessary to reach a prescribed target neighborhood.

Lemma 2. *If $\|z^0\| = 1$ and $c = \theta^\top z^0 > 0$, then $z^t \to \theta$. For an absolute constant C and every $0 < \epsilon < 1$,*

$$\inf\{t \geq 0 : \|z^t - \theta\| \leq \epsilon\} \leq C\{\log(2/c) + \log(2/\epsilon)\}.$$

For a unit starting vector, Lemma 2 replaces proximity to the target signal with positive directional alignment. Its iteration bound concerns the discrete population updates and does not give a finite-sample optimization complexity guarantee.

Proof. Decompose $z^t = a_t\theta + v_t$, with $v_t^\top\theta = 0$, and put $s_t = \|z^t\|^2$. Direct substitution gives

$$a_{t+1} = \frac{11 - 3s_t}{8} a_t, \qquad v_{t+1} = \frac{9 - 3s_t}{8} v_t.$$

For $0 \le s \le 1$, the function $s(11-3s)^2/64$ is increasing and equals one at $s=1$. Hence

$$s_{t+1} \le \frac{s_t(11-3s_t)^2}{64} \le 1, \qquad a_{t+1} \ge a_t \ge c, \qquad \frac{\|v_t\|}{a_t} \le \frac{1}{c}\left(\frac{9}{11}\right)^t.$$

These inequalities hold by induction from $s_0 = 1$. If $c = 1$, the path is constant at θ, and the conclusion follows. Assume $0 < c < 1$. Put $\rho = 9/11$ and $T_0 = \lceil \log(2/c)/\log(11/9) \rceil$. For $t \ge T_0$, $\|v_t\|/a_t \le 1/2$. As long as $a_t < 1/\sqrt{2}$, this gives $s_t \le 5/8$ and $a_{t+1} \ge (73/64)a_t$. Thus some $T_1 \le C\log(2/c)$ satisfies $T_1 \ge T_0$ and $a_{T_1} \ge 1/\sqrt{2}$.

For $t \ge T_1$, put $e_t = 1 - a_t^2$. Because $1 - s_t = e_t - \|v_t\|^2 \ge 0$, the first recurrence gives

$$e_{t+1} \le \left(1 - \frac{3}{4}a_t^2\right)e_t + \frac{3}{4}a_t^2\|v_t\|^2 \le \frac{5}{8}e_t + \frac{3}{4}\|v_t\|^2.$$

Also $\|v_{T_1+k}\|^2 \le \rho^{2k}/4$. Summing the recurrence and using $e_{T_1} \le 1/2$ gives

$$e_{T_1+k} \le \frac{1}{2}\left(\frac{5}{8}\right)^k + \frac{3}{16}\sum_{j=0}^{k-1}\left(\frac{5}{8}\right)^{k-1-j}\rho^{2j} \le \frac{203}{43}\rho^{2k}.$$

Thus $\left\|z^{T_1+k} - \theta\right\|^2 = (1-a_{T_1+k})^2 + \|v_{T_1+k}\|^2 \le 5\rho^{2k}$. Taking $k \ge \log(\sqrt{5}/\epsilon)/\log(11/9)$ proves both claims. This completes the proof. □

K.1.2 The empirical path and stopping rule

We can precisely approximate the finite initial path without drawing any additional sample. The procedure uses all target observations at every single iteration. We first formally establish this discrete approximation and then tightly control the remaining optimization error.

Proof of Theorem 8. Choose the sign of θ^*_{S} such that $c = (\theta^*_{\mathsf{S}})^\top\theta > 0$. For analysis, choose the sign of the source spectral estimate in the same direction. The empirical updates are odd and the reported region is sign-symmetric. These choices do not change the procedure. Independence and $\mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty$ give

$$\frac{1}{n_{\mathsf{S}}}\sum_i Y_{\mathsf{S},i}X_{\mathsf{S},i}X_{\mathsf{S},i}^\top \xrightarrow{\mathbb{P}} I_p + 2\theta^*_{\mathsf{S}}(\theta^*_{\mathsf{S}})^\top, \qquad \frac{1}{n_{\mathsf{S}}}\sum_i Y_{\mathsf{S},i} \xrightarrow{\mathbb{P}} 1.$$

Each matrix entry is integrable. For $p \ge 2$, the population eigengap is two; for $p = 1$, the sign-aligned unit eigenvector is exact. The raw spectral estimate hence converges, after sign alignment, to θ^*_{S}.

Let $G_n(u) = u - \nabla F_{\mathsf{T},n}(u)/8$. For fixed p, its coefficients are averages of fourth-order Gaussian monomials and entries of $Y_{\mathsf{T}}XX^\top$. They are integrable under the target moment condition. Convergence of these finitely many empirical coefficients gives, for every fixed $B < \infty$,

$$\sup_{\|u\|\le B}\|G_n(u) - G(u)\| \xrightarrow{\mathbb{P}} 0, \qquad \sup_{\|u\|\le B}\left\|\nabla^2 F_{\mathsf{T},n}(u) - \nabla^2 F_{\mathsf{T}}(u)\right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0.$$

Define the unrestricted empirical iterates $u^{t+1} = G_n(u^t)$ from the aligned source estimate, without stopping. Continuity of G, the last bounds, and induction imply that, for every fixed integer T,

$$\max_{0\le t\le T}\left\|u^t - z^t\right\| \xrightarrow{\mathbb{P}} 0, \qquad z^0 = \theta^*_{\mathsf{S}}.$$

The induction remains localized to a fixed ball containing both the finite population path and a unit margin. The corresponding failure probability tends to zero. Crucially, we do not assume any statistical independence between a target iterate and the target observations.

If $c < 1$, Lemma 2 gives a finite, nonrandom T with $\left\|z^T - \theta\right\| < 1/16$. Thus $\left\|u^T - \theta\right\| < 1/8$ with probability tending to one. Before this time the population gradients are nonzero. Indeed, $a_t > 0$

and $v_t \neq 0$ at every finite time, whereas a stationary point with a positive signal component must have $s_t = 1$ and $v_t = 0$. Thus

$$\min_{0\le t<T} \left\|\nabla F_{\mathsf{T}}(z^t)\right\| > 0, \qquad \mathbb{P}\left\{\min_{0\le t<T} \left\|\nabla F_{\mathsf{T},n}(u^t)\right\| > \frac{1}{n\log(n+1)}\right\} \longrightarrow 1.$$

The second claim follows directly from the finite-path approximation and uniform gradient consistency. Thus our implemented stopping rule does not terminate during this initial phase, with probability tending to one. When $c = 1$, we set $T = 0$; in this case, source consistency immediately provides basin entry.

Proposition 17 and uniform Hessian consistency now give a unique local stationary point $\widehat{\theta}_{\mathsf{T}}$, with $\left\|\widehat{\theta}_{\mathsf{T}} - \theta\right\| = O_{\mathbb{P}}(n^{-1/2})$. Theorem 5 applies from time T, with contraction $15/16$ on the ball of radius $1/6$. On these events,

$$\left\|u^{T+k} - \widehat{\theta}_{\mathsf{T}}\right\| \le \left(\frac{15}{16}\right)^k \left(\frac{1}{8} + 2\left\|S_n\right\|\right), \qquad \left\|u^{T+k} - \widehat{\theta}_{\mathsf{T}}\right\| \le 2\left\|\nabla F_{\mathsf{T},n}(u^{T+k})\right\|,$$

where $S_n = \nabla F_{\mathsf{T},n}(\theta)$. The score satisfies $\sqrt{n}S_n \Rightarrow \mathsf{N}(0, \nu_{\mathsf{T}}^2 H/2)$, by Gaussian fourth moments and noise independence. Let $b_n = \lceil 64\log(n+1) + \{\log(n+1)\}^2 \rceil$ be the cap in (14). It satisfies $n\log(n+1)(15/16)^{b_n - T} \to 0$. The gradient threshold is reached before the cap, and the reported aligned output satisfies

$$\left\|\widetilde{u}_n - \widehat{\theta}_{\mathsf{T}}\right\| \le \frac{2}{n\log(n+1)} = o(n^{-1/2}).$$

The stationary-point expansion proved in Corollary 8 strictly guarantees the asserted normal limit. That expansion inherently describes the local root without using any source-target distance bound. Instead, our preceding path argument formally establishes the necessary basin entry.

Put $h_n = \widetilde{u}_n - \theta$. On $\|h_n\| \le 1$, the residual expansion in Proposition 19 gives

$$\left|\widehat{\nu}_n^2 - \frac{1}{n}\sum_i \varepsilon_{\mathsf{T},i}^2\right| \le 6\left\|h_n\right\| \frac{1}{n}\sum_i \left|\varepsilon_{\mathsf{T},i}\right| \left\|X_i\right\|^2 + 9\left\|h_n\right\|^2 \frac{1}{n}\sum_i \left\|X_i\right\|^4 = o_{\mathbb{P}}(1).$$

Both sample averages on the right remain bounded in probability. The sample mean of $\varepsilon_{\mathsf{T}}^2$ strictly converges to ν_{T}^2, which establishes formal variance consistency. Concurrently, $2I_p + 4\widehat{u}_n\widehat{u}_n^\top$ converges to H in probability. The aligned target Wald statistic converges in distribution to χ_p^2, whereas the corresponding statistic for the incorrect sign diverges. Slutsky's theorem and the continuity of the chi-square distribution prove our stated coverage limit. This completes the proof. □

This same theoretical conclusion applies to an entirely independent uniform unit start in any fixed dimension. Conditional on such a start, its true alignment remains nonzero with probability one, allowing our preceding finite-path argument to apply. We then integrate this conditional convergence via the bounded convergence theorem to guarantee the unconditional normal limit and the required coverage. We note that this observation requires an increasing target sample size and the stated iteration cap of order $(\log n)^2$. It emphatically does not compare fixed-iteration errors nor assess total computational costs.

K.2 Inference after Checking the Target Fit

A small target gradient alone cannot strictly guarantee that the optimization path has successfully reached one of the two valid target solutions. In particular, the gradient precisely at zero always evaluates to zero. We introduce the empirical target Hessian specifically to exclude such degenerate fits. The formal procedure detailed below requires absolutely no bound on the source-target distance, nor does it require any source noise moment assumption.

We explicitly treat any nonfinite iterate as a rejected fit. For each evaluated source starting value, we apply the step size $1/8$ and the finite stopping rule (14). We accept a completed fit u when

$$\|u\| \le 2, \qquad \|\nabla F_n(u)\| \le \frac{1}{n\log(n+1)}, \qquad \lambda_{\min}\{\nabla^2 F_n(u)\} \ge \frac{1}{2}, \tag{9}$$

where we define $F_n = F_{\mathsf{T},n}$ and assume $n = n_{\mathsf{T}} \geq 2$. We compute the Hessian strictly from the directly observed target criterion. We explicitly do not use the theoretical model-based matrix $2I_p + 4uu^\top$, because that theoretical matrix remains positive definite for every possible u. Analysts may provide any finite list of candidate source starts, and they may freely use the target data to select among any accepted fits.

If the procedure accepts no source fit, we actively compute the spectral starting value directly from the target observations using the identical formula. We then apply the identical target updates, stopping rule, and target check. We define $\widehat{u}_n^c$ as the finally selected accepted fit. If the check rejects all fits, we assign $\widehat{u}_n^c = 0$ and we conservatively report the full space $\mathbb{R}^p$ as our confidence region. For any formally accepted fit, we construct the region using (15) and (16) with $\widehat{u}_n^c$ replacing $\widehat{u}_n$. We also report $\mathbb{R}^p$ if the computed residual variance is exactly zero or undefined. We denote this resulting confidence region by $\mathcal{C}_n^c(\beta)$. We define $\widetilde{u}_n^c$ as the specific sign of $\widehat{u}_n^c$ closest to the true θ_{T}^*. We use this aligned sign strictly for mathematical analysis, not for algorithmic execution.

Theorem 4 systematically replaces the stringent source accuracy premise with a fully observed target check and a prescribed target restart. It expressly permits arbitrary target-dependent selection because, after proper sign alignment, every accepted fit mathematically approaches the exact same target estimator.

Theorem 4 in the main manuscript gives the distribution and coverage of this procedure. We prove it below.

Theorem 4 deliberately imposes no conditions on either the generating source distribution or the underlying number of source observations. The source estimator may be arbitrarily inaccurate, and the chosen source start may even depend directly on the target sample. Our subsequent proof derives uniform target gradient and Hessian consistency using only finite population moments. It then proves that every accepted fit is at distance at most $2/\{n\log(n+1)\}$ from the unique target local estimator after sign alignment, with probability tending to one. The procedure requires no independent validation sample and no statistical correction for selecting among multiple accepted fits. The explicit target restart guarantees that the formally reported region forms an ellipsoid union, with probability tending to one. However, this target restart may necessitate an entirely separate target fit. We thus caution that this specific theorem does not formally guarantee any reduction in total computational complexity. Finally, this particular statistical guarantee remains strictly asymptotic for a fixed dimension; it is not a fully finite-sample coverage statement.

The explicitly observed Hessian effectively separates a legitimate target solution from any spurious stationary point exhibiting negative curvature. We first formally establish a strict quantitative geometric property of the population phase criterion. Uniform empirical convergence then transfers this crucial property to every individual accepted empirical fit. Our rigorous argument relies heavily on the specific mathematical structure of the Gaussian phase model; it does not constitute a generic assertion about arbitrary nonconvex losses. Sun, Qu and Wright (2018) extensively studied the population and empirical geometry underlying phase retrieval, while Mei, Bai and Montanari (2018) analyzed uniform derivatives of general nonconvex empirical criteria. Our specific result below gives explicit constants sufficient for the proposed observed check and the subsequent statistical expansion.

K.2.1 A POPULATION GRADIENT AND CURVATURE BOUND

We denote $\theta = \theta_{\mathsf{T}}^*$, explicitly fixing $\|\theta\| = 1$. We subtract the strictly parameter-independent term from the population loss. The resulting criterion is

$$Q(u) = \frac{3}{4}\|u\|^4 - \frac{1}{2}\|u\|^2 - (u^\top\theta)^2.$$

Its gradient and Hessian are the population derivatives of F_n. A point with a small gradient and positive curvature must be close to one of the two target signals.

Lemma 3. *If $u \in \mathbb{R}^p$ satisfies*

$$\|\nabla Q(u)\| \leq \frac{1}{4}, \qquad \lambda_{\min}\{\nabla^2 Q(u)\} \geq \frac{1}{4},$$

then $d_\pm(u,\theta) \leq \|\nabla Q(u)\|$.

Lemma 3 uses neither a known neighborhood of θ nor a bound on $\|u\|$. The norm restriction in (9) is instead used to control the difference between sample and population derivatives on a fixed set.

Proof. Replace u by $-u$ if necessary, and write $u = s\theta + v$, where $s \geq 0$, $v^\top \theta = 0$, and $q = \|v\|$. Put $A = 3(s^2 + q^2) - 1$ and $\gamma = \|\nabla Q(u)\|$. Direct differentiation gives

$$\nabla Q(u) = (A-2)s\theta + Av, \qquad \theta^\top \nabla^2 Q(u)\theta = A - 2 + 6s^2.$$

Suppose $s \leq 1/2$. The curvature condition implies $A \geq 9/4 - 6s^2 \geq 3/4$, which gives $q \leq 4\gamma/3 \leq 1/3$. Hence,

$$A = 3(s^2 + q^2) - 1 \leq 3\left(\frac{1}{4} + \frac{1}{9}\right) - 1 = \frac{1}{12},$$

which contradicts $A \geq 3/4$. Thus $s > 1/2$, and

$$|A - 2| \leq 2\gamma, \qquad A \geq \frac{3}{2}, \qquad q \leq \frac{2\gamma}{3}, \qquad |s^2 + q^2 - 1| \leq \frac{2\gamma}{3}.$$

It follows that

$$|s - 1| \leq \frac{2\gamma/3 + 4\gamma^2/9}{s+1} \leq \frac{4\gamma}{9} + \frac{8\gamma^2}{27} \leq \frac{14\gamma}{27}.$$

Hence

$$d_\pm(u,\theta)^2 = (s-1)^2 + q^2 \leq \left\{\left(\frac{14}{27}\right)^2 + \left(\frac{2}{3}\right)^2\right\}\gamma^2 = \frac{520}{729}\gamma^2 \leq \gamma^2.$$

This completes the proof. □

Proof of Theorem 4. Step 1: Uniform target control.

We completely derive the necessary sample properties without a priori assuming that any observed fit lies within the actual target region. For a fixed p, we strictly define for this proof

$$D_n = \sup_{\|u\| \leq 2} \left\|\nabla^2 F_n(u) - \nabla^2 Q(u)\right\|_{\mathrm{op}}.$$

The empirical Hessian has entries

$$\{\nabla^2 F_n(u)\}_{jk} = 3\sum_{a,b=1}^{p} u_a u_b \left(\frac{1}{n}\sum_{i=1}^{n} X_{i,a}X_{i,b}X_{i,j}X_{i,k}\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i X_{i,j} X_{i,k}.$$

Each coefficient fundamentally admits a finite absolute expectation. Gaussian fourth moments give finite expectations for the coefficients in the first sum. For the specific noise term in the second sum, statistical independence directly gives

$$\mathbb{E}|\varepsilon_i X_{i,j} X_{i,k}| = \mathbb{E}|\varepsilon_i|\,\mathbb{E}|X_{i,j}X_{i,k}| < \infty.$$

The finite number of coefficients guarantees that their sample averages converge almost surely. For $\|u\| \leq 2$, we have $\sum_{a,b} |u_a u_b| \leq 4p$, while a standard matrix operator norm cannot exceed p times its absolute maximum entry. These bounds formally prove $D_n \to 0$ almost surely. Because both gradients precisely evaluate to zero at the origin, direct integration along the segment from zero to u establishes

$$\sup_{\|u\| \leq 2} \|\nabla F_n(u) - \nabla Q(u)\| \leq 2D_n, \qquad \|\nabla F_n(\theta)\| \leq D_n. \tag{10}$$

In particular, the target-sample event $D_n \leq 1/256$ has an asymptotic probability tending to one under our stated population moment conditions. We emphasize that no source-related quantity enters this specific event.

For sample sizes n strictly satisfying $\{n\log(n+1)\}^{-1} \leq 1/256$, any algorithmic point accepted by (9) simultaneously satisfies on this event

$$\|\nabla Q(u)\| \leq \frac{3}{256}, \qquad \lambda_{\min}\{\nabla^2 Q(u)\} \geq \frac{1}{2} - D_n \geq \frac{1}{4}.$$

Lemma 3 provides $d_{\pm}(u,\theta) \leq 3/256 < 1/8$, holding simultaneously for all accepted points. Crucially, this robust conclusion does not require any subsequent union bound over the user-provided candidate list.

Our explicit population phase curvature calculation in Appendix O gives $I_p \preceq \nabla^2 Q(u) \preceq (37/4) I_p$ on the restricted ball $\mathbb{B}_{1/6}$. Concurrently, $1/8 + 4D_n \leq 9/64 < 1/6$. The rigorous deterministic argument in Theorem 5 thus yields a unique target stationary point $\widehat{\theta}_{\mathsf{T}}$ in $\mathbb{B}_{1/6}$ such that $\left\|\widehat{\theta}_{\mathsf{T}} - \theta\right\| \leq 2\left\|\nabla F_n(\theta)\right\|$. The empirical sample Hessian bounds strictly from below by $I_p/2$ on this identical ball. For any given accepted point, we define $\widetilde{u}$ as its precise sign closest to θ. The direct segment connecting $\widetilde{u}$ to $\widehat{\theta}_{\mathsf{T}}$ demonstrably lies inside the designated ball, and

$$\begin{aligned}\frac{1}{2}\left\|\widetilde{u} - \widehat{\theta}_{\mathsf{T}}\right\|^2 &\leq (\widetilde{u} - \widehat{\theta}_{\mathsf{T}})^\top \{\nabla F_n(\widetilde{u}) - \nabla F_n(\widehat{\theta}_{\mathsf{T}})\} \\ &\leq \left\|\widetilde{u} - \widehat{\theta}_{\mathsf{T}}\right\| \left\|\nabla F_n(\widetilde{u})\right\|.\end{aligned}$$

The strict evenness of F_n and condition (9) explicitly enforce the uniform bound

$$\left\|\widetilde{u} - \widehat{\theta}_{\mathsf{T}}\right\| \leq \frac{2}{n\log(n+1)} = o(n^{-1/2}). \tag{11}$$

This geometric property explains exactly why any target-dependent choice among accepted fits structurally cannot alter the resulting first-order statistical distribution. We emphasize that our observed empirical check acts as a computable verification step, not as an abstract population assumption.

Step 2: Formal acceptance by the target restart.

The observed target spectral matrix and the empirical mean response jointly satisfy

$$\frac{1}{n}\sum_{i=1}^{n} Y_i X_i X_i^\top \longrightarrow I_p + 2\theta\theta^\top, \qquad \frac{1}{n}\sum_{i=1}^{n} Y_i \longrightarrow 1 \quad \text{almost surely.}$$

The exact same finite first-moment calculation directly proves both asymptotic limits. The leading population eigenspace explicitly remains one-dimensional; hence, continuity of the corresponding spectral projection strictly guarantees that the target spectral starting value approaches the unidentifiable set $\{\theta, -\theta\}$. We resolve sample ties by choosing a measurable leading eigenvector via any fixed rule. For $p = 1$, we only select its respective sign. Thus the assigned target start definitively enters one of the signed balls of radius $1/8$ with probability tending to one.

On the preceding target events, our fixed step size $1/8$ iteratively contracts the error by at most $15/16$. We define $S_n = \nabla F_n(\theta)$, allowing the proof of Proposition 19 to guarantee

$$\left\|\theta^t - \widehat{\theta}_{\mathsf{T}}\right\| \leq \left(\frac{15}{16}\right)^t \left(\frac{1}{8} + 2\left\|S_n\right\|\right), \qquad \left\|\nabla F_n(\theta^t)\right\| \leq \frac{39}{4}\left\|\theta^t - \widehat{\theta}_{\mathsf{T}}\right\|.$$

We align the explicit sign of the target sequence exclusively for our theoretical analysis. The explicit iteration cap is $b_n = \lceil 64\log(n+1) + \{\log(n+1)\}^2 \rceil$, as in (14), and satisfies $n\log(n+1)(15/16)^{b_n} \to 0$. Thus the actual target sequence successfully attains the required gradient threshold well before triggering the cap, with probability tending to one. The algorithmic output converges in strict sign-invariant distance to θ; concurrently, its Euclidean output norm converges to one while its empirical Hessian approaches $H = 2I_p + 4\theta\theta^\top$. We bound the smallest eigenvalue of this limit matrix by at least two. The remaining two inequalities in (9) structurally hold with probability tending to one. If the procedure previously accepted a given source fit, it has already returned an accepted fit. If the check rejects all provided source fits, the subsequent target restart successfully passes acceptance on the identical target event. This argument formally proves algorithmic acceptance with probability tending to one, holding uniformly for any finite measurable source list.

Step 3: Deriving normal limit and confidence coverage.

We compute the score exactly at the true target as

$$S_n = -\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta) X_i, \qquad \mathbb{E}\{\varepsilon_i^2 (X_i^\top\theta)^2 X_i X_i^\top\} = \nu_{\mathsf{T}}^2 (I_p + 2\theta\theta^\top) = \frac{\nu_{\mathsf{T}}^2}{2} H.$$

The standard multivariate central limit theorem readily applies under only our minimal finite second noise moment assumption. The explicitly integrated sample Hessian between $\widehat{\theta}_{\mathsf{T}}$ and θ correctly tends to H, while its smallest eigenvalue remains bounded safely away from zero with probability tending to one. The fundamental stationary equation directly implies

$$\sqrt{n}(\widehat{\theta}_{\mathsf{T}} - \theta) = -H^{-1}\sqrt{n}S_n + o_{\mathbb{P}}(1) \Rightarrow \mathsf{N}(0, V_{\mathsf{T}}).$$

For any successfully accepted output, (11) directly enforces this identical statistical limit. Because the strict probability of yielding no accepted fit asymptotically tends to zero, assigning a zero estimate on that rare event cannot theoretically disrupt convergence in distribution.

We let $\widehat{\nu}^2_{n,c}$ and $\widehat{H}_{n,c}$ denote the specified estimates defined in (15) when evaluated specifically at $\widehat{u}^c_n$, and we explicitly define $h_n = \widetilde{u}^c_n - \theta$. On the event $\|h_n\| \le 1$, the calculated residual variance at our accepted output obeys

$$\left| \widehat{\nu}^2_{n,c} - \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \right| \le 6\,\|h_n\|\,\frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i|\,\|X_i\|^2 + 9\,\|h_n\|^2\,\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^4.$$

Both sample averages strictly remain bounded in probability, and we have established $\|h_n\| = O_{\mathbb{P}}(n^{-1/2})$. This ensures $\widehat{\nu}^2_{n,c} \to \nu^2_{\mathsf{T}}$ in probability and identically verifies $\widehat{H}_{n,c} = 2I_p + 4\widehat{u}^c_n(\widehat{u}^c_n)^\top \to H$. We purposefully require no fourth noise moment.

Slutsky's theorem precisely gives

$$\frac{2n}{\widehat{\nu}^2_{n,c}}(\widetilde{u}^c_n - \theta)^\top \widehat{H}_{n,c}(\widetilde{u}^c_n - \theta) \Rightarrow \chi^2_p.$$

The parallel quadratic form evaluated for the opposite, incorrect sign diverges to infinity upon scalar multiplication by n. The minimum of the two sign-specific quadratic forms thus has the same chi-square limit. We report the exceptional region $\mathbb{R}^p$ only with probability asymptotically tending to zero. Formal continuity at the critical value $\chi^2_{p,1-\beta}$ rigorously proves Theorem 4.

This completes the proof. □

K.3 Unknown Gaussian Design Covariance and Signal Scale

The empirical target check can successfully be executed without assuming a previously known design covariance or a strict unit-norm signal constraint. We consistently estimate the design covariance and true signal scale solely from the given target observations. These computable estimates functionally change only the optimization coordinates, leaving the fundamental target estimating equation entirely intact. We allow conditional mean-zero noise under the moment and nondegeneracy conditions of Theorem 9. The corresponding noise variance may depend on the sensing vector.

We assume the observed target sample consists of independent copies obeying

$$X \sim \mathsf{N}(0, \Sigma), \qquad Y = (X^\top \theta^*)^2 + \varepsilon, \qquad \Sigma \succ 0, \quad \theta^* \neq 0,$$

where we assume neither Σ nor θ^* is theoretically known. We then define the following explicit population quantities strictly for theoretical analysis:

$$\tau = (\theta^*)^\top \Sigma \theta^*, \qquad H = 2\tau\Sigma + 4\Sigma\theta^*(\theta^*)^\top\Sigma, \qquad J = \mathbb{E}\{\varepsilon^2 (X^\top\theta^*)^2 X X^\top\}.$$

The limiting target score covariance remains explicitly nondegenerate whenever $J \succ 0$. The specified weighted noise moment given in the upcoming theorem provides a mathematically sufficient condition for bounding the finite target score variance. It emphatically does not represent a computable check on any physically observed sample.

K.3.1 An observed normalization and confidence region

We algorithmically compute $\widehat{\Sigma}_n = n^{-1}\sum_i X_i X_i^\top$ and $\widehat{\tau}_n = n^{-1}\sum_i Y_i$. If the resulting initial matrix lacks strict positive definiteness or if $\widehat{\tau}_n \le 0$, we set the final estimate to zero and report $\mathbb{R}^p$. Whenever both empirical checks pass, we subsequently define

$$Z_i = \widehat{\Sigma}_n^{-1/2} X_i, \qquad R_i = Y_i/\widehat{\tau}_n, \qquad T_n = \sqrt{\widehat{\tau}_n}\,\widehat{\Sigma}_n^{-1/2}, \qquad G_n(v) = \frac{1}{4n}\sum_{i=1}^{n}\{(Z_i^\top v)^2 - R_i\}^2.$$

All matrix square roots explicitly denote unique symmetric positive definite square roots. For any generic source starting vector u_0, we strictly initialize the normalized target iteration at $T_n^{-1}u_0$. We apply the fixed step size $1/8$, iteration cap $b_n = \lceil 64\log(n+1)+\{\log(n+1)\}^2\rceil$, and rigorous gradient threshold $\{n\log(n+1)\}^{-1}$, consistently using G_n to replace F_n inside (14). We explicitly apply (9) over the newly normalized coordinates, repeatedly using G_n. If the check formally rejects every tested source fit, we immediately initialize from a unit leading eigenvector of $n^{-1}\sum_i R_i Z_i Z_i^\top$ and deterministically repeat this exact same normalized target fit. This target restart strictly implements the raw target spectral rule because the normalization inherently enforces $n^{-1}\sum_i R_i = 1$. Our theoretical result fully permits any arbitrary finite measurable candidate list and any target-dependent selection among formally accepted fits.

For an accepted fit $\widehat{v}_n$, report $\widehat{u}_n = T_n\widehat{v}_n$. Let $r_i = Y_i - (X_i^\top\widehat{u}_n)^2$, and compute

$$\widehat{H}_n = \frac{1}{n}\sum_{i=1}^n\{3(X_i^\top\widehat{u}_n)^2 - Y_i\}X_iX_i^\top, \qquad \widehat{J}_n = \frac{1}{n}\sum_{i=1}^n r_i^2(X_i^\top\widehat{u}_n)^2X_iX_i^\top.$$

If no fit is accepted or either matrix is not positive definite, report $\mathbb{R}^p$. When both matrices are positive definite, let $\widehat{V}_n = \widehat{H}_n^{-1}\widehat{J}_n\widehat{H}_n^{-1}$ and report

$$\mathcal{C}_n^\Sigma(\beta) = \left\{u : n\min_{s\in\{-1,1\}}(u - s\widehat{u}_n)^\top\widehat{V}_n^{-1}(u - s\widehat{u}_n) \le \chi^2_{p,1-\beta}\right\}. \tag{12}$$

Use $\mathcal{C}_n^\Sigma(\beta) = \mathbb{R}^p$ in the exceptional cases and assign zero to an undefined estimate and I_p to an undefined covariance estimate. This gives a finite procedure for every sample. All normalization, fitting, and covariance calculations use the original target observations.

The theorem establishes coverage of (12) without treating the empirically normalized rows as independent observations. Let $\widetilde{u}_n$ be the sign of $\widehat{u}_n$ closest to θ^*. This alignment is used only in the analysis.

Theorem 9. *For fixed p, assume the stated Gaussian target model and*

$$\mathbb{E}(\varepsilon \mid X) = 0, \qquad \mathbb{E}|\varepsilon| < \infty, \qquad \mathbb{E}(\varepsilon^2\|X\|^4) < \infty, \qquad J \succ 0.$$

For every finite measurable list of source starts, acceptance has probability tending to one and

$$\sqrt{n}(\widetilde{u}_n - \theta^*) \Rightarrow \mathsf{N}(0, H^{-1}JH^{-1}), \qquad \widehat{V}_n \xrightarrow{\mathbb{P}} H^{-1}JH^{-1}, \qquad \mathbb{P}\{\theta^* \in \mathcal{C}_n^\Sigma(\beta)\} \longrightarrow 1-\beta.$$

The design covariance and signal scale in Theorem 9 need not be known. The noise may be heteroskedastic. No source accuracy or source moment condition is imposed. The covariance uses only target scores. Estimating the normalization does not add a first-order term because the accepted fit approximates a stationary point of the original target criterion. The theorem is pointwise in the fixed population parameters; it is not uniform over matrices approaching singularity or signals approaching zero.

Proof. We write the explicit empirical criterion as $F_n(u) = (4n)^{-1}\sum_i\{(X_i^\top u)^2 - Y_i\}^2$. We first formally analyze this original, unnormalized criterion. For any population calculation, we explicitly subtract the strictly parameter-independent term $(4n)^{-1}\sum_i Y_i^2$ before computing any expectations. The resulting continuous expectation and its higher-order derivatives remain finite under our stated moments, even if the marginal response moment $\mathbb{E}Y^2$ is infinite. Our standard Gaussian fourth-moment calculation and the conditional mean-zero assumption $\mathbb{E}(\varepsilon \mid X) = 0$ rigorously establish that the population Hessian evaluated precisely at θ^* exactly equals H. Because $\Sigma \succ 0$ and $\theta^* \neq 0$, we have $\tau > 0$ and $H \succ 0$. On every bounded compact set, the empirical Hessian functions strictly as a polynomial in u. Its exact coefficients comprise sample averages of fourth-degree Gaussian products and cross-terms of the form $Y_iX_{i,j}X_{i,k}$. These sample coefficients uniformly admit finite absolute expectations because

$$\mathbb{E}(|\varepsilon|\,\|X\|^2) \le \{\mathbb{E}(\varepsilon^2\|X\|^4)\}^{1/2} < \infty.$$

Our previous finite-coefficient argument from (10) thereby guarantees uniform empirical Hessian consistency strictly on every fixed compact set. Because the population Hessian remains continuous

and strictly positive definite precisely at θ^*, we can identify a fixed local neighborhood uniformly exhibiting positive curvature. The corresponding statistical score clearly satisfies

$$\nabla F_n(\theta^*) = -\frac{1}{n}\sum_i \varepsilon_i (X_i^\top \theta^*) X_i, \qquad \mathbb{E}\left\|\varepsilon (X^\top \theta^*) X\right\|^2 \le \|\theta^*\|^2 \, \mathbb{E}(\varepsilon^2 \|X\|^4) < \infty.$$

Applying the argument of Theorem 5 on a sufficiently small fixed ball gives a unique local stationary point $\widehat{\theta}_n$ with probability tending to one. Combining the integrated Hessian equation with the standard multivariate central limit theorem immediately produces

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) = -H^{-1}\sqrt{n}\nabla F_n(\theta^*) + o_{\mathbb{P}}(1) \Rightarrow \mathsf{N}(0, H^{-1} J H^{-1}). \tag{13}$$

We define this root completely arbitrarily on the rare asymptotic event where it fails to exist.

The empirical sample covariance correctly converges almost surely to Σ. The Gaussian model and $\mathbb{E}|\varepsilon| < \infty$ imply $\mathbb{E}|Y| < \infty$. Thus the empirical sample mean response converges almost surely to τ, and our normalization is defined with probability tending to one. This gives

$$T_n \longrightarrow T = \sqrt{\tau}\, \Sigma^{-1/2}, \qquad \bar{\theta} = T^{-1}\theta^*, \qquad \left\|\bar{\theta}\right\| = 1.$$

The algorithmically normalized objective strictly obeys the exact coordinate identities

$$G_n(v) = \widehat{\tau}_n^{-2} F_n(T_n v), \qquad \nabla G_n(v) = \widehat{\tau}_n^{-2} T_n^\top \nabla F_n(T_n v).$$

The derivatives of G_n converge uniformly on $\|v\| \le 2$ to the derivatives of $Q(v) = 3\|v\|^4/4 - \|v\|^2/2 - (v^\top \bar{\theta})^2$. To rigorously verify this limit, we can algebraically express each individual normalized coefficient precisely as a finite product combining $\widehat{\tau}_n^{-1}$, explicit entries of $\widehat{\Sigma}_n^{-1/2}$, and exactly one original empirical coefficient whose convergence we already established. This expansion strictly yields finitely many products, while our theoretical limiting normalization matrix remains unambiguously nonsingular. These sample limits perfectly match the exact Gaussian population coefficients describing Q, because the strictly conditional mean-zero noise functionally contributes zero to the limiting expectation. Our specific proof safely permits every normalized row to statistically depend on the exact same empirical sample covariance and identical mean response.

Lemma 3 and our preceding argument justifying (11) concurrently prove that, with probability asymptotically tending to one, every single accepted normalized fit rigorously falls within a margin $2/\{n\log(n+1)\}$ of the strictly unique normalized local root, strictly after enforcing correct sign alignment. The explicitly transformed original root $T_n^{-1}\widehat{\theta}_n$ successfully converges to $\bar{\theta}$ and remains exactly stationary for the normalized objective G_n. Uniqueness identifies it with that normalized root. Thus, simultaneously over every accepted fit, we obtain

$$\left\|\widetilde{u}_n - \widehat{\theta}_n\right\| \le \frac{2\|T_n\|_{\mathrm{op}}}{n\log(n+1)} = o_{\mathbb{P}}(n^{-1/2}).$$

We explicitly do not require any convergence rate for $\widehat{\tau}_n - \tau$.

The computed normalized target spectral matrix satisfies

$$\frac{1}{n}\sum_i R_i Z_i Z_i^\top = \widehat{\tau}_n^{-1}\widehat{\Sigma}_n^{-1/2}\left(\frac{1}{n}\sum_i Y_i X_i X_i^\top\right)\widehat{\Sigma}_n^{-1/2} \longrightarrow I_p + 2\bar{\theta}\bar{\theta}^\top.$$

Its corresponding leading eigenvector thereby successfully enters the targeted signed $1/8$-ball with probability rigorously tending to one. The precise local curvature constants, fixed optimization step, and defined algorithmic cap detailed strictly in Supplementary Section K.2 identically guarantee successful acceptance of the target restart with probability tending to one. Our derived equation (13) thus definitively establishes the asserted normal limit for our final reported statistical estimate.

We must finally theoretically justify our empirical sandwich variance estimate computed explicitly at the finalized data-dependent fit. Uniform Hessian consistency and the algorithmic convergence $\widetilde{u}_n \to \theta^*$ directly yield the probability limit $\widehat{H}_n \to H$. We formally define $h_n = \widetilde{u}_n - \theta^*$. On the event $\|h_n\| \le 1$, the following bounds hold for a constant depending only on θ^*:

$$|(X_i^\top \widetilde{u}_n)^2 - (X_i^\top \theta^*)^2| \le C\|h_n\|\|X_i\|^2, \qquad |X_i^\top \widetilde{u}_n| \le C\|X_i\|.$$

Algebraically expanding the squared empirical residual inside $\widehat{J}_n$ directly gives

$$\left\| \widehat{J}_n - \frac{1}{n} \sum_i \varepsilon_i^2 (X_i^\top \theta^*)^2 X_i X_i^\top \right\|_{\mathrm{op}}$$
$$\leq C \left\| h_n \right\| \frac{1}{n} \sum_i \{ \varepsilon_i^2 \left\| X_i \right\|^4 + |\varepsilon_i| \left\| X_i \right\|^6 \} + C \left\| h_n \right\|^2 \frac{1}{n} \sum_i \left\| X_i \right\|^8 .$$

Our weighted second-moment condition makes the first sample average bounded in probability. Concurrently, the standard Cauchy–Schwarz inequality bounds the cross-term via

$$\mathbb{E}(|\varepsilon| \left\| X \right\|^6) \leq \{ \mathbb{E}(\varepsilon^2 \left\| X \right\|^4) \}^{1/2} \{ \mathbb{E} \left\| X \right\|^8 \}^{1/2} < \infty.$$

This bound tends to zero in probability. The matrix average converges to J by the law of large numbers, which gives $\widehat{J}_n \xrightarrow{\mathbb{P}} J$. These limits and positive definiteness of H and J give sandwich consistency and make the probability of exceptional covariance cases tend to zero. Statistical studentization of the derived asymptotic normal limit thereby proves proper chi-square convergence. We rigorously separate the opposite, incorrect sign from the true target by exactly $2 \left\| \theta^* \right\| + o_{\mathbb{P}}(1)$, ensuring its corresponding empirical quadratic statistic reliably diverges to infinity. Taking the minimum of the two sign-specific quadratic forms and using continuity of the chi-square distribution establishes our coverage conclusion. This completes the proof. □

The specified weighted moment condition can validly hold even when the marginal target noise variance is infinite. To construct an explicit example, let $p = 1$, $X \sim \mathsf{N}(0, 1)$, $\theta^* \neq 0$, and define ζ as an independent fair random sign. We set

$$\varepsilon = \zeta |X|^{-3/4} \mathbf{1}\{0 < |X| \leq 1\}.$$

The underlying standard Gaussian density remains strictly bounded from above and strictly away from zero inside any small neighborhood around the origin. Direct integration there yields exactly $\mathbb{E}|\varepsilon| < \infty$, $\mathbb{E}\varepsilon^2 = \infty$, and $0 < \mathbb{E}(\varepsilon^2 X^4) < \infty$. The corresponding conditional mean strictly equals zero, while $J = (\theta^*)^2 \mathbb{E}(\varepsilon^2 X^4) > 0$. Thus, Theorem 9 applies. This example satisfies the conditions for the sandwich procedure in this appendix. Its noise depends on the design and has infinite variance. It does not satisfy the conditions of Proposition 19.

L Quotient Metrics and Nonidentifiability

Many nonconvex statistical models have nonidentifiability. In rank-one phase retrieval, θ and $-\theta$ define the same signal. In low-rank factor models, $UV^\top$ is unchanged under rotations of the factors. The quotient extension is not automatic for an arbitrary group action. Suppose $\mathcal{G}$ acts by Euclidean isometries and a smooth local slice selects a unique representative near the target equivalence class. Define

$$d_{\mathcal{G}}(\theta, \theta') = \inf_{g \in \mathcal{G}} \left\| \theta - g\theta' \right\| .$$

The basin event becomes $d_{\mathcal{G}}(\widehat{\theta}_{\mathsf{S}}, \theta^*_{\mathsf{T}}) < r$. Equivariance of the target iteration and curvature orthogonal to the equivalence orbit do not alone identify the update in slice coordinates. The coordinate representation of the iteration must also be analyzed. The contraction bound in Remark 1 gives an optimization conclusion when that bound and invariance of the local region have been proved. A first-order conclusion further requires the expansion of the target-sample fixed point.

For phase retrieval, the group is $\{I_p, -I_p\}$ and $d_{\mathcal{G}}$ is the signed metric $d_{\pm}$ of Section I.1. The reported errors and success indicators in Section T use this metric, and a source initializer is counted as successful even when its sign differs from the target representative. Section O verifies the sign-aligned argument directly, without relying on a general quotient claim.

M Empirical-Process Sufficient Conditions

Directional tail bounds for a smooth loss control the target score and the local empirical Hessian. The following results derive (5); the population curvature in Condition 2 must be verified for the chosen model. Comparable directional score and Hessian tail assumptions are used by Mei, Bai and Montanari (2018) to control empirical derivatives.

Condition 3 (Smooth finite-dimensional loss)**.** *For almost every Z_T, $\ell_\mathsf{T}(\cdot; Z_\mathsf{T})$ is twice continuously differentiable on $\mathbb{B}_R$. There are constants $K_1, K_2, K_3 < \infty$ such that*

$$\sup_{\|v\|=1} \left\| v^\top \{\nabla \ell_\mathsf{T}(\theta_\mathsf{T}^*; Z_\mathsf{T}) - \nabla L_\mathsf{T}(\theta_\mathsf{T}^*)\} \right\|_{\psi_2} \le K_1,$$

and, for all $\theta \in \mathbb{B}_R$,

$$\sup_{\|v\|=\|w\|=1} \left\| v^\top \{\nabla^2 \ell_\mathsf{T}(\theta; Z_\mathsf{T}) - \nabla^2 L_\mathsf{T}(\theta)\} w \right\|_{\psi_1} \le K_2,$$

and, for all $\theta, \theta' \in \mathbb{B}_R$,

$$\left\| \nabla^2 \ell_\mathsf{T}(\theta; Z_\mathsf{T}) - \nabla^2 \ell_\mathsf{T}(\theta'; Z_\mathsf{T}) \right\|_{\mathrm{op}} \le K_3 \|\theta - \theta'\|$$

almost surely.

Condition 3 places a directional tail bound on the score at the target parameter, a uniform directional tail bound on the local Hessian, and an almost-sure Lipschitz bound on the sample Hessian. The score condition is imposed only at θ_T^* because Theorem 5 uses it only there. These bounds also justify differentiation under expectation. In fact,

$$\sup_{\theta \in \mathbb{B}_R} \left\| \nabla^2 \ell_\mathsf{T}(\theta; Z_\mathsf{T}) \right\|_{\mathrm{op}} \le \left\| \nabla^2 \ell_\mathsf{T}(\theta_\mathsf{T}^*; Z_\mathsf{T}) \right\|_{\mathrm{op}} + K_3 R,$$

and

$$\sup_{\theta \in \mathbb{B}_R} \|\nabla \ell_\mathsf{T}(\theta; Z_\mathsf{T})\| \le \|\nabla \ell_\mathsf{T}(\theta_\mathsf{T}^*; Z_\mathsf{T})\| + R\{\left\| \nabla^2 \ell_\mathsf{T}(\theta_\mathsf{T}^*; Z_\mathsf{T}) \right\|_{\mathrm{op}} + K_3 R\}.$$

The right sides have finite expectations under the directional tail bounds. Dominated difference quotients give $\mathbb{E}\nabla \ell_\mathsf{T}(\theta; Z_\mathsf{T}) = \nabla L_\mathsf{T}(\theta)$ and $\mathbb{E}\nabla^2 \ell_\mathsf{T}(\theta; Z_\mathsf{T}) = \nabla^2 L_\mathsf{T}(\theta)$. An ε-net of a set is a finite subset such that every point of the set lies within Euclidean distance ε of some point of the subset. In Lemmas 4 and 5 and Proposition 11, C denotes a sufficiently large numerical constant. The next two lemmas derive the empirical bounds used by the target theorem.

The following lemma controls the target score at the population parameter.

Lemma 4. *Under Condition 3, for every $u \ge 1$,*

$$\mathbb{P}\left\{ \|\nabla L_{\mathsf{T},n_\mathsf{T}}(\theta_\mathsf{T}^*) - \nabla L_\mathsf{T}(\theta_\mathsf{T}^*)\| > C K_1 \sqrt{\frac{p+u}{n_\mathsf{T}}} \right\} \le e^{-u}.$$

Lemma 4 gives the tolerance τ_{n_T} in (5). Uniform Hessian control requires a second net over the local parameter ball.

Proof. For a unit vector v,

$$v^\top \{\nabla \ell_\mathsf{T}(\theta_\mathsf{T}^*; Z_\mathsf{T}) - \nabla L_\mathsf{T}(\theta_\mathsf{T}^*)\}$$

has sub-Gaussian norm bounded by a multiple of K_1. For a $1/2$-net $\mathcal{N}$ of the unit sphere in $\mathbb{R}^p$,

$$\|\nabla L_{\mathsf{T},n_\mathsf{T}}(\theta_\mathsf{T}^*) - \nabla L_\mathsf{T}(\theta_\mathsf{T}^*)\| \le 2 \sup_{v \in \mathcal{N}} v^\top \{\nabla L_{\mathsf{T},n_\mathsf{T}}(\theta_\mathsf{T}^*) - \nabla L_\mathsf{T}(\theta_\mathsf{T}^*)\}.$$

Choose at most 5^p net points (Vershynin, 2018, Corollary 4.2.13). Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.2) and a union bound complete the proof. □

The following lemma establishes that uniform control.

Lemma 5. *Under Condition 3, for every $u \ge 1$ and $n_\mathsf{T} \ge 2$, with probability at least $1 - e^{-u}$,*

$$\sup_{\theta \in \mathbb{B}_R} \left\| \nabla^2 L_{\mathsf{T},n_\mathsf{T}}(\theta) - \nabla^2 L_\mathsf{T}(\theta) \right\|_{\mathrm{op}} \le C \left[K_2 \left(\sqrt{\frac{p \log n_\mathsf{T} + u}{n_\mathsf{T}}} + \frac{p \log n_\mathsf{T} + u}{n_\mathsf{T}} \right) + \frac{K_3 R}{\sqrt{n_\mathsf{T}}} \right].$$

Lemma 5 gives Δ_{n_T} in (5). The following proposition records the joint probability after a union bound.

Proof. Let $\mathcal{N}_\theta$ be an ε-net of $\mathbb{B}_R$ with $\varepsilon = R\, n_{\mathsf{T}}^{-1/2}$ and let $\mathcal{N}_v$ be a $1/4$-net of the unit sphere, with cardinalities $|\mathcal{N}_\theta| \le (3R/\varepsilon)^p = (3n_{\mathsf{T}}^{1/2})^p$ and $|\mathcal{N}_v| \le 9^p$. Write $D(\theta) = \nabla^2 L_{\mathsf{T},n_{\mathsf{T}}}(\theta) - \nabla^2 L_{\mathsf{T}}(\theta)$. For $\theta \in \mathcal{N}_\theta$ and $v, w \in \mathcal{N}_v$, $v^\top D(\theta) w$ averages independent centered terms with ψ_1-norms at most $2K_2$. Bernstein's inequality (Vershynin, 2018, Theorem 2.8.1) gives, for $s \ge 0$,

$$\mathbb{P}\left\{ |v^\top D(\theta) w| > C K_2 \left(\sqrt{\frac{s}{n_{\mathsf{T}}}} + \frac{s}{n_{\mathsf{T}}} \right) \right\} \le 2e^{-s}.$$

Take $s = 3p\log(9n_{\mathsf{T}}) + u$ and a union bound over $|\mathcal{N}_\theta|\,|\mathcal{N}_v|^2 \le (9^2 \cdot 3)^p n_{\mathsf{T}}^{p/2}$ triples. The inequality $\|D(\theta)\|_{\mathrm{op}} \le 2\max_{v,w\in\mathcal{N}_v} |v^\top D(\theta) w|$ extends the bound beyond $\mathcal{N}_v$. For $\theta \in \mathbb{B}_R$, choose $\theta' \in \mathcal{N}_\theta$ with $\|\theta - \theta'\| \le \varepsilon$. Condition 3 bounds both empirical and population Hessian increments, giving $\|D(\theta) - D(\theta')\|_{\mathrm{op}} \le 2K_3\varepsilon = 2K_3 R\, n_{\mathsf{T}}^{-1/2}$. This completes the proof. □

Proposition 11. *Assume Condition 3 and $\nabla L_{\mathsf{T}}(\theta_{\mathsf{T}}^*) = 0$. Fix $u \ge 1$ and $n_{\mathsf{T}} \ge 2$. If*

$$C K_1 \sqrt{\frac{p+u}{n_{\mathsf{T}}}} \le \tau_{n_{\mathsf{T}}},$$

and

$$C\left[K_2 \left(\sqrt{\frac{p\log n_{\mathsf{T}} + u}{n_{\mathsf{T}}}} + \frac{p\log n_{\mathsf{T}} + u}{n_{\mathsf{T}}} \right) + \frac{K_3 R}{\sqrt{n_{\mathsf{T}}}} \right] \le \Delta_{n_{\mathsf{T}}},$$

then (5) *holds with $\eta_{n_{\mathsf{T}}} = 2e^{-u}$.*

Proposition 11 controls both stochastic errors with probability at least $1 - 2e^{-u}$. The score bound determines the local estimation error, while the Hessian bound preserves the population curvature on the target sample.

Proof. Lemmas 4 and 5 give the stated $\tau_{n_{\mathsf{T}}}$ and $\Delta_{n_{\mathsf{T}}}$ tolerances. A union bound completes the proof. □

M.1 Finite-Moment Conditions for the Target Limit

Local Hessian consistency suffices for the limiting distribution. We derive it from population moment conditions. The estimating-equation argument is standard in smooth M-estimation (van der Vaart, 1998, Chapter 5); we also verify uniform convergence of the target iterates. Fix p and $R_1 > 0$ with $\mathbb{B}_{R_1} \subset \Theta$. Suppose that the loss is twice continuously differentiable on a neighborhood of this ball almost surely and that $\mathbb{E}|\ell_{\mathsf{T}}(\theta_{\mathsf{T}}^*; Z_{\mathsf{T}})| < \infty$. Write $g(z) = \nabla \ell_{\mathsf{T}}(\theta_{\mathsf{T}}^*; z)$ and $\Sigma = \mathbb{E}\{g(Z_{\mathsf{T}}) g(Z_{\mathsf{T}})^\top\}$, and write $n = n_{\mathsf{T}}$.

Proposition 12. *Assume*

$$\mathbb{E} g(Z_{\mathsf{T}}) = 0, \qquad \mathbb{E}\,\|g(Z_{\mathsf{T}})\|^2 < \infty, \qquad \mathbb{E} \sup_{\theta \in \mathbb{B}_{R_1}} \left\| \nabla^2 \ell_{\mathsf{T}}(\theta; Z_{\mathsf{T}}) \right\|_{\mathrm{op}} < \infty,$$

and $\mathbb{E}\nabla^2 \ell_{\mathsf{T}}(\theta_{\mathsf{T}}^*; Z_{\mathsf{T}}) \succ 0$. For every $0 < \alpha < 2/\lambda_{\max}(H)$, there are $0 < r < R < R_1$ and $0 < \kappa < 1$ such that

$$\mathbb{P}\{u_n \in \mathbb{B}_r\} \to 1, \quad \sqrt{n}\,\kappa^{t_n} \to 0 \quad \Longrightarrow \quad \sqrt{n}\{\theta^{t_n}(u_n) - \theta_{\mathsf{T}}^*\} \Rightarrow \mathsf{N}(0, H^{-1}\Sigma H^{-1}).$$

Proposition 12 does not require exponential score tails or a deterministic Lipschitz bound for individual sample Hessians. Positive curvature at the target parameter and an integrable local Hessian bound suffice. The conclusion concerns fixed dimension and does not give the explicit finite-sample remainder of Theorem 5.

Proof. Write $A(\theta, z) = \nabla^2 \ell_{\mathsf{T}}(\theta; z)$ and $G(z) = \sup_{\theta \in \mathbb{B}_{R_1}} \|A(\theta, z)\|_{\mathrm{op}}$. The moment conditions and gradient bound $\|g(Z_{\mathsf{T}})\| + R_1 G(Z_{\mathsf{T}})$ justify $\nabla L_{\mathsf{T}}(\theta_{\mathsf{T}}^*) = 0$ and $\nabla^2 L_{\mathsf{T}}(\theta) = \mathbb{E} A(\theta, Z_{\mathsf{T}})$. Dominated convergence gives population Hessian continuity; H is positive definite by assumption.

For $\delta > 0$, put

$$\omega_\delta(z) = \sup_{\substack{\theta,\theta' \in \mathbb{B}_{R_1} \\ \|\theta-\theta'\| \le \delta}} \|A(\theta, z) - A(\theta', z)\|_{\mathrm{op}}.$$

Uniform continuity on the compact ball and $\omega_\delta \le 2G$ give $\mathbb{E}\omega_\delta(Z_\mathsf{T}) \to 0$ as $\delta \downarrow 0$. Let P_n and P denote the empirical and population averages, and let $\mathcal{N}_\delta$ be a finite δ-net of the ball. Then

$$\sup_{\theta \in \mathbb{B}_{R_1}} \|(P_n - P)A(\theta, \cdot)\|_{\mathrm{op}} \le \max_{v \in \mathcal{N}_\delta} \|(P_n - P)A(v, \cdot)\|_{\mathrm{op}} + P_n\omega_\delta + P\omega_\delta.$$

Almost-sure convergence of the net entries and ω_δ bounds the limit superior by $2P\omega_\delta$ almost surely. Take $\delta = 1/j$, $j \ge 1$, and then $j \to \infty$ for uniform Hessian convergence. The central limit theorem gives $\sqrt{n}S_n \Rightarrow \mathsf{N}(0, \Sigma)$.

Choose $0 < m < \lambda_{\min}(H)$ and $\lambda_{\max}(H) < L < 2/\alpha$. Population Hessian continuity gives $0 < R < R_1$ such that its eigenvalues on $\mathbb{B}_R$ lie in $[m, L]$. Set $r = R/2$. Choose deterministic $\tau_n, \Delta_n, \eta_n \to 0$ sufficiently slowly to obtain (5). For large n, $\Delta_n \le m/2$ and $r + 4\tau_n/m \le R$. The factor κ_n converges to $\max\{|1 - \alpha m|, |1 - \alpha L|\} < 1$. Choose a fixed $\kappa < 1$ exceeding this limit. Then $\sqrt{n}\,\kappa_n^{t_n} r \to 0$, and Corollary 4 proves the result. This completes the proof. □

N SOURCE-TAIL CALIBRATION

The localization bound in (6) concerns estimation around the source parameter, independently of the shift to the target. For a source estimator with an asymptotically linear representation, write

$$\widehat{\theta}_\mathsf{S} - \theta^*_\mathsf{S} = \frac{1}{n_\mathsf{S}} \sum_{i=1}^{n_\mathsf{S}} \psi_\mathsf{S}(Z_{\mathsf{S},i}) + \rho_{n_\mathsf{S}},$$

where ψ_S is the influence function and ρ_{n_S} is the remainder. Let $0 \le K_\mathsf{S} < \infty$ and $b_\mathsf{S}(s) \ge 0$ be deterministic bounds. The following proposition controls localization by the fluctuation of the influence function and the remainder.

Proposition 13. *Fix $s \ge 1$. Assume $\mathbb{E}\psi_\mathsf{S}(Z_\mathsf{S}) = 0$ and*

$$\sup_{\|v\|=1} \left\|v^\top \psi_\mathsf{S}(Z_\mathsf{S})\right\|_{\psi_2} \le K_\mathsf{S}, \qquad \mathbb{P}\{\|\rho_{n_\mathsf{S}}\| > b_\mathsf{S}(s)\} \le e^{-s}.$$

Then, for a numerical constant $C < \infty$,

$$\mathbb{P}\left\{\left\|\widehat{\theta}_\mathsf{S} - \theta^*_\mathsf{S}\right\| > CK_\mathsf{S}\sqrt{\frac{p+s}{n_\mathsf{S}}} + b_\mathsf{S}(s)\right\} \le 2e^{-s}.$$

If the remainder bound holds for every $s \ge 1$, Proposition 13 verifies (6) with

$$a_\mathsf{S}(s) = CK_\mathsf{S}\sqrt{\frac{p + s + \log 2}{n_\mathsf{S}}} + b_\mathsf{S}(s + \log 2).$$

The sufficient source-success condition in Proposition 3 becomes

$$\delta + CK_\mathsf{S}\sqrt{\frac{p + s + \log 2}{n_\mathsf{S}}} + b_\mathsf{S}(s + \log 2) \le r.$$

The source estimation error and the population shift must fit within the target radius.

Proof. For any fixed unit vector v,

$$v^\top n_\mathsf{S}^{-1} \sum_{i=1}^{n_\mathsf{S}} \psi_\mathsf{S}(Z_{\mathsf{S},i})$$

is sub-Gaussian with norm bounded by $CK_\mathsf{S} n_\mathsf{S}^{-1/2}$. Let $\mathcal{N}$ be a $1/2$-net of the unit sphere with $|\mathcal{N}| \le 5^p$. Then

$$\left\|\frac{1}{n_\mathsf{S}} \sum_{i=1}^{n_\mathsf{S}} \psi_\mathsf{S}(Z_{\mathsf{S},i})\right\| \le 2 \sup_{v \in \mathcal{N}} v^\top \frac{1}{n_\mathsf{S}} \sum_{i=1}^{n_\mathsf{S}} \psi_\mathsf{S}(Z_{\mathsf{S},i}).$$

A union bound over $\mathcal{N}$ gives

$$\mathbb{P}\left\{\left\|\frac{1}{n_{\mathsf{S}}}\sum_{i=1}^{n_{\mathsf{S}}}\psi_{\mathsf{S}}(Z_{\mathsf{S},i})\right\| > CK_{\mathsf{S}}\sqrt{\frac{p+s}{n_{\mathsf{S}}}}\right\} \le e^{-s}.$$

Combining this event with $\{\|\rho_{n_{\mathsf{S}}}\| \le b_{\mathsf{S}}(s)\}$ gives the claim. □

For a model identified up to a group of Euclidean isometries $\mathcal{G}$, let

$$d_{\mathcal{G}}(\theta,\theta') = \inf_{g\in\mathcal{G}} \|\theta - g\theta'\|.$$

Let c be a local chart defined on $\{\theta : d_{\mathcal{G}}(\theta,\theta^*_{\mathsf{T}}) \le R\}$. The following proposition gives the corresponding entry bound when chart distance from the target equals quotient distance.

Proposition 14. *Assume $r \le R$ and*

$$\|c(\theta) - c(\theta^*_{\mathsf{T}})\| = d_{\mathcal{G}}(\theta,\theta^*_{\mathsf{T}})$$

on this set. If

$$\mathbb{P}\{d_{\mathcal{G}}(\widehat{\theta}_{\mathsf{S}},\theta^*_{\mathsf{S}}) > a_{\mathsf{S}}(s)\} \le e^{-s}, \qquad d_{\mathcal{G}}(\theta^*_{\mathsf{S}},\theta^*_{\mathsf{T}}) + a_{\mathsf{S}}(s) \le r,$$

then the source initializer lies in the target chart of radius r with probability at least $1 - e^{-s}$.

Proposition 14 concerns entry into the stated chart, not contraction of the target iteration within it. For phase retrieval, sign alignment satisfies the chart identity. Rotation-invariant models require their own chart and iteration analysis.

Proof. On the source event,

$$d_{\mathcal{G}}(\widehat{\theta}_{\mathsf{S}},\theta^*_{\mathsf{T}}) \le d_{\mathcal{G}}(\widehat{\theta}_{\mathsf{S}},\theta^*_{\mathsf{S}}) + d_{\mathcal{G}}(\theta^*_{\mathsf{S}},\theta^*_{\mathsf{T}}) \le r.$$

The chart identity then gives $\left\|c(\widehat{\theta}_{\mathsf{S}}) - c(\theta^*_{\mathsf{T}})\right\| \le r$. □

O PHASE-RETRIEVAL VERIFICATION

The quartic phase-retrieval loss permits explicit bounds for target curvature, empirical derivatives, and source spectral localization. The validation and Gaussian risk benchmarks are separate results and are not claimed as consequences of the phase-retrieval calculation.

Let $X \sim \mathsf{N}(0, I_p)$, let $\theta^* \in \mathbb{R}^p$, and define

$$Y = (X^\top\theta^*)^2 + \varepsilon, \qquad \mathbb{E}(\varepsilon \mid X) = 0.$$

For the curvature calculation, set $\varepsilon = 0$ and consider

$$F(u) = \frac{1}{4}\mathbb{E}\left[\{(X^\top u)^2 - Y\}^2\right].$$

Mean-zero noise with a finite second moment adds $\mathbb{E}(\varepsilon^2)/4$ to the population risk and does not change its derivatives.

The following moment identities determine the population risk and Hessian.

Lemma 6. *For any $a, b \in \mathbb{R}^p$,*

$$\mathbb{E}(X^\top a)^4 = 3\|a\|^4,$$

and

$$\mathbb{E}\{(X^\top a)^2(X^\top b)^2\} = \|a\|^2\|b\|^2 + 2(a^\top b)^2.$$

Lemma 6 reduces the quartic population criterion to ordinary inner products. The following proposition gives the resulting derivatives.

Proof. The first identity is the fourth moment of a centered normal random variable with variance $\|a\|^2$. For the second, Isserlis' formula gives

$$\mathbb{E}(X_iX_jX_kX_l) = \mathbf{1}\{i=j\}\mathbf{1}\{k=l\} + \mathbf{1}\{i=k\}\mathbf{1}\{j=l\} + \mathbf{1}\{i=l\}\mathbf{1}\{j=k\}.$$

Multiplying by $a_ia_jb_kb_l$ and summing over all indices yields the three terms $\|a\|^2\|b\|^2$, $(a^\top b)^2$, and $(a^\top b)^2$. □

Proposition 15. *For the noiseless part of the phase-retrieval population objective,*

$$F(u) = \frac{1}{4}\left[3\|u\|^4 - 2\|u\|^2\|\theta^*\|^2 - 4(u^\top\theta^*)^2 + 3\|\theta^*\|^4\right].$$

In addition,

$$\nabla F(u) = 3\|u\|^2 u - \|\theta^*\|^2 u - 2(u^\top\theta^*)\theta^*,$$

and

$$\nabla^2 F(u) = 6uu^\top + (3\|u\|^2 - \|\theta^*\|^2)I_p - 2\theta^*(\theta^*)^\top.$$

For a nonzero signal, Proposition 15 gives a stationary point with positive curvature. The next proposition gives a certified radius and explicit constants for a unit-norm signal.

Proof. Expanding the square gives

$$4F(u) = \mathbb{E}(X^\top u)^4 - 2\mathbb{E}\{(X^\top u)^2(X^\top\theta^*)^2\} + \mathbb{E}(X^\top\theta^*)^4.$$

Lemma 6 gives the formula for $F(u)$. Differentiating that expression with respect to u gives the gradient and Hessian. □

Proposition 16. *Assume $\|\theta^*\| = 1$. For all u, u' with $\|u-\theta^*\| \le 1/20$ and $\|u'-\theta^*\| \le 1/20$,*

$$I_p \preceq \nabla^2 F(u) \preceq 7I_p, \qquad \left\|\nabla^2 F(u) - \nabla^2 F(u')\right\|_{\mathrm{op}} \le 19\|u-u'\|.$$

The population objective satisfies the curvature and Hessian-Lipschitz parts of Condition 2 with $R = 1/20$, $m = 1$, $L = 7$, and $M = 19$.

Proposition 16 verifies the population part of Condition 2 on the sign-aligned ball of radius $1/20$. We write $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$.

Proof. At $u = \theta^*$, Proposition 15 gives

$$\nabla^2 F(\theta^*) = 4\theta^*(\theta^*)^\top + 2I_p.$$

The eigenvalue in the direction θ^* is 6, and every eigenvalue orthogonal to θ^* is 2. This gives

$$2I_p \preceq \nabla^2 F(\theta^*) \preceq 6I_p.$$

The Hessian formula in Proposition 15 gives, for any u, u',

$$\nabla^2 F(u) - \nabla^2 F(u') = 6(uu^\top - u'u'^\top) + 3(\|u\|^2 - \|u'\|^2)I_p.$$

Because $\left\|uu^\top - u'u'^\top\right\|_{\mathrm{op}} \le (\|u\|+\|u'\|)\|u-u'\|$ and $|\,\|u\|^2 - \|u'\|^2\,| \le (\|u\|+\|u'\|)\|u-u'\|$,

$$\left\|\nabla^2 F(u) - \nabla^2 F(u')\right\|_{\mathrm{op}} \le 9(\|u\|+\|u'\|)\|u-u'\| \le 9\cdot\frac{42}{20}\|u-u'\| \le 19\|u-u'\|$$

on the stated ball, where $\|u\| \vee \|u'\| \le 21/20$. Taking $u' = \theta^*$ gives

$$\left\|\nabla^2 F(u) - \nabla^2 F(\theta^*)\right\|_{\mathrm{op}} \le 9\cdot\frac{41}{20}\cdot\frac{1}{20} = \frac{369}{400} < 1,$$

which implies $I_p \preceq \nabla^2 F(u) \preceq 7I_p$ on the ball. This completes the proof. □

A direct lower bound for the Hessian permits a larger target ball without changing the target iteration. The following proposition establishes the required curvature bounds by a directional calculation.

Proposition 17. *Assume* $\|\theta^*\| = 1$. *For* $\|u - \theta^*\| \vee \|u' - \theta^*\| \le 1/6$,

$$I_p \preceq \nabla^2 F(u) \preceq \frac{37}{4} I_p, \qquad \left\|\nabla^2 F(u) - \nabla^2 F(u')\right\|_{\mathrm{op}} \le 21 \|u - u'\|.$$

The population bounds in Proposition 17 and the concentration bound in Lemma 9 yield the entry radius $1/8$ in Theorem 7. The remaining distance to the boundary accommodates the error of the local empirical estimator.

Proof. Write $u = \theta^* + h$ and fix $\|v\| = 1$. Put $t = v^\top \theta^*$ and $z = (1+8t^2)^{1/2} \in [1, 3]$. The Hessian in Proposition 15 gives

$$v^\top \nabla^2 F(u) v = 2 + 4t^2 + 6(\theta^* + 2tv)^\top h + 3\|h\|^2 + 6(v^\top h)^2.$$

Here $\|\theta^* + 2tv\| = z$. For $0 < R \le 1/6$ and $\|h\| \le R$, the function $x \mapsto -6zx + 3x^2$ is decreasing on $[0, R]$. Thus

$$v^\top \nabla^2 F(u) v \ge \frac{3}{2} + \frac{z^2}{2} - 6Rz + 3R^2 \ge 2 - 6R + 3R^2.$$

The last inequality follows because $z \mapsto z^2/2 - 6Rz$ is increasing for $z \ge 1$. At $R = 1/6$ this bound equals $13/12 > 1$. The Hessian difference formula in the proof of Proposition 16 gives

$$\left\|\nabla^2 F(u) - \nabla^2 F(\theta^*)\right\|_{\mathrm{op}} \le 9(2 + R)R \le \frac{13}{4}.$$

Because $\nabla^2 F(\theta^*) \preceq 6 I_p$, the upper bound is $37 I_p/4$. Finally, $\|u\| \vee \|u'\| \le 7/6$ implies

$$\left\|\nabla^2 F(u) - \nabla^2 F(u')\right\|_{\mathrm{op}} \le 9(\|u\| + \|u'\|) \|u - u'\| \le 21 \|u - u'\|.$$

This completes the proof. □

The source spectral estimator uses the leading direction of the empirical second moment weighted by the responses and estimates its scale from their mean. For $n \ge 2$ independent observations $Y_i = (X_i^\top \theta^*)^2 + \varepsilon_i$, define

$$\widehat{M} = \frac{1}{n} \sum_{i=1}^n Y_i X_i X_i^\top, \qquad \overline{Y} = \frac{1}{n} \sum_{i=1}^n Y_i,$$

let $\widehat{v}$ be a unit leading eigenvector of $\widehat{M}$, and set

$$\widehat{\theta}_{\mathrm{sp}} = \{\max(\overline{Y}, 0)\}^{1/2} \widehat{v}.$$

The following proposition verifies the source localization needed in Proposition 3.

Proposition 18. *Assume* $X_i \sim \mathsf{N}(0, I_p)$, $\|\theta^*\| = 1$, *and* ε_i *independent of* X_i *with* $\mathbb{E}\varepsilon_i = 0$ *and* $\|\varepsilon_i\|_{\psi_2} \le \sigma$. *For* $s \ge 1$, *there is a constant* $C_\sigma < \infty$, *depending only on* σ, *such that, whenever* $C_\sigma b_{n,p}(s) \le 1/4$,

$$\mathbb{P}\{d_\pm(\widehat{\theta}_{\mathrm{sp}}, \theta^*) > C_\sigma b_{n,p}(s)\} \le 6e^{-s}.$$

Proposition 18 gives $a_{\mathsf{S}}(s) = C_\sigma b_{n_{\mathsf{S}},p}(s + \log 6)$ for a unit-norm source signal when this bound is at most $1/4$. Thus, a fixed signed source-target shift below the certified radius gives basin entry once the source sample makes this bound smaller than the remaining radius.

Proof. The Gaussian fourth-moment identity gives

$$\mathbb{E}\widehat{M} = I_p + 2\theta^*(\theta^*)^\top, \qquad \mathbb{E}\overline{Y} = 1.$$

For $p \ge 2$, the eigengap is 2. Put $\ell = s + \log(8n)$ and truncate observation i on

$$A_i = \{|X_i^\top \theta^*| \le C\sqrt{\ell},\ |\varepsilon_i| \le C\sigma\sqrt{\ell}\}.$$

Choose the constants such that $\mathbb{P}(A_i^c) \le 4e^{-\ell}$ and $\mathbb{P}(\cap_i A_i) \ge 1 - e^{-s}/2$. On A_i, $|Y_i| \le C_\sigma \ell$. For a unit vector v, put $T_i(v) = Y_i (v^\top X_i)^2 \mathbf{1}_{A_i}$. For every integer $k \ge 2$, Cauchy–Schwarz gives

$$\mathbb{E}|T_i(v)|^k \le (C_\sigma \ell)^{k-2} (\mathbb{E}Y_i^4)^{1/2} (\mathbb{E}|v^\top X_i|^{4k})^{1/2} \le k! C_\sigma^k \ell^{k-2}.$$

Here $\mathbb{E}Y_i^4 \le C_\sigma$, and $\{\mathbb{E}|v^\top X_i|^{4k}\}^{1/2} \le (Ck)^k \le k!C^k$. By $\mathbb{E}|T_i(v) - \mathbb{E}T_i(v)|^k \le 2^k\mathbb{E}|T_i(v)|^k$, the exponential series gives

$$\log \mathbb{E}\exp\{\lambda(T_i(v) - \mathbb{E}T_i(v))\} \le \frac{C_\sigma\lambda^2}{1 - C_\sigma\ell|\lambda|}, \qquad |\lambda| < (C_\sigma\ell)^{-1}.$$

Independence and Chernoff's inequality imply, for every $x \ge 1$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_i\{T_i(v) - \mathbb{E}T_i(v)\}\right| > C_\sigma\left(\sqrt{\frac{x}{n}} + \frac{\ell x}{n}\right)\right\} \le 2e^{-x}.$$

Use a $1/4$-net of at most 9^p points, $x = s + p\log 9 + \log 4$, and, for symmetric B, $\|B\|_{\mathrm{op}} \le 2\max_v |v^\top Bv|$. Uniformly over unit v, the truncation bias satisfies

$$|\mathbb{E}\{Y_i(v^\top X_i)^2\mathbf{1}_{A_i^c}\}| \le \{\mathbb{E}Y_i^2(v^\top X_i)^4\}^{1/2}\mathbb{P}(A_i^c)^{1/2} \le C_\sigma e^{-\ell/2}.$$

This bias is at most $C_\sigma\sqrt{(p+s)/n}$, with $\ell \le C(s + \log n)$. On $\cap_i A_i$, truncation changes no observation. Hence, with probability at least $1 - 4e^{-s}$,

$$\left\|\widehat{M} - \mathbb{E}\widehat{M}\right\|_{\mathrm{op}} \le C_\sigma\left\{\sqrt{\frac{p+s}{n}} + \frac{(s+\log n)(p+s)}{n}\right\}.$$

The scalar Bernstein inequality applied to $(X_i^\top\theta^*)^2 - 1 + \varepsilon_i$ also gives

$$|\overline{Y} - 1| \le C_\sigma\left\{\sqrt{\frac{s}{n}} + \frac{s}{n}\right\}$$

with probability at least $1 - 2e^{-s}$.

Choose the sign of $\widehat{v}$ to ensure that $(\widehat{v})^\top\theta^* \ge 0$. For $p = 1$, the signed eigenvector error is zero. For $p \ge 2$, under the stated upper bound on $b_{n,p}(s)$, the Davis–Kahan theorem (Davis and Kahan, 1970) and the eigengap give

$$\|\widehat{v} - \theta^*\| \le C\left\|\widehat{M} - \mathbb{E}\widehat{M}\right\|_{\mathrm{op}}.$$

The same event has $\overline{Y} > 0$, and $|\sqrt{\overline{Y}} - 1| \le |\overline{Y} - 1|$. Hence

$$d_\pm(\widehat{\theta}_{\mathrm{sp}}, \theta^*) \le |\sqrt{\overline{Y}} - 1| + \sqrt{\overline{Y}}\,\|\widehat{v} - \theta^*\| \le C_\sigma b_{n,p}(s).$$

A union bound completes the proof. □

O.1 Fixed-Dimensional Phase Inference under Moment Conditions

The source need not be estimated at the target's rate to preserve the target limiting distribution. Fix the dimension and unit-norm source and target signals. Let the errors be centered and independent of the Gaussian designs, with $\mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty$ and $\mathbb{E}\varepsilon_{\mathsf{T}}^2 < \infty$. Here θ_{S}^* is the source signal defined by $\mathbb{E}(Y_{\mathsf{S}} \mid X_{\mathsf{S}}) = (X_{\mathsf{S}}^\top\theta_{\mathsf{S}}^*)^2$; the source moment estimator does not require a finite expected squared loss. Use the raw source spectral estimator defined above. Let $u_{n_{\mathsf{T}}}$ be its sign representative closest to θ_{T}^*, and let $\theta^t(u_{n_{\mathsf{T}}})$ follow target gradient descent with step $1/8$. This sign choice is only for stating the limit: changing the sign of the initial value changes the sign of every iterate. Write $\nu_{\mathsf{T}}^2 = \mathbb{E}\varepsilon_{\mathsf{T}}^2$ and $H = 2I_p + 4\theta_{\mathsf{T}}^*(\theta_{\mathsf{T}}^*)^\top$.

Corollary 8. *Let p be fixed and $n_{\mathsf{S}}, n_{\mathsf{T}} \to \infty$. If*

$$d_\pm(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*) < \frac{1}{8}, \qquad \sqrt{n_{\mathsf{T}}}\left(\frac{15}{16}\right)^{t_{n_{\mathsf{T}}}} \to 0,$$

then

$$\sqrt{n_{\mathsf{T}}}\{\theta^{t_{n_{\mathsf{T}}}}(u_{n_{\mathsf{T}}}) - \theta_{\mathsf{T}}^*\} \Rightarrow \mathsf{N}\left(0, \frac{\nu_{\mathsf{T}}^2}{2}H^{-1}\right).$$

Corollary 8 imposes no relative growth rate on the source and target sample sizes. A finite first absolute source-noise moment and a strict shift margin give source consistency and entry. Finite target-noise variance gives the limit; the stronger tail assumption in Proposition 18 is used for its explicit finite-sample probability bound.

Proof. Gaussian moments and $\mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty$ make $Y_{\mathsf{S}} X_{\mathsf{S}} X_{\mathsf{S}}^\top$ integrable entrywise. Almost-sure convergence of the sample averages gives

$$\widehat{M}_{\mathsf{S}} \longrightarrow I_p + 2\theta^*_{\mathsf{S}}(\theta^*_{\mathsf{S}})^\top, \qquad \overline{Y}_{\mathsf{S}} \longrightarrow 1 \quad \text{almost surely.}$$

For $p \geq 2$, the population eigengap is 2; for $p = 1$, the signed eigenvector error is zero. The spectral perturbation argument in Proposition 18 gives $d_\pm(\widehat{\theta}_{\mathsf{S}}, \theta^*_{\mathsf{S}}) \to 0$ almost surely. The triangle inequality and the strict shift margin imply $\mathbb{P}\{u_{n_{\mathsf{T}}} \in \mathbb{B}_{1/8}\} \to 1$.

Put $\theta^* = \theta^*_{\mathsf{T}}$, and write the loss of one target observation as $\ell(u; X, Y) = \{(X^\top u)^2 - Y\}^2/4$. The expected loss at the target is $\nu^2_{\mathsf{T}}/4 < \infty$, and its score is $g = -\varepsilon_{\mathsf{T}}(X^\top \theta^*)X$. Independence and Lemma 6 give

$$\mathbb{E}g = 0, \qquad \mathbb{E}(gg^\top) = \nu^2_{\mathsf{T}}\{I_p + 2\theta^*(\theta^*)^\top\} = \frac{\nu^2_{\mathsf{T}}}{2} H.$$

On $\mathbb{B}_{1/6}$, differentiation yields the integrable bound

$$\sup_{u \in \mathbb{B}_{1/6}} \left\| \nabla^2 \ell(u; X, Y) \right\|_{\mathrm{op}} \leq C \left\| X \right\|^4 + |\varepsilon_{\mathsf{T}}| \left\| X \right\|^2.$$

Proposition 17 uses only centered independent noise and bounds the Hessian between I_p and $37I_p/4$. Apply the finite-net proof of Proposition 12 for uniform Hessian consistency and the central limit theorem for the computed covariance.

Choose deterministic tolerances $\Delta_{n_{\mathsf{T}}} \leq 1/2$ and $\tau_{n_{\mathsf{T}}} \leq 1/160$ with joint failure probability tending to zero. The local theorem applies with $m = 1$, $L = 37/4$, $R = 1/6$ and $r = 1/8$. At step $1/8$ its contraction factor is at most $15/16$. Corollary 4 gives the claimed covariance $H^{-1}\mathbb{E}(gg^\top)H^{-1} = (\nu^2_{\mathsf{T}}/2)H^{-1}$. Entry uses $n_{\mathsf{S}} \to \infty$; the target limit uses $n_{\mathsf{T}} \to \infty$. This completes the proof. □

The next two lemmas verify the stochastic part of Condition 2 for the quartic loss. The proofs use truncation and net arguments only; they give crude powers of $p + \log n_{\mathsf{T}}$ in the higher-order terms. The recovery results of Candès, Li and Soltanolkotabi (2015); Sun, Qu and Wright (2018) use different empirical controls. Throughout, write $n = n_{\mathsf{T}}$, $\theta^* = \theta^*_{\mathsf{T}}$, and $F_n = F_{\mathsf{T},n_{\mathsf{T}}}$, and let $s \geq 1$.

Lemma 7. *In the model of Theorem 7, there is a constant $C_\sigma < \infty$ depending only on σ such that, with probability at least $1 - 5e^{-s}$,*

$$\|\nabla F_n(\theta^*)\| \leq C_\sigma \left\{ \sqrt{\frac{p+s}{n}} + \frac{(p + s + \log n)^{5/2}}{n} \right\}.$$

Lemma 7 bounds the displacement of the empirical stationary point from the target parameter. The leading term has the usual root-n score order, while the second term is the price of the elementary truncation argument used here.

Proof. Direct differentiation gives

$$\nabla F_n(\theta^*) = -\frac{1}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \theta^*) X_i =: -\frac{1}{n}\sum_{i=1}^n \xi_i.$$

Put $t = s + \log n$. The bound of Laurent and Massart (2000) and a union bound give $\max_{i \leq n} \|X_i\|^2 \leq 3(p + t)$ with probability at least $1 - e^{-s}$. The sub-Gaussian tail of ε_i gives $\max_{i \leq n} |\varepsilon_i| \leq C\sigma\sqrt{t}$ with probability at least $1 - 2e^{-s}$. Let $\widetilde{\xi}_i$ be ξ_i truncated by these bounds for observation i. Then $\left\|\widetilde{\xi}_i\right\| \leq C\sigma(p+t)^{3/2} =: B$, and $\mathbb{E}\xi_i = 0$ gives

$$\left\|\mathbb{E}\widetilde{\xi}_i\right\| = \|\mathbb{E}\xi_i \mathbf{1}\{\text{truncation fails}\}\| \leq \{\mathbb{E}\,\|\xi_i\|^2\}^{1/2}\{3e^{-t}\}^{1/2} \leq C\sigma\sqrt{p}\, e^{-t/2},$$

which is bounded by $C\sigma\sqrt{(p+s)/n}$ because $e^{-t/2} \le n^{-1/2}$. For any unit vector v, independence of ε_i and Lemma 6 give $\mathbb{E}\{(v^\top \xi_i)^2\} \le C\sigma^2$. Bernstein's inequality for bounded variables and a $1/2$-net $\mathcal{N}$ of the unit sphere with $|\mathcal{N}| \le 5^p$ give, with $t' = p\log 5 + s$,

$$\left\| \frac{1}{n}\sum_{i=1}^n (\widetilde{\xi}_i - \mathbb{E}\widetilde{\xi}_i) \right\| \le 2 \sup_{v\in\mathcal{N}} \left| \frac{1}{n}\sum_{i=1}^n v^\top(\widetilde{\xi}_i - \mathbb{E}\widetilde{\xi}_i) \right| \le C\left\{ \sigma\sqrt{\frac{t'}{n}} + \frac{Bt'}{n} \right\}$$

with probability at least $1 - 2e^{-s}$. A union bound gives failure probability at most $5e^{-s}$. Using $Bt' \le C\sigma(p+t)^{5/2}$ completes the proof. □

The following lemma controls the empirical curvature uniformly on the target neighborhood.

Lemma 8. *In the model of Theorem 7, there is a constant $C_\sigma < \infty$ depending only on σ such that, with probability at least $1 - 5e^{-s}$,*

$$\sup_{u:\|u-\theta^*\|\le 1/20} \left\|\nabla^2 F_n(u) - \nabla^2 F(u)\right\|_{\mathrm{op}} \le C_\sigma \left\{ \sqrt{\frac{p\log n + s}{n}} + \frac{(p+s+\log n)^2(p\log n + s)}{n} \right\}.$$

Lemma 8 ensures that the positive population curvature in Proposition 16 persists for the empirical criterion. This is the model-specific step that turns source localization into convergence to the target local estimator.

Proof. Direct differentiation gives $\nabla^2 F_n(u) = n^{-1}\sum_{i=1}^n \{3(X_i^\top u)^2 - Y_i\} X_i X_i^\top$. For unit v, w and u in the ball, $v^\top\{\nabla^2 F_n(u) - \nabla^2 F(u)\}w$ averages

$$Q_i(u,v,w) = \{3(X_i^\top u)^2 - (X_i^\top \theta^*)^2 - \varepsilon_i\}(X_i^\top v)(X_i^\top w) - \mathbb{E}\{\cdot\},$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation of the first term. Set $t = s + \log n$, and let $\mathcal{A}_i = \{\|X_i\|^2 \le 3(p+t), |\varepsilon_i| \le C\sigma\sqrt{t}\}$. By Lemma 7's tail bound, these events hold jointly over $i \le n$ with probability at least $1 - 3e^{-s}$. Put $W_i = Q_i \mathbf{1}_{\mathcal{A}_i} - \mathbb{E}(Q_i \mathbf{1}_{\mathcal{A}_i})$. For each fixed triple, these variables are independent and centered, $|W_i| \le C(1+\sigma)(p+t)^2 =: B_H$, and $\mathbb{E}W_i^2 \le C(1+\sigma^2)$ by the Cauchy-Schwarz inequality and Gaussian moment bounds with $\|u\| \le 21/20$. The truncation bias satisfies $|\mathbb{E}(Q_i \mathbf{1}_{\mathcal{A}_i^c})| \le C(1+\sigma)e^{-t/2} \le C(1+\sigma)n^{-1/2}$.

Let $\mathcal{N}_v$ be a $1/4$-net of the unit sphere with $|\mathcal{N}_v| \le 9^p$ and let $\mathcal{N}_u$ be a $1/n$-net of the ball $\{u : \|u - \theta^*\| \le 1/20\}$ with $|\mathcal{N}_u| \le n^p$ for $n \ge 2$. Bernstein's inequality for bounded variables with $t'' = 3p\log(9n) + s$ and a union bound over all triples give

$$\max_{u\in\mathcal{N}_u} \max_{v,w\in\mathcal{N}_v} \left| \frac{1}{n}\sum_{i=1}^n W_i(u,v,w) \right| \le C(1+\sigma)\left\{ \sqrt{\frac{t''}{n}} + \frac{B_H t''}{n} \right\}$$

with probability at least $1 - e^{-s}$. The operator-norm net inequality extends the bound over v, w. On $\bigcap_{i=1}^n \mathcal{A}_i$, add the bias to $n^{-1}\sum_i W_i$; it is bounded by the first term. For u and its net point u',

$$\nabla^2 F_n(u) - \nabla^2 F_n(u') = \frac{3}{n}\sum_{i=1}^n \{X_i^\top(u-u')\}\{X_i^\top(u+u')\} X_i X_i^\top,$$

whose operator norm is at most $C\,\|u-u'\| \max_{i\le n}\|X_i\|^4 \le C(p+t)^2 n^{-1}$ on the truncation event, and the population counterpart is at most $19\,\|u-u'\| \le 19n^{-1}$ by Proposition 16. Both errors are bounded by the second term. A union bound completes the proof. □

O.2 Concentration of Quadratic Gaussian Features

The empirical Hessian is determined by quadratic Gaussian features. We control their second-moment operator directly to obtain a sufficient target sample size closer to the quadratic lower bound below.

For $n \ge 2$ and $s \ge 1$, put $\ell = s + \log(32n)$ and $b = s + \log(8p^2)$. The constant C_σ below depends only on the population noise bound.

The following lemma bounds the target score and the deviation of the empirical Hessian uniformly on the target ball.

Lemma 9. *Assume the target model of Theorem 7. With probability at least* $1 - 10e^{-s}$,

$$\|\nabla F_n(\theta^*)\| \le C_\sigma \left\{ \sqrt{\frac{p+s}{n}} + \frac{\ell\sqrt{p+\ell}(p+s)}{n} \right\},$$

$$\sup_{\|u-\theta^*\|\le 1/6} \|\nabla^2 F_n(u) - \nabla^2 F_{\mathsf{T}}(u)\|_{\mathrm{op}} \le C_\sigma \left\{ (p+\ell)\sqrt{\frac{b}{n}} + \frac{(p+\ell)^2 b}{n} \right\}.$$

Lemma 9 gives uniform local curvature control at a sample size of quadratic order in the dimension, up to logarithmic factors. The estimator and empirical loss use all observations; truncation is used only to prove concentration.

Proof. For each observation, use

$$E_i = \{\|X_i\|^2 \le C(p+\ell), \quad |X_i^\top \theta^*| \le C\sqrt{\ell}, \quad |\varepsilon_i| \le C\sigma\sqrt{\ell}\}.$$

Gaussian and sub-Gaussian tail bounds give $\Pr(E_i^c) \le 5e^{-\ell}$ after increasing C. Hence $\Pr(\cap_i E_i) \ge 1 - 5ne^{-\ell}$. The score summand $-\varepsilon_i(X_i^\top\theta^*)X_i$ has norm at most $C_\sigma \ell\sqrt{p+\ell}$ on E_i, and for unit v,

$$\mathbb{E}\{\varepsilon_i(X_i^\top\theta^*)(X_i^\top v)\}^2 \le C_\sigma.$$

Cauchy–Schwarz bounds the centering error by $C_\sigma e^{-\ell/2}$ uniformly over v. Bernstein's inequality on a $1/2$-net of at most 5^p points gives the bound, with bias at most $C_\sigma\sqrt{(p+s)/n}$.

Let $\mathcal{S}_p$ be the space of real symmetric $p \times p$ matrices, with the Frobenius inner product, and let $I_{\mathcal{S}_p}$, also written I, denote its identity operator. On this space, put $d = p(p+1)/2$ and define Q by $QA = 2A + \mathrm{tr}(A)I_p$. The Gaussian fourth-moment identity gives

$$\mathbb{E}\{\langle XX^\top, A\rangle XX^\top\} = QA.$$

For $W = Q^{-1/2}(XX^\top)$, the uncentered second-moment operator is $\mathbb{E}(W \otimes W) = I_{\mathcal{S}_p}$, where $(W \otimes W)A = \langle W, A\rangle W$. Also, $\|W\|_F^2 \le \|X\|^4/2$.

For $B = Q^{-1/2}A$, $\|A\|_F = 1$, and eigenvalues λ_j of B,

$$\mathbb{E}(X^\top BX)^2 = 2\|B\|_F^2 + \{\mathrm{tr}(B)\}^2 = 1.$$

For independent standard normal variables g_j, direct expansion yields

$$\mathbb{E}\left\{\sum_j \lambda_j(g_j^2 - 1)\right\}^4 = 12\left(\sum_j \lambda_j^2\right)^2 + 48\sum_j \lambda_j^4.$$

Thus $\sup_{\|A\|_F=1} \mathbb{E}\langle W, A\rangle^4 \le C$.

Put $Z_i = (W_i \otimes W_i)1_{E_i}$. Then

$$0 \preceq Z_i \preceq C(p+\ell)^2 I, \qquad \mathbb{E}Z_i \preceq I, \qquad \mathbb{E}(Z_i - \mathbb{E}Z_i)^2 \preceq C(p+\ell)^2 I.$$

Use $Z_i^2 \preceq C(p+\ell)^2 Z_i$ and $\mathbb{E}(Z_i - \mathbb{E}Z_i)^2 = \mathbb{E}Z_i^2 - (\mathbb{E}Z_i)^2$ for the last inequality. The fourth-moment bound gives

$$\|\mathbb{E}Z_i - I\|_{\mathrm{op}} \le Ce^{-\ell/2}.$$

Matrix Bernstein's inequality (Tropp, 2012, Theorem 6.1) gives, on $\cap_i E_i$, except on an event of probability at most $2de^{-b}$,

$$\left\| \frac{1}{n}\sum_{i=1}^n W_i \otimes W_i - I \right\|_{\mathrm{op}} \le C\left\{ (p+\ell)\sqrt{\frac{b}{n}} + \frac{(p+\ell)^2 b}{n} + e^{-\ell/2} \right\}.$$

Denote the right side by D. Because $\|Q^{1/2}(uu^\top)\|_F = \sqrt{3}\|u\|^2$, it follows simultaneously for all $u, v \in \mathbb{R}^p$ that

$$\left| \frac{1}{n}\sum_i (X_i^\top u)^2(X_i^\top v)^2 - \mathbb{E}\{(X^\top u)^2(X^\top v)^2\} \right| \le 3D\|u\|^2\|v\|^2.$$

For $N_i = \varepsilon_i X_i X_i^\top 1_{E_i}$,

$$\|N_i\|_{\mathrm{op}} \le C_\sigma(p+\ell)\sqrt{\ell}, \quad \mathbb{E}(N_i - \mathbb{E}N_i)^2 \preceq C_\sigma p I, \quad \|\mathbb{E}N_i\|_{\mathrm{op}} \le C_\sigma e^{-\ell/2}.$$

Here the variance bound follows from independence and $\mathbb{E}\{\|X\|^2 XX^\top\} = (p+2)I$. Matrix Bernstein's inequality gives

$$\left\| \frac{1}{n}\sum_i \varepsilon_i X_i X_i^\top \right\|_{\mathrm{op}} \le C_\sigma \left\{ \sqrt{\frac{pb}{n}} + \frac{(p+\ell)\sqrt{\ell}\, b}{n} + e^{-\ell/2} \right\}.$$

Finally,

$$\nabla^2 F_n(u) = \frac{1}{n}\sum_i \{3(X_i^\top u)^2 - (X_i^\top \theta^*)^2 - \varepsilon_i\} X_i X_i^\top.$$

The quartic terms are bounded by $(9\|u\|^2 + 3)D$. On the stated ball, $\|u\| \le 7/6$. The noise bound is absorbed by the asserted Hessian bound. The score-net, matrix, and truncation probabilities sum to at most $10e^{-s}$. This completes the proof. □

O.3 Sample Size for Uniform Hessian Control

Uniform Hessian control on a ball of fixed radius requires at least quadratic sample growth in p. Sun, Qu and Wright (2018) discuss this obstruction for uniform concentration of quartic empirical operators. The calculation below uses the contribution of one Gaussian sensing vector.

Put $R = 1/20$, $v = X_1/\|X_1\|$, and $u_\pm = \theta^* \pm Rv$. These quantities are defined almost surely. The quadratic dependence of the Hessian on u gives

$$v^\top \left\{ \frac{\nabla^2 F_n(u_+) + \nabla^2 F_n(u_-)}{2} - \nabla^2 F_n(\theta^*) \right\} v = \frac{3R^2}{n}\sum_{i=1}^n (X_i^\top v)^4.$$

The corresponding population expression equals $9R^2$. The triangle inequality and the summand with $i = 1$ imply

$$\sup_{\|u-\theta^*\| \le R} \left\| \nabla^2 F_n(u) - \nabla^2 F_{\mathsf{T}}(u) \right\|_{\mathrm{op}} \ge \frac{3R^2}{2}\left(\frac{\|X_1\|^4}{n} - 3 \right).$$

Because $\mathbb{E}e^{-\|X_1\|^2/2} = 2^{-p/2}$, Markov's inequality gives $\mathbb{P}\{\|X_1\|^2 < p/2\} \le e^{p/4}2^{-p/2}$. Thus the uniform Hessian deviation tends to infinity in probability when $p \to \infty$ and $n = o(p^2)$. This excludes nearly linear sample sizes for a bounded uniform Hessian deviation on the fixed ball. The ball of radius $1/20$ is contained in the ball of radius $1/6$, and this obstruction also applies to the latter. The calculation does not restrict recovery guarantees based on a different property of the empirical criterion, including the local convexity and one-point convexity properties studied by Liu, Wang and Wu (2024).

P Confidence Regions after Target Refinement

The normal limit in Corollary 8 separates source accuracy from target uncertainty. We make the resulting confidence region computable from the target observations. We also replace a deterministic iteration count by a stopping rule based on the observed target gradient. Inference for sparse phase retrieval has been studied through corrected initial estimates (Yao, 2022). We consider fixed dimension and the target local estimator reached from a separate source sample. The source fit is not corrected or used in the target criterion.

P.1 Stopping and estimation of target variation

Use the model, independence, and unit-norm signals of Corollary 8. Write $n = n_{\mathsf{T}}$ and $F_n = F_{\mathsf{T},n}$. Starting at the raw source spectral estimator, apply the same target step $1/8$. Let

$$b_n = \left\lceil 64\log(n+1) + \{\log(n+1)\}^2 \right\rceil,$$

and define the stopping time and the reported estimate by

$$\widehat{t}_n = \min\left(\left\{t \in \{0,\dots,b_n\} : \left\|\nabla F_n(\theta^t)\right\| \le \frac{1}{n\log(n+1)}\right\} \cup \{b_n\}\right), \qquad \widehat{u}_n = \theta^{\widehat{t}_n}. \tag{14}$$

The cap gives a finite procedure for every sample. The gradient threshold controls the remaining optimization error near the target solution. The cap grows faster than the iterations needed for any fixed local contraction factor; its numerical value need not estimate that factor.

The target residuals give the variance estimate

$$\widehat{\nu}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\{Y_{\mathsf{T},i} - (X_{\mathsf{T},i}^\top \widehat{u}_n)^2\}^2, \qquad \widehat{H}_n = 2I_p + 4\widehat{u}_n\widehat{u}_n^\top. \tag{15}$$

The Hessian formula in (15) uses the unit-norm Gaussian model. It does not assume that the sample Hessian equals its expectation. Both estimates are unchanged if the sign of $\widehat{u}_n$ is reversed.

For $0 < \beta < 1$, let $\chi^2_{p,1-\beta}$ be the $1-\beta$ quantile of a chi-square distribution with p degrees of freedom. If $\widehat{\nu}_n^2 > 0$, define

$$\mathcal{C}_n(\beta) = \left\{u \in \mathbb{R}^p : \min_{s\in\{-1,1\}} (u - s\widehat{u}_n)^\top \widehat{H}_n (u - s\widehat{u}_n) \le \frac{\widehat{\nu}_n^2}{2n}\chi^2_{p,1-\beta}\right\}. \tag{16}$$

Set $\mathcal{C}_n(\beta) = \varnothing$ when the variance estimate is zero or an iterate is undefined. The probability of these cases tends to zero under the following conditions. The two ellipsoids represent the same signal up to its unidentifiable sign.

The following proposition establishes coverage from the source and target moment conditions.

Proposition 19. *Assume the stated phase model, fixed p, and*

$$n_{\mathsf{S}}, n \to \infty, \qquad \mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty, \qquad 0 < \nu_{\mathsf{T}}^2 = \mathbb{E}\varepsilon_{\mathsf{T}}^2 < \infty, \qquad d_{\pm}(\theta^*_{\mathsf{S}}, \theta^*_{\mathsf{T}}) < \frac{1}{8}.$$

For (14)–(16)*,*

$$\widehat{\nu}_n^2 \xrightarrow{\mathbb{P}} \nu_{\mathsf{T}}^2, \qquad \widehat{H}_n \xrightarrow{\mathbb{P}} H, \qquad \mathbb{P}\{\theta^*_{\mathsf{T}} \in \mathcal{C}_n(\beta)\} \longrightarrow 1-\beta.$$

Proposition 19 uses source information to reach the target estimator, while the target observations determine the size of the confidence region. It imposes no relative growth rate on n_{S} and n, and it does not require a fourth target-noise moment. Estimation of the target variance remains necessary; the source variance cannot replace it.

Proof. Write $\theta^* = \theta^*_{\mathsf{T}}$, and choose the sign of the source start toward θ^*. The proof of Corollary 8 gives entry into $\mathbb{B}_{1/8}$, uniform target Hessian consistency on $\mathbb{B}_{1/6}$, and a unique local stationary point $\widehat{\theta}_{\mathsf{T}}$, with probability tending to one. On these events, for all sufficiently large n,

$$\frac{1}{2}I_p \preceq \nabla^2 F_n(u) \preceq \frac{39}{4}I_p, \qquad u \in \mathbb{B}_{1/6},$$

and Theorem 5 gives

$$\left\|\theta^t - \widehat{\theta}_{\mathsf{T}}\right\| \le \left(\frac{15}{16}\right)^t \left(\frac{1}{8} + 2\left\|S_n\right\|\right).$$

Every segment between the iterate and the stationary point lies in the ball. Integrating the Hessian along that segment yields

$$\left\|\nabla F_n(\theta^t)\right\| \le \frac{39}{4}\left\|\theta^t - \widehat{\theta}_{\mathsf{T}}\right\|, \qquad \left\|\theta^t - \widehat{\theta}_{\mathsf{T}}\right\| \le 2\left\|\nabla F_n(\theta^t)\right\|.$$

Because $\|S_n\| = O_{\mathbb{P}}(n^{-1/2})$ and $(15/16)^{b_n} n\log(n+1) \to 0$, the gradient threshold is attained before the cap with probability tending to one. Thus, for the aligned output $\widetilde{u}_n$,

$$\left\|\widetilde{u}_n - \widehat{\theta}_{\mathsf{T}}\right\| \le \frac{2}{n\log(n+1)} = o(n^{-1/2}).$$

Corollary 8 and its stationary-point expansion give

$$\sqrt{n}(\widetilde{u}_n - \theta^*) \Rightarrow \mathsf{N}\left(0, \frac{\nu_{\mathsf{T}}^2}{2} H^{-1}\right). \tag{17}$$

Put $h_n = \widetilde{u}_n - \theta^*$ and $d_i = (X_i^\top \widetilde{u}_n)^2 - (X_i^\top \theta^*)^2$. On $\|h_n\| \le 1$, $|d_i| \le 3\,\|h_n\|\,\|X_i\|^2$, and

$$\left|\widehat{\nu}_n^2 - \frac{1}{n}\sum_{i=1}^n \varepsilon_{\mathsf{T},i}^2\right| \le 6\,\|h_n\|\,\frac{1}{n}\sum_{i=1}^n |\varepsilon_{\mathsf{T},i}|\,\|X_i\|^2 + 9\,\|h_n\|^2\,\frac{1}{n}\sum_{i=1}^n \|X_i\|^4.$$

Both averages on the right converge in probability to finite values. The first uses independence and the finite first absolute noise moment; the second uses Gaussian fourth moments. The average of $\varepsilon_{\mathsf{T},i}^2$ converges to ν_{T}^2. This proves variance consistency without a fourth noise moment. Continuity gives $\widehat{H}_n \to H$ in probability.

On $\widehat{\nu}_n^2 > 0$, (17) implies

$$\frac{2n}{\widehat{\nu}_n^2}(\widetilde{u}_n - \theta^*)^\top \widehat{H}_n(\widetilde{u}_n - \theta^*) \Rightarrow \chi_p^2.$$

The other sign has distance at least $2 - \|h_n\|$ from θ^*. Its quadratic form tends to infinity after multiplication by n, whereas the aligned quadratic form is tight. The minimum in (16) hence has the same limit. Continuity of the chi-square distribution proves the coverage conclusion. This completes the proof. □

P.2 Confidence intervals for prediction intensities

An observable phase response has mean $(x^\top \theta_{\mathsf{T}}^*)^2$ at sensing vector x. This mean is unchanged by the sign of the signal. Projection of $\mathcal{C}_n(\beta)$ gives intervals for these means without selecting a sign. When the region is nonempty, for $x \in \mathbb{R}^p$, write

$$a_n(x) = \left\{\frac{\widehat{\nu}_n^2}{2n}\chi_{p,1-\beta}^2 x^\top \widehat{H}_n^{-1} x\right\}^{1/2}.$$

The projected intensity interval is

$$\left[\{(|x^\top \widehat{u}_n| - a_n(x)) \vee 0\}^2,\ \{|x^\top \widehat{u}_n| + a_n(x)\}^2\right]. \tag{18}$$

Use empty intensity intervals when the confidence region is empty. These intervals have simultaneous asymptotic coverage at least $1 - \beta$ over all sensing vectors x, because the event $\theta_{\mathsf{T}}^* \in \mathcal{C}_n(\beta)$ implies coverage for every x.

To verify the projection, write $u = s\widehat{u}_n + z$ in either ellipsoid. Cauchy–Schwarz in the $\widehat{H}_n$ metric gives

$$|x^\top z| \le \{z^\top \widehat{H}_n z\}^{1/2}\{x^\top \widehat{H}_n^{-1} x\}^{1/2} \le a_n(x).$$

For $x \neq 0$, equality is attained with z proportional to $\widehat{H}_n^{-1}x$; for $x = 0$, the interval is $\{0\}$. Squaring the union of the two projected intervals gives exactly (18). For fixed x, $a_n(x) = O_{\mathbb{P}}(n^{-1/2})$. If $x^\top \theta_{\mathsf{T}}^* \neq 0$, the width is $O_{\mathbb{P}}(n^{-1/2})$. If $x^\top \theta_{\mathsf{T}}^* = 0$, (17) gives $|x^\top \widehat{u}_n| = O_{\mathbb{P}}(n^{-1/2})$, and the width is $O_{\mathbb{P}}(n^{-1})$. The projection remains valid at a zero intensity, where a first-order scalar delta-method variance would equal zero.

P.3 Limits of the projected intervals

The projection is conservative for a single sensing direction because it protects the whole signal. We quantify this effect and the different widths at zero and nonzero intensities. For a fixed nonzero x, define $\sigma_x^2 = (\nu_{\mathsf{T}}^2/2)x^\top H^{-1}x$ and $a(x) = \sigma_x\sqrt{\chi_{p,1-\beta}^2}$. Let $W_n(x)$ be the width of (18) when the region is nonempty. The probability of this event tends to one. Variance consistency and (17) give $\sqrt{n}\,a_n(x) \to a(x)$ in probability. If $x^\top \theta_{\mathsf{T}}^* \neq 0$, the two interval bounds are positive with probability

tending to one, and their difference is $4|x^\top \widehat{u}_n| a_n(x)$. If $x^\top \theta^*_{\mathsf{T}} = 0$, then $\sqrt{n}|x^\top \widehat{u}_n| \Rightarrow |Z_x|$, where $Z_x \sim \mathsf{N}(0, \sigma_x^2)$. The continuous mapping theorem gives

$$\begin{aligned} \sqrt{n}\, W_n(x) &\xrightarrow{\mathbb{P}} 4|x^\top \theta^*_{\mathsf{T}}| a(x), && x^\top \theta^*_{\mathsf{T}} \neq 0, \\ n W_n(x) &\Rightarrow \{|Z_x| + a(x)\}^2 - \{(|Z_x| - a(x)) \vee 0\}^2, && x^\top \theta^*_{\mathsf{T}} = 0. \end{aligned} \tag{19}$$

The first limit is constant. The second is random because the estimated linear prediction is of the same order as its uncertainty at a zero intensity.

Let Φ denote the standard normal distribution function. Coverage is equivalent to $\big||x^\top \widehat{u}_n| - |x^\top \theta^*_{\mathsf{T}}|\big| \leq a_n(x)$. At a nonzero intensity the sign of the aligned linear prediction agrees with its population sign with probability tending to one. At a zero intensity coverage is exactly the event $|x^\top \widehat{u}_n| \leq a_n(x)$. Both cases have limiting pointwise coverage

$$2\Phi\{\sqrt{\chi^2_{p,1-\beta}}\} - 1.$$

For $p = 4$ and $\beta = 0.05$, this probability is 0.99793. It exceeds the simultaneous signal-region coverage. The additional coverage is the cost of protecting every sensing direction by projection. Let e_j be the jth coordinate vector. For $\nu_{\mathsf{T}} = 0.35$, $\theta^*_{\mathsf{T}} = e_1$, and $x = e_1$, the constant in (19) is 1.24485. For $x = e_2$, the mean of the limiting random width equals 0.47171, obtained by integrating the stated function of $|Z_x|$. These are properties of the limiting distribution; no convergence of mean widths is claimed without an additional uniform-integrability argument.

Q The Population Target of Pooled Phase Estimation

Pooling changes the population criterion even when both studies follow the same phase-retrieval observation model. We compute its population minimizer and then compare inference at that minimizer with inference at the target signal. This gives a model-specific counterpart to the Gaussian comparison in Proposition 2.

Q.1 Population minimizer and target bias

Use the population criterion and its spectral quantities defined in Appendix H.

The following proposition identifies the parameter selected by pooling.

Proposition 20. *Assume centered independent noise with $\mathbb{E}|\varepsilon_{\mathsf{T}}| + \mathbb{E}|\varepsilon_{\mathsf{S}}| < \infty$, $0 < c \leq 1$, and $0 < w < 1$. Then*

$$\arg\min_{u \in \mathbb{R}^p} Q_w(u) = \{-\theta_w, \theta_w\}, \qquad \nabla^2 Q_w(\theta_w) = 2\{\lambda_+(w) I_p - A_w\} + 6\theta_w \theta_w^\top \succ 0.$$

If $c < 1$, then

$$d^2_{\pm}(\theta_w, \theta^*_{\mathsf{T}}) = 1 + \frac{1 + 2\lambda_+(w)}{3} - 2\left\{\frac{1 + 2\lambda_+(w)}{3} \frac{1 - w + wc^2 - \lambda_-(w)}{\lambda_+(w) - \lambda_-(w)}\right\}^{1/2} > 0.$$

If $c = 1$, then $\theta_w = \theta^*_{\mathsf{T}}$.

Proposition 20 shows two changes under pooling. The leading direction lies between the source and target directions, and its norm is less than one when the signals differ. Increasing the pooled sample size estimates this changed parameter more accurately. It does not remove the target bias.

Proof. Lemma 6 gives $\mathbb{E}(X^\top u)^4 = 3\|u\|^4$ and $\mathbb{E}\{Y_a (X^\top u)^2\} = \|u\|^2 + 2\{u^\top \theta^*_a\}^2$, which proves the expression for Q_w. If $c < 1$, use the basis θ^*_{T} and $(\theta^*_{\mathsf{S}} - c\theta^*_{\mathsf{T}})/(1 - c^2)^{1/2}$. On their span, A_w is

$$\begin{pmatrix} 1 - w + wc^2 & wc(1 - c^2)^{1/2} \\ wc(1 - c^2)^{1/2} & w(1 - c^2) \end{pmatrix}.$$

Its trace is one and its determinant is $w(1 - w)(1 - c^2)$, giving the two stated eigenvalues. Its positive off-diagonal entries imply that the leading eigenvector has positive coordinates. The leading eigenvalue is simple. For $c = 1$, $A_w = \theta^*_{\mathsf{T}} (\theta^*_{\mathsf{T}})^\top$, and the same leading-eigenvalue conclusion holds.

For $u = rv$, with $r \geq 0$ and $\|v\| = 1$,

$$Q_w(rv) = \frac{3}{4}r^4 - r^2\left(\frac{1}{2} + v^\top A_w v\right) \geq \frac{3}{4}r^4 - r^2\left(\frac{1}{2} + \lambda_+(w)\right).$$

Equality requires $v = \pm v_w$. Minimizing the right side over r^2 gives $r^2 = \{1 + 2\lambda_+(w)\}/3 > 0$, proving the global minimizer claim. Differentiation gives the stated Hessian. Its eigenvalue in direction v_w is $6\|\theta_w\|^2$; the remaining eigenvalues are $2\{\lambda_+(w) - \lambda_j(A_w)\} > 0$.

For $c < 1$, the leading spectral projection on the two-dimensional span is $(A_w - \lambda_-(w)I)/\{\lambda_+(w) - \lambda_-(w)\}$. Thus

$$\{v_w^\top \theta_{\mathsf{T}}^*\}^2 = \frac{1 - w + wc^2 - \lambda_-(w)}{\lambda_+(w) - \lambda_-(w)}.$$

Substitute this identity into $1 + \|\theta_w\|^2 - 2\|\theta_w\| v_w^\top \theta_{\mathsf{T}}^*$. When $c < 1$, $A_w\theta_{\mathsf{T}}^*$ has a nonzero component orthogonal to θ_{T}^*. Hence $v_w \neq \theta_{\mathsf{T}}^*$, which makes the sign-invariant distance positive. When $c = 1$, the formulas give $\theta_w = \theta_{\mathsf{T}}^*$. This completes the proof. □

Q.2 Sampling variation around the pooled minimizer

Use the pooled estimate and covariance defined in Appendix H.

For the limits, let

$$H_w = \nabla^2 Q_w(\theta_w), \qquad \Sigma_w = (1-w)\,\mathrm{Var}\{g_{\mathsf{T}}(\theta_w)\} + w\,\mathrm{Var}\{g_{\mathsf{S}}(\theta_w)\}, \qquad V_w = H_w^{-1}\Sigma_w H_w^{-1}.$$

If $\widehat{V}_P \succ 0$, define

$$\mathcal{C}_P(\beta) = \left\{u : N \min_{s\in\{-1,1\}} (u - s\widehat{u}_P)^\top \widehat{V}_P^{-1}(u - s\widehat{u}_P) \leq \chi^2_{p,1-\beta}\right\};$$

Use the empty set if the estimate or covariance is undefined, or if $\widehat{V}_P$ is not positive definite. Let $\widetilde{u}_P$ be the sign of $\widehat{u}_P$ closest to θ_{w_N}. The following proposition distinguishes coverage of the pooled parameter from coverage of the target signal.

Proposition 21. *Assume fixed p, the stated Gaussian designs and independent centered errors,*

$$n_{\mathsf{T}}, n_{\mathsf{S}} \to \infty, \qquad w_N \to w \in (0,1), \qquad 0 < d_\pm(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*) < \frac{1}{8},$$
$$0 < \nu_{\mathsf{T}}^2 < \infty, \qquad \mathbb{E}\varepsilon_{\mathsf{S}}^2 < \infty.$$

Then

$$\sqrt{N}(\widetilde{u}_P - \theta_{w_N}) \Rightarrow \mathsf{N}(0, V_w), \qquad \widehat{V}_P \xrightarrow{\mathbb{P}} V_w,$$

and

$$\mathbb{P}\{\theta_{w_N} \in \mathcal{C}_P(\beta)\} \to 1 - \beta, \qquad \mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_P(\beta)\} \to 0.$$

Proposition 21 uses a covariance estimate appropriate for the pooled estimating equation. Failure to cover the target signal is caused by its nonzero population displacement, not by using the target-only covariance for a pooled estimator. Source initialization has a different conclusion in Proposition 19.

Proof. Put $\delta = d_\pm(\theta_{\mathsf{S}}^*, \theta_{\mathsf{T}}^*)$, and align the two signals. The eigenvector equation and positive inner products show that v_w is a positive combination of the two signals. Its angle from either signal is at most their mutual angle, and hence its distance from either is at most δ. The eigenvalue formula gives

$$\lambda_-(w) \leq \frac{1}{2}(1 - c^2) \leq \frac{1}{2}\delta^2, \qquad 1 - \|\theta_w\| \leq \frac{2}{3}\lambda_-(w) \leq \frac{\delta^2}{3}.$$

Uniformly over $0 \leq w \leq 1$,

$$\max_{a\in\{\mathsf{T},\mathsf{S}\}} \|\theta_w - \theta_a^*\| \leq \delta + \frac{\delta^2}{3}, \qquad \frac{1}{32} + \delta + \frac{\delta^2}{3} < \frac{1}{6}.$$

The ball of radius $1/32$ centered at θ_{w_N} lies in the radius $1/6$ ball about each signal. Averaging Proposition 17 gives curvature between 1 and $37/4$ and Hessian Lipschitz constant 21 throughout this ball.

The finite-net argument in Proposition 12 applies to each study. Its integrable Hessian bound is $C\|X\|^4 + |\varepsilon_a|\|X\|^2$. The weighted empirical Hessian thus converges uniformly to the weighted population Hessian. Sample second moments and response means converge to $I_p + 2A_{w_N}$ and one. The pooled spectral initializer differs from v_{w_N} by $o_{\mathbb{P}}(1)$, whereas $\|v_{w_N} - \theta_{w_N}\| \le 1/192 < 1/64$. It enters the radius $1/64$ ball about θ_{w_N} with probability tending to one.

At θ_{w_N}, the weighted population score equals zero. Its empirical counterpart is a weighted sum of independent centered study scores. Each score is a polynomial in the Gaussian design and is linear in ε_a; finite second noise moments give finite score covariance. Applying the multivariate central limit theorem to their finitely many polynomial coefficients gives

$$\sqrt{N}\,\nabla F_{P,N}(\theta_{w_N}) \Rightarrow \mathsf{N}(0, \Sigma_w).$$

This calculation remains valid when $w_N - w$ converges slowly, because the centering is the exact population minimizer θ_{w_N}.

The local theorem with $r = 1/64$, $R = 1/32$, and the stopping calculation in Proposition 19 now give

$$\sqrt{N}(\widetilde{u}_P - \theta_{w_N}) = -H_w^{-1}\sqrt{N}\,\nabla F_{P,N}(\theta_{w_N}) + o_{\mathbb{P}}(1).$$

For covariance consistency, on a fixed bounded ball the score outer products are bounded entrywise by an integrable multiple of $\|X\|^8 + \varepsilon_a^2\|X\|^4$. Their finitely many empirical polynomial coefficients converge almost surely to their expectations. Consistency of $\widetilde{u}_P$, within-study centering, and continuity give $\widehat{\Sigma}_P \to \Sigma_w$ and $\widehat{H}_P \to H_w$ in probability. The target-noise component of Σ_w is positive definite, because $1 - w > 0$, $\nu_{\mathsf{T}}^2 > 0$, and $\theta_w \neq 0$. Thus $V_w \succ 0$ and the first coverage statement follows.

Proposition 20 gives $d_{\pm}(\theta_w, \theta_{\mathsf{T}}^*) > 0$. Continuity and $w_N \to w$ keep the target a positive distance from both centers. The maximum radius of either ellipsoid is $\{\chi^2_{p,1-\beta}\lambda_{\max}(\widehat{V}_P)/N\}^{1/2} = O_{\mathbb{P}}(N^{-1/2})$. This proves that target coverage tends to zero. This completes the proof. □

Q.3 Pooling without a small signal-separation condition

The small separation in Proposition 21 allows one common explicit ball for the two population criteria. It is not needed for fixed-dimensional inference at the pooled minimizer. The next result covers every fixed pair of distinct, nonorthogonal unit signals. All estimates and covariance matrices have the preceding definitions.

Proposition 22. *Assume the stated Gaussian phase model with fixed p, independent centered noise, and*

$$n_{\mathsf{T}}, n_{\mathsf{S}} \to \infty, \qquad w_N \to w \in (0,1), \qquad 0 < |(\theta_{\mathsf{S}}^*)^\top \theta_{\mathsf{T}}^*| < 1, \qquad 0 < \nu_{\mathsf{T}}^2 < \infty, \qquad \mathbb{E}\varepsilon_{\mathsf{S}}^2 < \infty.$$

Then

$$\sqrt{N}(\widetilde{u}_P - \theta_{w_N}) \Rightarrow \mathsf{N}(0, V_w), \qquad \widehat{V}_P \xrightarrow{\mathbb{P}} V_w,$$

and

$$\mathbb{P}\{\theta_{w_N} \in \mathcal{C}_P(\beta)\} \to 1 - \beta, \qquad \mathbb{P}\{\theta_{\mathsf{T}}^* \in \mathcal{C}_P(\beta)\} \to 0.$$

Proposition 22 and Theorem 8 compare the same two procedures beyond a small target neighborhood. Source initialization preserves target coverage, while pooling estimates a different parameter. The source and target noises may have different variances, and the pooled covariance is centered within each study.

Proof. Align the signals to have positive inner product and write $\lambda = \lambda_+(w)$, $v = v_w$, and $r_* = \|\theta_w\| = \sqrt{(1 + 2\lambda)/3}$. Then $1/2 < \lambda < 1$. For the population pooled gradient map with step $1/8$, a start at v stays on its span. Its radial update is

$$r_{t+1} = f(r_t), \qquad f(r) = r\left\{1 + \frac{1 + 2\lambda - 3r^2}{8}\right\}, \qquad r_0 = 1.$$

On $[r_*, 1]$,

$$\frac{\lambda}{4} \le f'(r) = \frac{9 + 2\lambda - 9r^2}{8} \le \frac{3 - 2\lambda}{4} < \frac{1}{2}.$$

Because $f(r_*) = r_*$, the interval is invariant and $0 < r_t - r_* \le 2^{-t}(1 - r_*)$. The population path reaches any fixed neighborhood of θ_w in finitely many iterations. Its gradient is nonzero at every finite time.

Proposition 20 gives a positive definite Hessian at θ_w. Its largest eigenvalue is $6\|\theta_w\|^2 = 2+4\lambda \le 6$. Continuity in u and w gives constants $m > 0$, $L < 16$, and a fixed ball about θ_w on which the Hessians of Q_{w_N} lie between mI_p and LI_p for all sufficiently large N. Their Lipschitz constants are bounded on this ball. Choose an inner radius less than one quarter of its radius. Because $\theta_{w_N} \to \theta_w$, balls of twice this inner radius about θ_{w_N} lie in the same curvature region for all large N. The population path can be chosen to enter half this inner radius. The local contraction factor for step $1/8$ is bounded away from one. These constants may depend on the fixed signals and the limiting source fraction.

The study-specific moment averages converge, giving a raw pooled spectral start converging to v. The empirical gradient maps converge uniformly on fixed compact sets to the pooled population map. Both claims follow from convergence of their finitely many empirical coefficients, as in the proof of Theorem 8. The finite-path induction gives entry into the local ball with probability tending to one. Nonzero population gradients before entry exclude premature stopping. Local contraction then makes the gradient threshold attainable before the cap. The stopped iterate differs from the local root by $O_{\mathbb{P}}\{1/(N\log(N+1))\}$.

At the exact population minimizer θ_{w_N}, the weighted population score is zero. Hence

$$\sqrt{N}\,\nabla F_{P,N}(\theta_{w_N}) = \sum_{a\in\{\mathsf{T},\mathsf{S}\}} \sqrt{\frac{n_a}{N}}\frac{1}{\sqrt{n_a}}\sum_{i=1}^{n_a}\{g_{a,i}(\theta_{w_N}) - \mathbb{E}g_a(\theta_{w_N})\} \Rightarrow \mathsf{N}(0, \Sigma_w).$$

For each study the score is a polynomial with finitely many coefficients having finite second moments. Their joint central limit theorem and $\theta_{w_N} \to \theta_w$ justify the last convergence without a rate condition on $w_N - w$. Taylor expansion of the local score yields

$$\sqrt{N}(\widetilde{u}_P - \theta_{w_N}) = -H_w^{-1}\sqrt{N}\,\nabla F_{P,N}(\theta_{w_N}) + o_{\mathbb{P}}(1).$$

On bounded parameter sets, score outer products have an integrable bound $C\{\|X\|^8 + \varepsilon_a^2\|X\|^4\}$. The corresponding coefficient averages, within-study centering, and Hessian consistency give $\widehat{V}_P \to V_w$ in probability. Its limit is positive definite because the target contribution has weight $1 - w > 0$, positive noise variance, and $\theta_w \neq 0$. The centered Wald limit proves coverage of θ_{w_N}. The target is a fixed positive distance from $\{\theta_w, -\theta_w\}$ by Proposition 20. Its coverage tends to zero because both confidence-region radii are $O_{\mathbb{P}}(N^{-1/2})$. This completes the proof. □

R Supplementary Model-Example Verification

The phase-retrieval calculations verify Theorem 7, and the Gaussian-mixture calculation verifies local target equivalence in fixed dimension. The low-rank and sparse-regression examples identify the additional model-specific properties needed for such a conclusion. In each model, source localization determines entry into the target convergence region, and source-data shift enters a pooled criterion.

R.1 Phase retrieval

For phase retrieval, the population calculation in Propositions 15 and 17 verifies curvature on the sign-aligned ball of radius $1/6$. Lemma 9 proves the concentration bound in (5) on this ball; the general sufficient condition of Proposition 11 does not apply here because the quartic loss has heavier than sub-exponential Hessian entries. The source estimator gives the source tail condition. Written in the sign-invariant distance, the relevant probability is

$$\mathbb{P}\{d_{\pm}(\widehat{\theta}_{\mathsf{S}}, \theta^*_{\mathsf{S}}) > a_{\mathsf{S}}(s)\} \le e^{-s}.$$

If

$$d_{\pm}(\theta^*_{\mathsf{S}}, \theta^*_{\mathsf{T}}) + a_{\mathsf{S}}(s) \le r,$$

then Proposition 14 places the source start inside the target chart, and Theorem 5 gives the target local expansion for the target-refitted estimator. Theorem 7 records the resulting end-to-end guarantee, and this sequence is the formal reason that the simulation reports both the source-initialization error and the target-refit error.

R.2 Low-rank factorization

Let $X = (U, V)$ denote a factor pair and let $X^* = (U^*, V^*)$. The rotation distance can be written as

$$d_{\mathrm{rot}}(X, X^*) = \min_{Q^\top Q = I_r} \{\|U - U^*Q\|_F^2 + \|V - V^*Q\|_F^2\}^{1/2}.$$

Local factor analyses control error after alignment under their sampling conditions (Tu et al., 2016; Sun and Luo, 2016). Source localization determines whether the initial factors enter the region of such a guarantee. A source factor estimator with tail

$$\mathbb{P}\{d_{\mathrm{rot}}(\widehat{X}_{\mathsf{S}}, X^*_{\mathsf{S}}) > a_{\mathsf{S}}(s)\} \le e^{-s}$$

enters the target region of radius r whenever

$$d_{\mathrm{rot}}(X^*_{\mathsf{S}}, X^*_{\mathsf{T}}) + a_{\mathsf{S}}(s) \le r.$$

The source-data comparison has a different mathematical form. Write $\Delta_M = M^*_{\mathsf{S}} - M^*_{\mathsf{T}}$. For source observations $Y = \langle A, M^*_{\mathsf{S}} \rangle + \varepsilon$ with $\mathbb{E}_{\mathsf{S}}(\varepsilon \mid A) = 0$, differentiation of their squared-error loss at M^*_{T} gives the population matrix score $-\mathbb{E}_{\mathsf{S}}\{\langle A, \Delta_M \rangle A\}$. This term depends on the source sensing distribution. Initialization transfer does not put source observations into the target score; source-target separation affects its probability of entering the target factor chart.

R.3 Gaussian mixture EM

For the balanced mixture with known identity covariance, assume that p is fixed and $\mu^*_{\mathsf{T}} \neq 0$. Explicit convergence regions under stronger signal conditions are given by Balakrishnan, Wainwright and Yu (2017). The following local calculation verifies the target expansion; its constants may depend on p and μ^*_{T}.

Up to a term independent of μ, the negative log-likelihood is

$$\ell_{\mathsf{T}}(\mu; z) = \frac{\|\mu\|^2}{2} - \log\cosh(z^\top \mu), \qquad \nabla \ell_{\mathsf{T}}(\mu; z) = \mu - z \tanh(z^\top \mu).$$

Thus, with $n = n_{\mathsf{T}}$,

$$M_{\mathsf{T},n}(\mu) = \mu - \nabla L_{\mathsf{T},n}(\mu), \qquad \nabla^2 L_{\mathsf{T},n}(\mu) = I_p - \frac{1}{n}\sum_{i=1}^{n} Z_{\mathsf{T},i} Z_{\mathsf{T},i}^\top \operatorname{sech}^2(Z_{\mathsf{T},i}^\top \mu).$$

Let $g(z) = \mu^*_{\mathsf{T}} - z \tanh(z^\top \mu^*_{\mathsf{T}})$. Differentiating the integral of the mixture density gives

$$\mathbb{E} g(Z_{\mathsf{T}}) = 0, \qquad H = \nabla^2 L_{\mathsf{T}}(\mu^*_{\mathsf{T}}) = \mathbb{E}\{g(Z_{\mathsf{T}}) g(Z_{\mathsf{T}})^\top\}.$$

Gaussian tails justify differentiation uniformly near μ^*_{T}. If $a^\top H a = 0$, positivity of the density and continuity of g give $a^\top g(z) = 0$ for all z. At $z = 0$, this gives $a^\top \mu^*_{\mathsf{T}} = 0$. For $a \neq 0$, taking $z = a + \mu^*_{\mathsf{T}}$ gives $a^\top g(z) = -\|a\|^2 \tanh(\|\mu^*_{\mathsf{T}}\|^2) \neq 0$, a contradiction. Hence H is positive definite. The Hessian formula and continuity give $0 < R < \|\mu^*_{\mathsf{T}}\|$ and $m > 0$ such that

$$m I_p \preceq \nabla^2 L_{\mathsf{T}}(\mu) \preceq I_p \quad \text{for } \|\mu - \mu^*_{\mathsf{T}}\| \le R.$$

The sample Hessian has Lipschitz constant at most $2n^{-1}\sum_i \|Z_{\mathsf{T},i}\|^3$, converging in probability to $2\mathbb{E}\|Z_{\mathsf{T}}\|^3$, which also bounds the population Lipschitz constant.

Directional Hessian summands are sub-exponential, bounded in absolute value by $C + |v^\top Z_{\mathsf{T}}||w^\top Z_{\mathsf{T}}|$. Apply Bernstein's inequality on an $n^{-1/2}$-net of the parameter ball and $1/4$-nets

of the unit spheres: at most $Cn^{p/2}9^{2p}$ triples. Exponent $(p+2)\log n$ gives failure probability tending to zero. Extend by the Lipschitz bounds as in Lemma 5. The score $g(Z_{\mathsf{T}})$ is directionally sub-Gaussian. For large C, with probability tending to one,

$$\|\nabla L_{\mathsf{T},n}(\mu_{\mathsf{T}}^*)\| \le C\sqrt{\frac{\log n}{n}}, \qquad \sup_{\|\mu-\mu_{\mathsf{T}}^*\|\le R} \left\|\nabla^2 L_{\mathsf{T},n}(\mu) - \nabla^2 L_{\mathsf{T}}(\mu)\right\|_{\mathrm{op}} \le C\sqrt{\frac{\log n}{n}}.$$

Set $\tau_{n_{\mathsf{T}}} = \Delta_{n_{\mathsf{T}}} = C\sqrt{\log n/n}$. For all sufficiently large n, Theorem 5 applies with $L=1$, $\alpha=1$, $r = R/2$, and $\theta_{\mathsf{T}}^* = \mu_{\mathsf{T}}^*$. It gives the unique local stationary point $\widehat{\mu}_{\mathsf{T}}$. On this event, the EM derivative lies between zero and $(1-m/2)I_p$. Integrating to $\widehat{\mu}_{\mathsf{T}}$ gives, for $\kappa = 1 - m/2 < 1$,

$$\|M_{\mathsf{T},n_{\mathsf{T}}}(\mu) - \widehat{\mu}_{\mathsf{T}}\| \le \kappa \|\mu - \widehat{\mu}_{\mathsf{T}}\|$$

throughout the local ball. In addition, $\sqrt{n}\,\nabla L_{\mathsf{T},n}(\mu_{\mathsf{T}}^*) \Rightarrow \mathsf{N}(0,H)$. The local estimator has asymptotic covariance H^{-1}. Corollary 4 gives the same limit for iterates whose starts enter the ball with probability tending to one and whose iteration counts satisfy $\sqrt{n}\,\kappa^{t_n} \to 0$. For a source bound $\mathbb{P}\{d_{\pm}(\widehat{\mu}_{\mathsf{S}}, \mu_{\mathsf{S}}^*) > a_{\mathsf{S}}(s)\} \le e^{-s}$, the inequality

$$d_{\pm}(\mu_{\mathsf{S}}^*, \mu_{\mathsf{T}}^*) + a_{\mathsf{S}}(s) \le r,$$

places the sign-aligned source start in this ball with probability at least $1 - e^{-s}$. If this entry probability tends to one, the sign-aligned source-initialized EM iterate has the same first-order limit as $\widehat{\mu}_{\mathsf{T}}$. The model gives local target equivalence without assuming contraction of the fitted EM map.

R.4 Folded-concave sparse regression

For folded-concave sparse regression, let $S_{\mathsf{T}} = \{j : \beta_{\mathsf{T},j}^* \neq 0\}$ be the target active support and let S_{T}^c be its complement in $\{1,\dots,p\}$. Subscripts restrict vectors and matrices to the indicated coordinates. Let $H_{S_{\mathsf{T}}S_{\mathsf{T}}}$ be the target oracle Hessian. In the oracle chart, the local coordinate is $\beta_{S_{\mathsf{T}}}$ and the inactive coordinates are held at zero. For a positive signal bound $b_{\min}$, a compact sufficient condition has the form

$$\lambda_{\min}(H_{S_{\mathsf{T}}S_{\mathsf{T}}}) \ge m, \qquad \min_{j\in S_{\mathsf{T}}} |\beta_{\mathsf{T},j}^*| \ge b_{\min},$$

combined with a local score bound and a local concavity bound for the penalty. For a vector v, write $\|v\|_\infty = \max_j |v_j|$. The source initializer is useful only if it enters this oracle chart. A typical sufficient event is

$$\left\|\widehat{\beta}_{\mathsf{S},S_{\mathsf{T}}} - \beta_{\mathsf{T},S_{\mathsf{T}}}^*\right\|_\infty \le b_{\min}/2, \qquad \widehat{\beta}_{\mathsf{S},S_{\mathsf{T}}^c} = 0$$

or a relaxed screened version followed by a target-only local refit. The source-data analogue is more than a Euclidean shift: it can change the score on inactive coordinates and the model-selection part of the target problem. This is why the general theory keeps initialization transfer and source-data transfer separate.

S Validation Implementation

Proposition 5 uses statistical separation and not an optimization stopping rule. In a nonconvex problem, a small gradient norm and a positive local Hessian show only that a point is locally stable for the empirical objective; they do not show that the point is the target local estimator. A validation rule must compare a candidate with a computable target-based reference.

For a target validation sample $Z_{\mathsf{T},i}^v$ of size n_v, independent of the sample that produced the candidate outputs, a natural validation loss is

$$V_n(\theta) = \frac{1}{n_v}\sum_{i=1}^{n_v} \ell_{\mathsf{T}}(\theta; Z_{\mathsf{T},i}^v).$$

With $\mathcal{A}_r = \{\theta \in \Theta_0 : \|\theta - \theta_{\mathsf{T}}^*\| \ge r\}$ as in Section D, Proposition 5 requires

$$\max_{\theta\in\{\bar{\theta},\widehat{\theta}_{\mathrm{ref}}\}} |V_n(\theta) - R_v(\theta)| \le \gamma_n, \qquad \inf_{\theta\in\mathcal{A}_r} R_v(\theta) - R_v(\widehat{\theta}_{\mathrm{ref}}) > 4\gamma_n.$$

Proposition 6 gives the first inequality when the population validation losses have uniformly bounded sub-exponential norms. Its conclusion controls only the two completed outputs and requires no covering argument. The second inequality concerns the fitted reference and requires a population loss gap. The proof of Proposition 5 uses two tolerances for concentration and two for the acceptance threshold.

For validation observations from the target distribution, $R_v = L_{\mathsf{T}}$. Suppose that the conditions of Proposition 3 hold with $\delta + a_{\mathsf{S}}(s_{\mathsf{S}}) \le r$ and $\mathbb{B}_R \subset \Theta_0$, and let the reference be $\widehat{\theta}_{\mathrm{ref}} = \theta^t(\widehat{\theta}_{\mathsf{S}})$. Define the deterministic reference-error bound

$$e_t = \frac{2\tau_{n_{\mathsf{T}}}}{m} + \kappa_{n_{\mathsf{T}}}^t \left(r + \frac{2\tau_{n_{\mathsf{T}}}}{m} \right).$$

Proposition 3 gives $\widehat{\theta}_{\mathrm{ref}} \in \mathbb{B}_R$ and $\left\| \widehat{\theta}_{\mathrm{ref}} - \theta_{\mathsf{T}}^* \right\| \le e_t$ with failure probability at most $\eta_{n_{\mathsf{T}}} + e^{-s_{\mathsf{S}}}$. Condition 2 and $\nabla L_{\mathsf{T}}(\theta_{\mathsf{T}}^*) = 0$ then give $R_v(\widehat{\theta}_{\mathrm{ref}}) - R_v(\theta_{\mathsf{T}}^*) \le L e_t^2/2$. Thus the deterministic population condition

$$\inf_{\theta \in \mathcal{A}_r} \{R_v(\theta) - R_v(\theta_{\mathsf{T}}^*)\} > \frac{L}{2} e_t^2 + 4\gamma_n$$

gives reference separation on the same event. Under the tail bound in Proposition 6, take $K = 2$, $s_v \ge 1$, and $\gamma_n = CK_v\{\sqrt{(s_v + \log 4)/n_v} + (s_v + \log 4)/n_v\}$. Both outputs lie in Θ_0 on this fitting event, where Proposition 6 bounds conditional failure by e^{-s_v}. The proof of Proposition 5 and a union bound give false-acceptance probability at most $\eta_{n_{\mathsf{T}}} + e^{-s_{\mathsf{S}}} + e^{-s_v}$. The population gap still requires verification; another parameter with the same loss can make it zero.

The quartic phase loss need not satisfy the sub-exponential tail bound, and this calibration is not asserted for the phase experiment. The phase selector experiment evaluates the rule in Proposition 10 by choosing between two completed candidates on independent validation data. It does not implement the one-sided rejection rule in Proposition 5. The remaining simulations report target error, while the real-data analysis reports target reconstruction error on target test images. These quantities do not verify Proposition 5.

If a source-initialized run is rejected, a practical procedure must use a target-only default or a separately validated restart. A risk statement also needs a bound for that replacement and for outputs accepted when validation concentration fails; the manuscript does not assume these bounds implicitly.

S.1 Validation by the Phase-Retrieval Loss

The phase-retrieval validation criterion estimates prediction loss. Its population value differs from the local quadratic parameter error. We derive this difference and control the sampling error of the validation contrast for the two completed candidates.

For the validation observations, let $Y = (X^\top \theta^*)^2 + \varepsilon$, where $\theta^* = \theta_{\mathsf{T}}^*$, $X \sim \mathsf{N}(0, I_p)$, and $\|\theta^*\| = 1$. Assume that ε is independent of X, $\mathbb{E}\varepsilon = 0$, and $\nu^2 = \mathbb{E}\varepsilon^2 < \infty$. Let u_a, u_b be computed independently of the validation sample. For each candidate, choose a sign minimizing its distance to θ^*, and let h_c be the signed candidate minus θ^*, for $c \in \{a, b\}$. Put $r_v = \max_{c \in \{a,b\}} \|h_c\|$ and $H = 2I_p + 4\theta^*(\theta^*)^\top$. The sign choice also minimizes the quadratic error because $H\theta^* = 6\theta^*$. Proposition 15 gives the identity

$$F_{\mathsf{T}}(u_c) - F_{\mathsf{T}}(\theta^*) = \frac{1}{2} \|h_c\|_H^2 + 3\{(\theta^*)^\top h_c\} \|h_c\|^2 + \frac{3}{4} \|h_c\|^4. \tag{20}$$

The constant noise contribution cancels in this difference. The last two terms explain why a comparison of prediction losses need not give the same ordering as a comparison of quadratic parameter errors.

Let V_{n_v} be the average quartic loss on n_v independent target observations, and let u_{sel} minimize V_{n_v} over the two candidates, with a fixed rule for ties. Let h_{sel} be its signed error. Write $\mathbb{E}_v$ and $\mathbb{P}_v$ for expectation and probability conditional on the fitted candidates. Define

$$W = \frac{\|u_a - u_b\|^2 \|u_a + u_b\|^2}{4} \left\{ 105\big(\|u_a\|^4 + \|u_b\|^4 + 4\big) + 12\nu^2 \right\}.$$

For $0 < \alpha_v < 1$, put $g_v = \{W/(n_v\alpha_v)\}^{1/2}$, and let $\Delta_v = |F_{\mathsf{T}}(u_a) - F_{\mathsf{T}}(u_b)|$. If $\Delta_v = 0$, both candidates have the same population loss and the selection regret is zero. The following proposition bounds selection error and expected excess loss.

Proposition 23. *Assume the stated validation model and independent validation observations. With conditional probability at least* $1 - \alpha_v$*,*

$$F_{\mathsf{T}}(u_{\mathrm{sel}}) \le \min_{c\in\{a,b\}} F_{\mathsf{T}}(u_c) + g_v, \qquad \|h_{\mathrm{sel}}\|_H^2 \le \min_{c\in\{a,b\}} \|h_c\|_H^2 + 2g_v + 12r_v^3 + 3r_v^4.$$

For $\Delta_v > 0$*,*

$$\mathbb{E}_v\{F_{\mathsf{T}}(u_{\mathrm{sel}}) - \min_{c\in\{a,b\}} F_{\mathsf{T}}(u_c)\} \le \frac{\Delta_v W}{W + n_v\Delta_v^2} \le \frac{1}{2}\sqrt{\frac{W}{n_v}}.$$

In all cases,

$$\mathbb{E}_v \|h_{\mathrm{sel}}\|_H^2 \le \min_{c\in\{a,b\}} \|h_c\|_H^2 + \sqrt{\frac{W}{n_v}} + 12r_v^3 + 3r_v^4.$$

Proposition 23 separates validation variation from the approximation of prediction loss by quadratic parameter error. Only finite noise variance is needed because the squared-noise term cancels in the validation contrast. For bounded candidate norms, validation variation decreases with the sample size and with agreement of the candidates up to sign. For a fixed positive loss gap, expected excess prediction loss is at most $W/(n_v\Delta_v)$. This bound requires only finite noise variance. The cubic and quartic terms are of smaller order than r_v^2 as both candidates approach the target. The expectation is conditional on the candidates; integrating it over the fitting data requires the corresponding moments of W and r_v.

Proof. For one validation observation, let $D = \{((X^\top u_a)^2 - Y)^2 - ((X^\top u_b)^2 - Y)^2\}/4$. Put $d = u_a - u_b$ and $z = u_a + u_b$. The difference of squares gives

$$D = \frac{1}{4}(X^\top d)(X^\top z)\big\{(X^\top u_a)^2 + (X^\top u_b)^2 - 2(X^\top\theta^*)^2 - 2\varepsilon\big\}.$$

For every w, Hölder's inequality and Gaussian eighth moments imply

$$\mathbb{E}\{(X^\top d)^2(X^\top z)^2(X^\top w)^4\} \le 105\,\|d\|^2\,\|z\|^2\,\|w\|^4.$$

Noise independence and Lemma 6 give $\mathbb{E}\{\varepsilon^2(X^\top d)^2(X^\top z)^2\} \le 3\nu^2\,\|d\|^2\,\|z\|^2$. Applying $(x_1 + x_2 + x_3 + x_4)^2 \le 4\sum_j x_j^2$ to the last factor of D yields

$$\mathrm{Var}_v(D) \le \mathbb{E}_v D^2 \le \frac{\|d\|^2\,\|z\|^2}{4}\{105(\|u_a\|^4 + \|u_b\|^4 + 4) + 12\nu^2\} = W.$$

Let $\overline{D}$ average the n_v contrasts and put $\mu_D = F_{\mathsf{T}}(u_a) - F_{\mathsf{T}}(u_b)$. Independence gives

$$\mathbb{P}_v\{|\overline{D} - \mu_D| > g_v\} \le \alpha_v, \qquad \mathbb{E}_v|\overline{D} - \mu_D| \le \sqrt{W/n_v}.$$

If $W = 0$, then $\overline{D} = \mu_D$ almost surely. Selecting the larger population loss gives a gap at most $|\overline{D} - \mu_D|$. Hence, for either c,

$$F_{\mathsf{T}}(u_{\mathrm{sel}}) \le F_{\mathsf{T}}(u_c) + |\overline{D} - \mu_D|.$$

Suppose $\Delta_v > 0$, and put $U = \mathrm{sgn}(\mu_D)(\overline{D} - \mu_D)$. Selecting the larger population loss implies $U \le -\Delta_v$. For any $c \ge 0$,

$$\mathbb{P}_v(U \le -\Delta_v) \le \frac{\mathbb{E}_v(U - c)^2}{(\Delta_v + c)^2} \le \frac{W/n_v + c^2}{(\Delta_v + c)^2}.$$

Taking $c = W/(n_v\Delta_v)$ gives

$$\mathbb{E}_v\{F_{\mathsf{T}}(u_{\mathrm{sel}}) - \min_{c\in\{a,b\}} F_{\mathsf{T}}(u_c)\} \le \frac{\Delta_v W}{W + n_v\Delta_v^2} \le \frac{1}{2}\sqrt{\frac{W}{n_v}}.$$

The last inequality follows from $W + n_v\Delta_v^2 \ge 2\Delta_v\sqrt{n_v W}$; when $W = 0$, the selection regret is zero.

For each candidate, the remainder in (20) has absolute value at most $3r_v^3 + 3r_v^4/4$. Choose c minimizing $\|h_c\|_H^2$ and substitute the identity to obtain

$$\|h_{\mathrm{sel}}\|_H^2 \le \|h_c\|_H^2 + 2\{F_{\mathsf{T}}(u_{\mathrm{sel}}) - \min_{j\in\{a,b\}} F_{\mathsf{T}}(u_j)\} + 12r_v^3 + 3r_v^4.$$

The probability and expectation bounds follow. This completes the proof. □

The expected loss admits a sharper comparison when one candidate has small excess prediction loss. Let $e_c = F_{\mathsf{T}}(u_c) - F_{\mathsf{T}}(\theta^*)$ for $c \in \{a, b, \mathrm{sel}\}$. Put $B = 1 \vee \|u_a\| \vee \|u_b\|$ and $A_B = \nu^2 + B^4 + 1$. Under the conditions of Proposition 23, for $0 < \eta \le 1$,

$$\mathbb{E}_v e_{\mathrm{sel}} \le (1+\eta)\min(e_a, e_b) + \frac{1890 A_B}{\eta n_v}.$$

For bounded candidate norms and fixed η, the additive remainder is of order n_v^{-1}. This is a conditional expectation over validation observations; averaging over fitted candidates also requires a finite fourth moment of their norms. It does not establish confidence coverage after selecting between the fits.

To prove the bound, write $q_c = (X^\top u_c)^2 - (X^\top\theta^*)^2$. Noise independence and Gaussian fourth moments give

$$\|u_a - u_b\|^2 \|u_a + u_b\|^2 \le \mathbb{E}(q_a - q_b)^2 \le 8(e_a + e_b), \qquad W \le 630 A_B (e_a + e_b).$$

Let $x = e_a \vee e_b$, $y = e_a \wedge e_b$, and $\Delta = x - y$. If $\Delta \le \eta y$, the excess over y is at most ηy. If $\Delta > \eta y$, the preceding selection-probability bound yields

$$\mathbb{E}_v e_{\mathrm{sel}} - y \le \frac{W}{n_v \Delta} \le \frac{630 A_B}{n_v}\left(1 + \frac{2}{\eta}\right) \le \frac{1890 A_B}{\eta n_v}.$$

These two cases prove the claim.

Prediction loss also controls sign-invariant distance globally:

$$d_\pm(u, \theta^*)^2 \le 4\{F_{\mathsf{T}}(u) - F_{\mathsf{T}}(\theta^*)\}, \qquad u \in \mathbb{R}^p. \tag{21}$$

To prove (21), choose the sign of u such that $(\theta^*)^\top u \ge 0$. For $r = \|u\| > 0$, put $c = (\theta^*)^\top u / r \in [0, 1]$. The fourth-moment identity gives

$$4\{F_{\mathsf{T}}(u) - F_{\mathsf{T}}(\theta^*)\} - d_\pm(u, \theta^*)^2 = 3r^4 - 3r^2 - 4r^2c^2 + 2rc + 2.$$

The expression is concave in c. At $c = 0$ it equals $3(r^2 - 1/2)^2 + 5/4$, and at $c = 1$ it equals $(r-1)^2(3r^2 + 6r + 2)$. Both are nonnegative. For $r = 0$, the population excess loss equals $3/4$, which completes the proof. Equation (21) and Proposition 23 give, on the same event,

$$d_\pm(u_{\mathrm{sel}}, \theta^*)^2 \le 4 \min_{c\in\{a,b\}} \{F_{\mathsf{T}}(u_c) - F_{\mathsf{T}}(\theta^*)\} + 4 g_v.$$

Thus one candidate with small population loss is sufficient for a small selected error when validation variation is small. The other candidate may lie outside the local target ball.

The quadratic-risk remainder cannot be omitted in general. For example, take $u_a = (11/10)\theta^*$ and $u_b = (89/100)\theta^*$. Their quadratic risks are $6/100 < 726/10000$, while their population excess losses are $1323/40000 > 12966723/400000000$, respectively. Even population validation can then select the larger quadratic error. Equation (20) specifies the correction needed to compare the two risks.

T Numerical Studies

The numerical studies compare source-data transfer and initialization transfer through target error and convergence. We report the phase-retrieval comparison under population differences, direct evaluation of the validation selector, a larger matrix-completion experiment, and a real-data analysis. The simulation designs in this section use 100 Monte Carlo repetitions, and the real-data analysis uses 100 target splits. Implementation details are given in Appendix W.

T.1 Phase-Retrieval Simulation

The target model is the phase-retrieval model of Section I.1 with $p = 24$,

$$Y_i = (X_i^\top \theta_{\mathsf{T}}^*)^2 + \varepsilon_i, \qquad X_i \sim \mathsf{N}(0, I_{24}), \quad \varepsilon_i \sim \mathsf{N}(0, 0.35^2),$$

and the target criterion $F_{\mathsf{T},n_{\mathsf{T}}}$. The target sample size is $n_{\mathsf{T}} = 55$. The source sample size is $n_{\mathsf{S}} = 320$. For an orthogonal unit vector v, the source truth is

$$\theta_{\mathsf{S}}^* = \frac{\theta_{\mathsf{T}}^* + \rho v}{\sqrt{1+\rho^2}}.$$

We use $\rho \in \{0.35, 0.55, 0.75\}$. Thus ρ is the design shift and the sign-invariant population distance defined in Section A is $\delta(\rho) = \{2 - 2/(1+\rho^2)^{1/2}\}^{1/2}$. The source estimator is the spectral initializer computed from the source sample. The main method runs target gradient descent from the source spectral initializer. The baselines are target spectral initialization, target random restarts, and a pooled source-target fit. Success means sign-invariant Euclidean error at most 0.35, the noise standard deviation. The prespecified source-distance proxy uses radius 0.85. The source-proxy column applies only to the source initializer. The experiment uses the stabilized source scale and backtracking described in Appendix W. Corollary 7 analyzes the empirical-response scale and the fixed step $1/8$.

Table 4: Phase-retrieval simulation with 100 repetitions per design shift. Error is sign-invariant Euclidean error to the target parameter. Numbers in parentheses represent standard deviations.

Design shift	Method	Mean error	Median error	Success	Source proxy
0.35	pooled data	0.295 (0.018)	0.295	1.00	–
0.35	source init, target refit	0.250 (0.087)	0.227	0.88	1.00
0.35	target random restart	0.373 (0.302)	0.245	0.71	–
0.35	target spectral	0.924 (0.538)	1.107	0.29	–
0.55	pooled data	0.443 (0.019)	0.442	0.00	–
0.55	source init, target refit	0.252 (0.089)	0.239	0.91	0.98
0.55	target random restart	0.430 (0.381)	0.268	0.70	–
0.55	target spectral	0.875 (0.545)	0.987	0.32	–
0.75	pooled data	0.563 (0.029)	0.567	0.00	–
0.75	source init, target refit	0.277 (0.156)	0.235	0.83	0.73
0.75	target random restart	0.371 (0.301)	0.256	0.73	–
0.75	target spectral	0.913 (0.528)	0.984	0.28	–

Table 4 shows smaller mean error for source-initialized target fitting than for pooling and both target-only starts at each reported shift. The target criterion remains the same across starting values. At $\rho = 0.75$, the source-proxy rate falls to 0.73 while the success rate of source initialization remains 0.83, which shows that the proxy is not a definition of target success. The experiment is not a finite-sample verification of Corollary 7: $n_{\mathsf{T}} = 55$ is below its conservative sample regime, and the radius 0.85 is a reporting proxy, not the certified radius $1/8$. The data-generating process and target criterion match the theorem, but the source scale, update rule, and sample size do not satisfy its sufficient statement.

T.2 Summary of additional comparisons

We also compare the two transfer candidates by validation. The fitting design is the same as in Table 2, and an independent target sample of size 100 selects the candidate with the smaller quartic loss. Risk is squared error in the population Hessian norm. Across the three reported shifts, the validation selector has mean risk 0.1465, compared with 0.1400 for the infeasible oracle, 0.1932 for always using source initialization, and 0.4567 for always pooling. Its correct-choice rate is 0.9133. Table 6 gives these risks by shift, and Table 7 shows the same adaptation for population quartic loss at four validation sample sizes.

We next consider an 80×80 positive semidefinite matrix of rank 5. The source eigenspace is rotated through an angle ϕ toward an orthogonal subspace. We observe 1,600 noisy target entries and

12,000 noisy source entries. The source estimator uses 30 source iterations. Source initialization, target spectral initialization, and pooling each receive 30 target iterations. The random comparator receives three 30-iteration target runs and is selected by target training loss. The pooled comparison fits the combined entries.

Table 5: Normalized target matrix error over 100 repetitions. The last two columns report paired source-initialization win rates.

ϕ	Pooled	Source init.	Target spectral	Random restart	vs. pooled	vs. spectral
0.25	0.320	0.273	0.304	0.392	1.00	0.94
0.45	0.555	0.275	0.301	0.392	1.00	0.92

Table 5 shows that source initialization has the smallest mean error at both shifts and beats pooling in every paired repetition. It beats target spectral initialization in 94 and 92 percent of repetitions, respectively, and random restarts in every repetition. At $\phi = 0.45$, its success rate is 1.00 and the pooling success rate is zero under the stated error threshold 0.35.

T.3 Validation Selection

This experiment compares the validation selector with the quadratic-risk oracle for the two phase candidates. The fitting sample has size 55, the source sample has size 320, and an independent target validation sample has size 100. The dimension, noise level, shifts, source estimator, and candidate-fitting algorithms equal those in the first phase experiment. The validation criterion is the target quartic loss. The oracle candidate minimizes squared error in the population Hessian norm, where $H = 4\theta^*_{\mathsf{T}}(\theta^*_{\mathsf{T}})^\top + 2I_p$.

Table 6: Validation selection over 100 repetitions at each design shift.

ρ	Selected	Oracle	Source	Pooled	Regret	Correct
0.35	0.1317	0.1233	0.1861	0.1861	0.0084	0.80
0.55	0.1563	0.1542	0.2298	0.4280	0.0021	0.97
0.75	0.1515	0.1424	0.1637	0.7561	0.0091	0.97

Table 6 reports mean Hessian-norm risk. Across the three reported shifts, the selector has mean risk 0.1465, compared with 0.1400 for the infeasible oracle, 0.1932 for always using source initialization, and 0.4567 for always pooling. The correct-choice rate is 0.9133 and the mean regret is 0.0065. This experiment evaluates the candidate-selection rule. It does not verify a numerical value of the uniform approximation tolerance γ_n in Proposition 10. Proposition 23 gives a separate conditional bound for the implemented quartic-loss comparison.

T.4 Population Loss and Validation Size

Quartic validation compares population prediction losses. We evaluate this criterion directly to distinguish its oracle from the quadratic-risk oracle in Table 6. For each of the same 100 independent fitting repetitions at each shift, we keep both fitted candidates fixed and use nested validation samples, independent of the fitting data, with sizes $n_v \in \{50, 100, 400, 1600\}$. The sample of size 100 and the candidates coincide with those used in Table 6.

We calculate each candidate's population excess loss $F_{\mathsf{T}}(u) - F_{\mathsf{T}}(\theta^*_{\mathsf{T}})$ from the Gaussian fourth-moment identity, without estimating it on the validation sample. Equation (20) relates this loss to the quadratic risk. The oracle in Table 7 minimizes population excess loss over the two fitted candidates. Regret is the selected loss minus this minimum. The Monte Carlo standard error, denoted by MCSE, is the standard deviation of the 100 paired regrets divided by $\sqrt{100}$.

Table 7: Population quartic loss of phase-retrieval validation.

ρ	n_v	Source init.	Pooled	Selected	Oracle	Regret MCSE
0.35	50	0.0917	0.0862	0.0661	0.0605	0.0013
0.35	100	0.0917	0.0862	0.0634	0.0605	0.0007
0.35	400	0.0917	0.0862	0.0613	0.0605	0.0003
0.35	1600	0.0917	0.0862	0.0610	0.0605	0.0003
0.55	50	0.1033	0.1819	0.0811	0.0768	0.0020
0.55	100	0.1033	0.1819	0.0768	0.0768	0.0000
0.55	400	0.1033	0.1819	0.0769	0.0768	0.0001
0.55	1600	0.1033	0.1819	0.0768	0.0768	0.0000
0.75	50	0.0760	0.2886	0.0733	0.0716	0.0014
0.75	100	0.0760	0.2886	0.0719	0.0716	0.0003
0.75	400	0.0760	0.2886	0.0717	0.0716	0.0001
0.75	1600	0.0760	0.2886	0.0717	0.0716	0.0001

Table 7 shows that validation has smaller mean prediction loss than either fixed choice at each reported shift and validation size. Averaged equally over the three shifts, the mean loss at $n_v = 100$ is 0.07072 for selection, compared with 0.06964 for the population-loss oracle, 0.09035 for source initialization, and 0.18558 for pooling. The mean regret decreases from 0.00384 at $n_v = 50$ to 0.00020 at $n_v = 1600$. This comparison measures selection under the prediction loss estimated by validation.

T.5 Low-Rank Matrix Completion

The larger nonlinear experiment considers an 80×80 positive semidefinite matrix of rank 5. Let $Q, V \in \mathbb{R}^{80\times 5}$ have orthonormal columns and satisfy $Q^\top V = 0$, and let $D \in \mathbb{R}^{5\times 5}$ be diagonal with entries equally spaced from 1 to 0.5. The target and source matrices are

$$M_{\mathsf{T}}^* = QDQ^\top, \qquad M_{\mathsf{S}}^* = (\cos\phi\, Q + \sin\phi\, V)D(\cos\phi\, Q + \sin\phi\, V)^\top.$$

Each response is a uniformly sampled matrix entry with independent Gaussian noise of standard deviation 0.01. The target sample contains 1,600 entries and the source sample contains 12,000 entries. The normalized source-target Frobenius distances at $\phi = 0.25, 0.45$ are $0.3499, 0.6151$.

The factorized target criterion minimizes squared entry error over matrices $UU^\top$. A scaled spectral estimator followed by 30 source iterations produces the source estimator. Source initialization, target spectral initialization, and pooling each receive 30 target iterations. The random comparison receives three 30-iteration target runs and is selected by target training loss. The pooled comparison minimizes the combined source-target criterion. Success means normalized target matrix error at most 0.35.

Table 8: Low-rank matrix-completion experiment over 100 repetitions.

ϕ	Method	Mean	Median	90% q.	Success
0.25	pooled data	0.3196	0.3199	0.3243	1.00
0.25	source initialization	0.2727	0.2731	0.2883	1.00
0.25	target spectral	0.3043	0.2861	0.3518	0.89
0.25	target random restart	0.3920	0.3856	0.4578	0.26
0.45	pooled data	0.5549	0.5549	0.5602	0.00
0.45	source initialization	0.2754	0.2758	0.2885	1.00
0.45	target spectral	0.3010	0.2895	0.3175	0.92
0.45	target random restart	0.3916	0.3902	0.4805	0.23

Table 8 shows that source initialization has the smallest mean error at $\phi = 0.25$ and 0.45. At $\phi = 0.45$, source initialization succeeds in every repetition and pooling fails in every repetition. The paired comparisons in Appendix Y show that source initialization beats target spectral initialization in 94 and 92 percent of repetitions, respectively, and random restarts in every repetition. The random comparison receives three times as many target iterations as each other method.

T.6 Nonnegative Matrix Factorization of Digits

The public handwritten-digits data from Pedregosa et al. (2011) provide a nonconvex factorization example with a natural source-target split. The source task consists of digits 0–4. The target task consists of digits 5–9. For each split, 30 target images are used for target training and the remaining target images are used for testing. The estimator is a rank-16 nonnegative matrix factorization (NMF) dictionary (Lee and Seung, 1999; 2000). The source dictionary is used only as the initial dictionary for fitting the target NMF. The baselines use one target random start, target-only NNDSVDa, eight random starts selected by target training loss, and pooled source-target fitting. We measure test reconstruction by root mean squared error (RMSE). The source dictionary is estimated from 400 source images within each split before the target fits. Each run receives at most 200 iterations. NNDSVDa replaces zero entries after nonnegative double singular value decomposition by the training-data mean. The eight-start comparison receives at most 1,600 target iterations in total.

Table 9: Digits NMF real-data analysis with 100 target splits, reporting test reconstruction RMSE and total target-stage iteration counts. Numbers in parentheses represent standard deviations.

Method	Mean	Median	90% quantile	Target iter.
source init	0.1503 (0.0042)	0.1505	0.1550	85.48
target random	0.1523 (0.0039)	0.1523	0.1570	200.00
Target NNDSVDa	0.1542 (0.0042)	0.1547	0.1593	200.00
Target 8 random starts	0.1535 (0.0040)	0.1536	0.1588	1599.37
pooled dictionary	0.1592 (0.0038)	0.1591	0.1644	200.00

Table 9 shows lower test RMSE for source initialization than all four comparators under the stated fitting budgets. The advantage persists against target NNDSVDa and eight random starts, although the latter receives more target iterations. The source fit has a separate 500-iteration cap. Target NNDSVDa, the single random fit, and pooling reach their 200-iteration caps on every split. The result does not compare fully converged stationary points or total computation costs.

U Numerical Evaluation of Phase-Retrieval Inference

We compare uncertainty about the target signal after source initialization with uncertainty about the parameter estimated by pooling. The distinction is statistical: the two procedures use different estimating equations, even when their optimization errors are small. The experiments use the estimators in Proposition 19, Proposition 21, and Theorem 3. We measure coverage at both population parameters and report target error separately.

U.1 Common design and reported quantities

Let e_j be the jth coordinate vector in $\mathbb{R}^4$. We set

$$\theta^*_{\mathsf{T}} = e_1, \qquad \theta^*_{\mathsf{S}} = (1-\delta^2/2)e_1 + \delta(1-\delta^2/4)^{1/2}e_2.$$

Both signals have norm one and their sign-invariant distance is exactly δ. All sensing vectors are independent standard Gaussian vectors. Noise is independent of the design and across observations and studies. Every configuration uses 200 independent repetitions. Within a repetition, all three procedures use the same source and target observations.

Source initialization uses the raw source spectral estimate and then refines the target criterion. The oracle target start uses θ^*_{T} and serves as a benchmark for the local target solution. Pooling uses the sample-size-weighted criterion and its raw spectral start. In Tables 10–16, these procedures are denoted by Source, Oracle, and Pooled.

All refinements use step $1/8$ and (14). The sample size in the stopping rule is n_{T} for target refinement and $n_{\mathsf{T}}+n_{\mathsf{S}}$ for pooling. We do not truncate responses, modify the spectral scale, or use backtracking in these experiments. The target confidence region is (16) with $\beta = 0.05$. Pooling uses the within-study centered covariance and confidence region defined in Appendix Q.

Coverage is the fraction of repetitions in which a region contains its specified parameter. Pool T reports coverage of θ^*_{T}; Pool P reports coverage of θ_{w_N}. All other coverage columns concern θ^*_{T}. Target error is $d_{\pm}(\widehat{u}, \theta^*_{\mathsf{T}})$.

For a reported proportion $\widehat{q}$, the Monte Carlo standard error is $\{\widehat{q}(1-\widehat{q})/199\}^{1/2}$. For a reported mean, it is the sample standard deviation divided by $\sqrt{200}$. Figure error bars are 1.96 times the corresponding Monte Carlo standard error. They describe simulation uncertainty, not confidence intervals for the signal.

U.2 Target coverage and estimation error

We use $n_{\mathsf{T}} \in \{512, 2048, 8192\}$, $n_{\mathsf{S}} = \lceil 64\sqrt{n_{\mathsf{T}}} \rceil$ or $4n_{\mathsf{T}}$, and $\delta \in \{0.04, 0.10\}$. Source and target noise are either Gaussian with variance 0.35^2 or Student-t_3 divided by $\sqrt{3}$ and multiplied by 0.35. The second noise distribution has the same variance and no fourth moment. We include all 24 configurations and every procedure in each configuration.

Table 10: Signal-region coverage under Gaussian noise.

n_{T}	n_{S}	δ	Source	Oracle	Pool T	Pool P
512	1449	0.04	0.940	0.940	0.000	0.965
512	1449	0.10	0.930	0.940	0.000	0.930
512	2048	0.04	0.940	0.945	0.000	0.900
512	2048	0.10	0.935	0.935	0.000	0.905
2048	2897	0.04	0.965	0.965	0.000	0.945
2048	2897	0.10	0.950	0.950	0.000	0.935
2048	8192	0.04	0.970	0.970	0.000	0.950
2048	8192	0.10	0.925	0.925	0.000	0.935
8192	5793	0.04	0.950	0.950	0.000	0.945
8192	5793	0.10	0.945	0.945	0.000	0.945
8192	32768	0.04	0.940	0.940	0.000	0.920
8192	32768	0.10	0.925	0.925	0.000	0.950

Table 11: Signal-region coverage under Student-t_3 noise.

n_{T}	n_{S}	δ	Source	Oracle	Pool T	Pool P
512	1449	0.04	0.920	0.925	0.000	0.960
512	1449	0.10	0.915	0.915	0.000	0.970
512	2048	0.04	0.940	0.940	0.000	0.950
512	2048	0.10	0.920	0.920	0.000	0.960
2048	2897	0.04	0.960	0.960	0.000	0.925
2048	2897	0.10	0.925	0.925	0.000	0.955
2048	8192	0.04	0.910	0.915	0.000	0.945
2048	8192	0.10	0.925	0.925	0.000	0.945
8192	5793	0.04	0.950	0.950	0.000	0.930
8192	5793	0.10	0.945	0.945	0.000	0.950
8192	32768	0.04	0.960	0.960	0.000	0.980
8192	32768	0.10	0.960	0.960	0.000	0.965

Tables 10 and 11 show that source initialization gives nearly the same target coverage as the oracle start. Its coverage ranges from 0.910 to 0.970, while oracle coverage ranges from 0.915 to 0.970. The small-sample deviations from 0.95 also occur for the oracle procedure. Thus the comparison concerns preservation of target inference after source initialization; Proposition 19 does not assert exact finite-sample coverage.

Pooling has zero target coverage in all 24 configurations, while coverage of its own population parameter ranges from 0.900 to 0.980. These calculations use the covariance of the pooled estimating equation. The difference cannot be explained by applying a target covariance formula to a pooled estimator. Proposition 20 identifies the nonzero parameter displacement responsible for this comparison.

In Tables 12–13, S denotes $n_{\mathsf{S}} = \lceil 64\sqrt{n_{\mathsf{T}}} \rceil$, L denotes $n_{\mathsf{S}} = 4n_{\mathsf{T}}$, and G denotes Gaussian noise.

Table 12: Signed target error under finite-variance noise. Numbers in parentheses represent standard deviations.

n_{T}	Size	δ	Noise	Source	Oracle	Pooled
512	S	0.04	G	0.0134 (0.0051)	0.0133 (0.0051)	0.0300 (0.0036)
512	S	0.04	t_3	0.0132 (0.0059)	0.0131 (0.0059)	0.0301 (0.0039)
512	S	0.10	G	0.0137 (0.0055)	0.0136 (0.0055)	0.0737 (0.0049)
512	S	0.10	t_3	0.0141 (0.0103)	0.0140 (0.0103)	0.0747 (0.0053)
512	L	0.04	G	0.0138 (0.0052)	0.0137 (0.0052)	0.0324 (0.0040)
512	L	0.04	t_3	0.0129 (0.0055)	0.0128 (0.0056)	0.0327 (0.0034)
512	L	0.10	G	0.0133 (0.0050)	0.0132 (0.0050)	0.0804 (0.0043)
512	L	0.10	t_3	0.0128 (0.0061)	0.0127 (0.0061)	0.0796 (0.0040)
2048	S	0.04	G	0.0064 (0.0025)	0.0064 (0.0025)	0.0239 (0.0026)
2048	S	0.04	t_3	0.0060 (0.0024)	0.0060 (0.0024)	0.0239 (0.0029)
2048	S	0.10	G	0.0067 (0.0026)	0.0067 (0.0026)	0.0589 (0.0033)
2048	S	0.10	t_3	0.0064 (0.0025)	0.0064 (0.0025)	0.0589 (0.0030)
2048	L	0.04	G	0.0061 (0.0024)	0.0061 (0.0024)	0.0321 (0.0018)
2048	L	0.04	t_3	0.0068 (0.0028)	0.0068 (0.0028)	0.0321 (0.0019)
2048	L	0.10	G	0.0067 (0.0027)	0.0067 (0.0027)	0.0801 (0.0022)
2048	L	0.10	t_3	0.0064 (0.0026)	0.0064 (0.0026)	0.0800 (0.0019)
8192	S	0.04	G	0.0033 (0.0013)	0.0033 (0.0013)	0.0166 (0.0015)
8192	S	0.04	t_3	0.0033 (0.0012)	0.0033 (0.0012)	0.0167 (0.0017)
8192	S	0.10	G	0.0032 (0.0012)	0.0032 (0.0012)	0.0414 (0.0017)
8192	S	0.10	t_3	0.0035 (0.0013)	0.0035 (0.0013)	0.0413 (0.0020)
8192	L	0.04	G	0.0034 (0.0013)	0.0034 (0.0013)	0.0319 (0.0009)
8192	L	0.04	t_3	0.0033 (0.0012)	0.0033 (0.0012)	0.0321 (0.0009)
8192	L	0.10	G	0.0034 (0.0012)	0.0034 (0.0012)	0.0800 (0.0011)
8192	L	0.10	t_3	0.0031 (0.0012)	0.0031 (0.0012)	0.0801 (0.0010)

Table 12 shows that the source-initialized target error decreases with target sample size and remains close to the oracle error. Pooled error is larger in every configuration. With linear source growth, its leading error is the fixed population displacement. With square-root source growth, that displacement decreases, but need not be negligible relative to the target confidence-region radius. The next experiment examines smaller coefficients of square-root source growth, for which nonzero target coverage can be observed.

U.3 Optimization error at the target scale

The stopping rule must make numerical error negligible relative to statistical error. We compare the source-initialized output and the oracle target output from the same target sample. Across the eight source-size, shift, and noise configurations at each target size, the mean of $\sqrt{n_{\mathsf{T}}} d_{\pm}(\widehat{u}_{\mathrm{Source}}, \widehat{u}_{\mathrm{Oracle}})$ is 0.00427, 0.00168, and 0.00069 at $n_{\mathsf{T}} = 512, 2048, 8192$, respectively. The corresponding maximum values over all 1,600 repetitions at each size are 0.01044, 0.00353, and 0.00123. These differences are small on the target statistical scale and decrease with sample size, as required by the stopping argument in Proposition 19.

Every one of the 14,400 fits in this experiment reaches the gradient threshold before the iteration cap, and every confidence matrix is positive definite. Meeting a gradient threshold alone does not certify the target basin, which is why we compare the output with the oracle target output.

U.4 Intervals for prediction intensities

The sign of a recovered signal has no effect on its predicted intensity. We evaluate (18) at e_1, e_2, and $x_o = (e_1 + e_2)/\sqrt{2}$. Their true intensities are 1, 0, and $1/2$. Let $W(x)$ denote the width of the interval at x. Joint coverage means that all three true intensities are contained in their intervals in the same repetition.

Table 13 shows simultaneous inference for an active direction, a direction orthogonal to the signal, and an oblique direction. The projection construction protects all directions on the signal-region

Table 13: Joint intensity coverage and mean interval widths. Numbers in parentheses represent standard deviations.

n_T	Size	δ	Noise	Coverage	$10^3 W(e_1)$	$10^4 W(e_2)$	$10^3 W(x_o)$
512	S	0.04	G	0.990	54.72 (1.79)	9.20 (3.25)	54.73 (1.80)
512	S	0.04	t_3	0.985	53.97 (9.26)	9.05 (4.67)	53.95 (9.29)
512	S	0.10	G	0.990	54.89 (1.75)	9.41 (3.45)	54.88 (1.75)
512	S	0.10	t_3	0.995	57.35 (46.40)	14.85 (82.82)	57.37 (46.45)
512	L	0.04	G	0.995	54.93 (1.79)	9.66 (3.40)	54.90 (1.81)
512	L	0.04	t_3	0.990	54.25 (8.98)	9.31 (5.64)	54.24 (8.99)
512	L	0.10	G	0.995	54.60 (1.80)	9.13 (3.07)	54.62 (1.82)
512	L	0.10	t_3	0.985	53.27 (7.18)	8.82 (4.37)	53.26 (7.19)
2048	S	0.04	G	1.000	27.48 (0.38)	2.35 (0.80)	27.48 (0.39)
2048	S	0.04	t_3	0.990	26.93 (2.16)	2.19 (0.78)	26.93 (2.16)
2048	S	0.10	G	0.995	27.47 (0.43)	2.34 (0.82)	27.46 (0.45)
2048	S	0.10	t_3	0.990	27.07 (2.43)	2.17 (0.79)	27.07 (2.42)
2048	L	0.04	G	1.000	27.49 (0.41)	2.25 (0.69)	27.49 (0.42)
2048	L	0.04	t_3	0.985	26.88 (2.00)	2.24 (0.87)	26.88 (1.99)
2048	L	0.10	G	0.985	27.46 (0.39)	2.30 (0.78)	27.45 (0.40)
2048	L	0.10	t_3	0.985	27.01 (2.15)	2.27 (0.79)	27.01 (2.14)
8192	S	0.04	G	1.000	13.75 (0.10)	0.58 (0.17)	13.75 (0.10)
8192	S	0.04	t_3	1.000	13.83 (2.21)	0.62 (0.37)	13.83 (2.21)
8192	S	0.10	G	1.000	13.74 (0.11)	0.56 (0.18)	13.74 (0.11)
8192	S	0.10	t_3	0.995	13.71 (1.15)	0.61 (0.23)	13.71 (1.15)
8192	L	0.04	G	0.995	13.76 (0.10)	0.57 (0.19)	13.76 (0.10)
8192	L	0.04	t_3	0.990	13.76 (0.92)	0.59 (0.19)	13.76 (0.92)
8192	L	0.10	G	0.995	13.74 (0.11)	0.61 (0.19)	13.74 (0.11)
8192	L	0.10	t_3	0.995	13.78 (2.02)	0.56 (0.28)	13.78 (2.02)

coverage event, and its three-direction coverage can exceed the nominal signal-region coverage. The zero-intensity intervals remain nonnegative. They do not use a first-order variance that would be zero at the true signal.

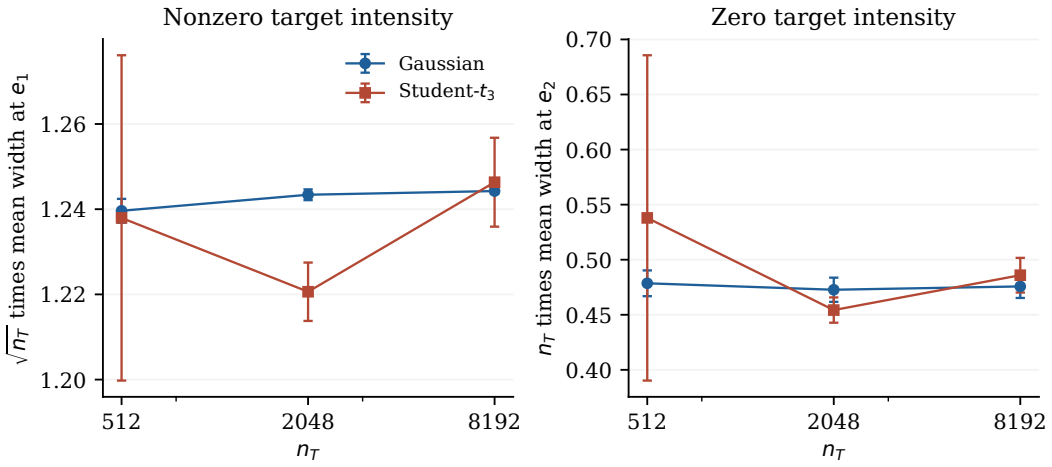

Figure 2: Scaled intensity-interval widths under Gaussian and Student-t_3 noise.

Figure 2 compares the width scales in (19). Each point averages the four source-size and shift configurations under one target-noise distribution, giving 800 independent repetitions. Multiplication by $\sqrt{n_T}$ gives nearly constant mean width at e_1, while multiplication by n_T gives nearly constant mean width at e_2. Under Gaussian noise, the mean widths at e_1 are 0.0548, 0.0275, and 0.0137, and those at e_2 are 0.00094, 0.00023, and 0.00006. The wider Monte Carlo error bar for Student-t_3 noise at the smallest sample size arises from variability of the residual variance estimate; no fourth noise moment is imposed in Proposition 19.

U.5 Coverage at square-root source sample sizes

We fix $\delta = 0.10$, use Gaussian noise with variance 0.35^2, and set $n_{\mathsf{S}} = \lceil a\sqrt{n_{\mathsf{T}}} \rceil$, with $a \in \{0.5, 1, 2, 4, 8\}$ and $n_{\mathsf{T}} \in \{4096, 16384\}$. These ten configurations compare empirical coverage with the noncentral chi-square limit in Theorem 3. The signal separation is fixed; only the number of observations changes.

Table 14: Signal-region coverage at square-root source sample sizes.

n_{T}	a	Source	Oracle	Pool T	Pool P	Limit
4096	0.5	0.955	0.960	0.950	0.955	0.946
4096	1	0.965	0.965	0.940	0.970	0.933
4096	2	0.935	0.935	0.860	0.945	0.876
4096	4	0.910	0.910	0.640	0.930	0.591
4096	8	0.965	0.965	0.140	0.935	0.030
16384	0.5	0.950	0.950	0.940	0.945	0.946
16384	1	0.940	0.940	0.925	0.935	0.933
16384	2	0.960	0.960	0.870	0.960	0.876
16384	4	0.930	0.930	0.630	0.915	0.591
16384	8	0.970	0.970	0.060	0.960	0.030

Table 14 shows that the source-initialized target coverage remains close to its oracle counterpart as the source sample increases. At $n_{\mathsf{T}} = 16384$, pooled target coverage decreases from 0.940 at $a = 0.5$ to 0.060 at $a = 8$. Coverage of the pooled population parameter is 0.945 and 0.960 in these two configurations. The limiting target coverage is 0.946 and 0.030, respectively. The limit does not give exact finite-sample coverage: at $a = 8$, the pooled target coverage is 0.140 when $n_{\mathsf{T}} = 4096$ and 0.060 when $n_{\mathsf{T}} = 16384$. The larger sample moves the observed coverage toward the limiting value.

Figure 1 shows the continuous coverage curve from Theorem 3 and the simulated proportions. The source-initialized procedure continues to use the target estimating equation. Its uncertainty is centered at the target signal, even as the source sample improves the initial estimate. Pooling uses the additional observations in the estimating equation and uses the source population displacement at the target statistical scale.

Table 15: Target error at square-root source sample sizes. Numbers in parentheses represent standard deviations.

n_{T}	a	Source	Oracle	Pooled
4096	0.5	0.0046 (0.0017)	0.0046 (0.0017)	0.0047 (0.0018)
4096	1	0.0047 (0.0017)	0.0047 (0.0017)	0.0049 (0.0018)
4096	2	0.0048 (0.0019)	0.0048 (0.0019)	0.0056 (0.0022)
4096	4	0.0048 (0.0018)	0.0048 (0.0018)	0.0074 (0.0026)
4096	8	0.0047 (0.0018)	0.0047 (0.0018)	0.0116 (0.0027)
16384	0.5	0.0023 (0.0009)	0.0023 (0.0009)	0.0024 (0.0009)
16384	1	0.0023 (0.0009)	0.0023 (0.0009)	0.0024 (0.0009)
16384	2	0.0023 (0.0008)	0.0023 (0.0008)	0.0026 (0.0011)
16384	4	0.0023 (0.0009)	0.0023 (0.0009)	0.0037 (0.0013)
16384	8	0.0023 (0.0008)	0.0022 (0.0008)	0.0064 (0.0014)

Table 15 reports target error for the same repetitions. The source error remains close to the oracle error, while pooled error increases with a.

U.6 A source distribution with infinite variance

The source moment condition in Proposition 19 is weaker than the target moment condition. We examine this distinction with $\delta = 0.04$, $n_{\mathsf{T}} \in \{2048, 8192, 32768\}$, and $n_{\mathsf{S}} = \lceil 64\sqrt{n_{\mathsf{T}}} \rceil$. Source noise is either Gaussian with standard deviation 0.35 or sZ, where Z has a Student-$t_{1.5}$ distribution and $s = 0.35\sqrt{2/\pi}/\mathbb{E}|Z| = 0.1365966766$. This choice matches the first absolute source-noise

moment. The Student-$t_{1.5}$ variance is infinite, and s is a scale parameter. Target noise is Gaussian or variance-matched Student-t_3, as above.

Table 16: Signal-region coverage with finite or infinite source-noise variance.

n_T	Source noise	Target noise	Source	Oracle	Pool T	Pool P
2048	G	G	0.935	0.940	0.000	0.960
2048	G	t_3	0.955	0.955	0.000	0.940
2048	$t_{1.5}$	G	0.950	0.955	0.040	0.960
2048	$t_{1.5}$	t_3	0.965	0.965	0.045	0.985
8192	G	G	0.950	0.950	0.000	0.925
8192	G	t_3	0.965	0.965	0.000	0.970
8192	$t_{1.5}$	G	0.930	0.935	0.015	0.925
8192	$t_{1.5}$	t_3	0.955	0.955	0.015	0.970
32768	G	G	0.940	0.940	0.000	0.960
32768	G	t_3	0.945	0.945	0.000	0.930
32768	$t_{1.5}$	G	0.980	0.980	0.000	0.985
32768	$t_{1.5}$	t_3	0.945	0.945	0.000	0.970

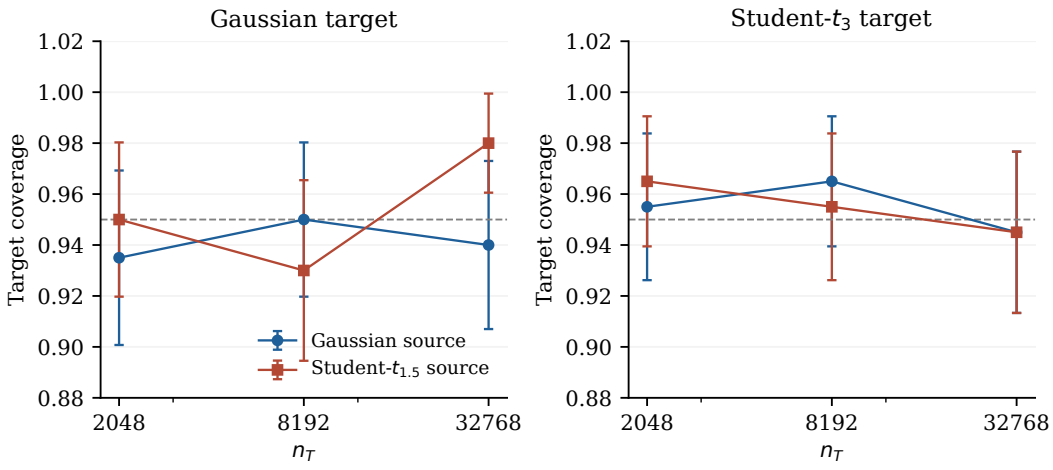

Figure 3: Target coverage under finite and infinite source-noise variances.

Table 16 and Figure 3 show target coverage between 0.930 and 0.980 under Student-$t_{1.5}$ source noise. The oracle target coverage in these configurations lies between 0.935 and 0.980. No source variance estimate enters the target confidence region. The comparison supports the different source and target roles in Proposition 19; it does not require the source estimate to have a normal limit or a target-scale convergence rate.

U.7 Evaluation of the implemented estimating equations

The three inference experiments use the same quartic criterion and fixed update. Computation uses its exact finite-dimensional sufficient statistics. For observations (X_i, Y_i), define $q_i = \mathrm{vec}(X_i X_i^\top)$, where vec stacks the matrix entries in the same fixed order, and set

$$T_n = \frac{1}{n}\sum_i q_i q_i^\top, \qquad B_n = \frac{1}{n}\sum_i Y_i X_i X_i^\top, \qquad c_n = \frac{1}{n}\sum_i Y_i^2.$$

For $q(u) = \mathrm{vec}(uu^\top)$, let $M_n(u)$ be the matrix whose vectorization is $T_n q(u)$. Direct differentiation gives

$$F_n(u) = \{q(u)^\top T_n q(u) - 2u^\top B_n u + c_n\}/4,$$
$$\nabla F_n(u) = M_n(u)u - B_n u, \qquad \nabla^2 F_n(u) = 3M_n(u) - B_n.$$

These identities use all observations and introduce no approximation to the criterion, gradient, or Hessian. The pooled statistics are the sample-size-weighted averages of the source and target statistics. The within-study score covariances used for pooling are evaluated from the observations at the reported estimate.

The numerical procedure reports the first iterate meeting the gradient threshold, or the capped iterate if the threshold is not met. Undefined iterates and undefined confidence matrices are recorded as failed coverage. All 27,600 fits in the three experiments have finite values, reach the gradient threshold, and give positive definite confidence matrices. This statement is not a claim of global recovery: in one Student-$t_{1.5}$ source repetition the raw source response mean is negative, the spectral start is zero, and the fit remains at the zero stationary point. The uncertainty comparison is tied to the population parameter specified for each estimating equation.

V Target Checks and Inference Beyond Close Source Signals

We examine whether the added guarantees apply to their computed estimates. The first study tests source initialization outside the local target ball. The second tests the target check when the source noise has no mean. The third tests estimation of the design covariance and signal scale. All 16 configurations use 200 independent replications, with $p = 4$, $n \in \{1024, 4096\}$, and $n_{\mathsf{S}} = 4n$. Every fit uses step $1/8$, the gradient threshold and cap in (14), and the stated covariance formula. The stopping rule uses sample size n for target fitting and $n + n_{\mathsf{S}}$ for pooling. Within each configuration, competing fits use the same target observations and, when needed, the same source observations. A failed raw fit contributes zero to target coverage. For the checked procedure, we also record whether the reported region is $\mathbb{R}^p$. This never occurred in these experiments.

V.1 Nonorthogonal source directions

We set the target signal to the first coordinate vector and choose unit source signals with target inner products 0.995, 0.8 and 0.2. Their sign-invariant distances are 0.1, $\sqrt{0.4}$ and $\sqrt{1.6}$, respectively. Only the first is inside the sufficient radius $1/8$. Both measurement designs are standard Gaussian. Target noise is independent normal with standard deviation 0.35. Source noise is either the same normal noise or 0.35 times a Student variable with 1.5 degrees of freedom. The latter has a finite first absolute moment and infinite variance. The raw spectral source estimate is not stabilized.

Table 17 compares 95% target confidence regions. The column Source uses the raw source start, and Checked applies (9) and the target restart. Source coverage lies between 0.925 and 0.970, including the two directions outside the local ball. This tests the nonlocal conclusion of Theorem 8. The check restarts one of the 2400 source fits, where the raw spectral scale is zero, and all checked fits are accepted after at most one restart. Pooling has target coverage at most 0.010, despite a within-study centered sandwich covariance. The close-signal Student case tests the infinite-variance conclusion of Corollary 6. The two more distant Student cases are empirical comparisons beyond that corollary's small-distance condition. We do not claim nominal coverage of the pooled parameter when the source variance is infinite.

Table 17: Target coverage with nonorthogonal source signals.

n	Alignment	Source noise	Source	Checked	Pooled
1024	0.995	Normal	0.935	0.935	0.000
1024	0.995	$t_{1.5}$	0.940	0.940	0.010
1024	0.800	Normal	0.970	0.970	0.000
1024	0.800	$t_{1.5}$	0.950	0.950	0.000
1024	0.200	Normal	0.945	0.945	0.000
1024	0.200	$t_{1.5}$	0.945	0.945	0.000
4096	0.995	Normal	0.925	0.925	0.000
4096	0.995	$t_{1.5}$	0.955	0.955	0.000
4096	0.800	Normal	0.940	0.940	0.000
4096	0.800	$t_{1.5}$	0.940	0.940	0.000
4096	0.200	Normal	0.935	0.935	0.000
4096	0.200	$t_{1.5}$	0.955	0.960	0.000

These results concern inference after the gradient threshold is reached. Source initialization does not lower the target estimator's first-order variance, in agreement with Corollary 4.

V.2 A SOURCE DISTRIBUTION WITHOUT A MEAN

We keep the alignment at 0.8 and replace the source noise by 0.35 times a Student variable with 0.8 degrees of freedom. This symmetric distribution has no finite mean and is outside the model conditions of Theorem 8. The target observations still satisfy the conditions of Theorem 4. The experiment tests whether the observed check can prevent unreliable source fits from determining inference.

Table 18 shows the resulting improvement. Raw source-initialized coverage is 0.410 and 0.345. After checking and restarting, coverage is 0.920 and 0.945, close to the nominal level. The restart proportions are 0.560 and 0.640. All checked regions are finite ellipsoid unions; the improved coverage is not obtained by reporting $\mathbb{R}^p$. Every divergent fit and zero start is included in these proportions. The pooled population criterion need not exist in this design, and we report only its coverage of the target signal.

Table 18: Target coverage with source noise without a finite mean.

n	Source	Checked	Pooled	Restart
1024	0.410	0.920	0.060	0.560
4096	0.345	0.945	0.090	0.640

V.3 UNKNOWN COVARIANCE, SIGNAL SCALE AND HETEROSKEDASTIC NOISE

We next evaluate the procedure of Theorem 9. Let the Gaussian design covariance have entries $\Sigma_{jk} = 0.6^{|j-k|}$. The target and source signals equal $1.5\Sigma^{-1/2}e_1$ and $1.5\Sigma^{-1/2}(0.8e_1 + 0.6e_2)$, where e_j is the jth coordinate vector. Their squared signal scale is 2.25. The source noise is independent normal with standard deviation 0.35. For $Z = \Sigma^{-1/2}X$, target noise equals $0.35(1 + 0.5Z_2^2)\zeta/\sqrt{3}$, where ζ has a Student t_3 distribution and is independent of X. Its conditional mean is zero and its conditional variance depends on the design. The procedure estimates the covariance and scale from the same target observations used for fitting, and forms the sandwich matrix in the original coordinates. The source start is computed by the analogous source normalization and spectral rule. The pooled comparison normalizes the combined observations and centers its scores separately within each study.

Table 19 gives coverage 0.945 and 0.950 for the checked source fit. Pooled target coverage is zero at both sample sizes. Thus estimating the normalization preserves target inference in these settings, including heteroskedastic noise. This comparison evaluates the normalized procedure, not the unit-scale residual variance formula used in the earlier experiments. For checked coverage in all three studies, Monte Carlo standard errors are at most 0.020.

Table 19: Target coverage with estimated normalization and heteroskedastic noise.

n	Checked source	Pooled
1024	0.945	0.000
4096	0.950	0.000

W NUMERICAL IMPLEMENTATION DETAILS

The phase-retrieval simulation uses backtracking gradient descent on the target quartic loss with random seed 20260709. The initial step size is 0.08, the iteration cap is 1500, the gradient tolerance is 10^{-7}, and each update allows at most 18 backtracking attempts. We also stop when no step is accepted or when the absolute loss change is at most 10^{-7} times the larger of one and the absolute preceding loss. A step is accepted only when the candidate norm is at most 10 and the finite loss

is at most the preceding loss plus 10^{-10}. The fixed step $1/8$ in Theorem 7 is not the algorithm used in the simulation. The source initializer is the spectral estimator: the leading eigenvector of $n^{-1}\sum_i Y_i X_i X_i^\top$, scaled by the square root of the average positive part of the responses. The target-random baseline uses eight random restarts and selects the run with the lowest target training loss. The pooled baseline fits the phase-retrieval objective after stacking source and target samples. Proposition 18 analyzes the consistent scale based on the untruncated response mean; the positive-part scale is used only in the experiment. For centered noise, $\mathbb{E}\max(Y,0) = 1 + \mathbb{E}\max(-Y,0)$, which can exceed one. The two scales need not have the same probability limit. The real-data analysis uses random seed 20260711 and the implementation of nonnegative matrix factorization in Pedregosa et al. (2011). The source dictionary is estimated once within each split with an iteration cap of 500. The transferred NMF method initializes the target dictionary at the source dictionary and then fits only the target training images; the pooled comparator fits one dictionary to the stacked source and target training images and evaluates it on the target test images. All target NMF fits use the same 200-iteration budget, and the results describe fixed-budget performance; they do not claim that every comparator reached its limiting stationary point.

The validation experiment uses seed 20260932 and the same backtracking candidate fits as the first phase experiment. Its validation sample is independent of the fitting and source samples.

The matrix-completion experiment uses seed 20260933. Entry observations are aggregated into count and response-sum matrices, and the sum of squared responses is included. This preserves the squared-error loss and gradient exactly. Source estimation and every target candidate use an initial step 0.2 with backtracking and a 30-iteration budget. The random baseline performs three target restarts and its reported iteration count is the total across all restarts. The production environment used Python 3.9.6, NumPy 2.0.2, pandas 2.3.3, and scikit-learn 1.5.2.

X Expanded Numerical Protocol

Within each numerical repetition, the competing target refits use the same target observations. Their differences measure the effects of the starting point or the inclusion of source observations under the specified optimization budget. The protocols specify source fitting, target updates and validation. The tables compare target accuracy and validation selection.

X.1 Phase-retrieval simulation protocol

For each repetition and each design shift $\rho \in \{0.35, 0.55, 0.75\}$, the script generates a unit-norm target vector θ_{T}^*. It then draws an orthogonal direction v and sets the source vector proportional to

$$\theta_{\mathsf{T}}^* + \rho v.$$

The source vector is renormalized, which gives actual sign-invariant distance $\delta(\rho) = \{2 - 2/(1 + \rho^2)^{1/2}\}^{1/2}$. Source and target samples are generated independently. The source spectral initializer is computed from the source sample, and all target refits use the same target sample for that repetition. Every gradient refit uses initial step size 0.08, at most 1500 iterations, gradient tolerance 10^{-7}, and at most 18 backtracking attempts per update. The acceptance and loss-change stopping rules are stated in Section W. These settings differ from the fixed step in Theorem 7, and the experiment is not presented as a finite-sample verification of that theorem.

The four reported methods are:

1. source initialization followed by target refitting;
2. target spectral initialization followed by target refitting;
3. eight random target restarts, selecting the lowest target training loss;
4. pooled source-target fitting.

Table 20 reports the error, success, and source-proxy summary. The script's random-restart iteration field records the cumulative iterations only through the restart that attained the selected loss, not the total cost of all eight restarts; we omit that noncomparable field.

Table 20: Phase-retrieval estimation error and success rates.

Design shift	Method	Mean	Median	90% q.	Success	Proxy
0.35	pooled data	0.2949	0.2951	0.3183	1.00	–
0.35	source init	0.2505	0.2272	0.3708	0.88	1.00
0.35	target random restart	0.3734	0.2451	0.7948	0.71	–
0.35	target spectral	0.9236	1.1072	1.5657	0.29	–
0.55	pooled data	0.4427	0.4416	0.4668	0.00	–
0.55	source init	0.2519	0.2389	0.3330	0.91	0.98
0.55	target random restart	0.4295	0.2682	1.0946	0.70	–
0.55	target spectral	0.8747	0.9867	1.5111	0.32	–
0.75	pooled data	0.5626	0.5668	0.5938	0.00	–
0.75	source init	0.2765	0.2355	0.4033	0.83	0.73
0.75	target random restart	0.3712	0.2559	0.8707	0.73	–
0.75	target spectral	0.9131	0.9838	1.5410	0.28	–

X.2 Sparse phase-retrieval protocol

The sparse phase-retrieval experiment uses base seed 20260923. For each shift and repetition, independent seed roles generate the signals, the source sample, a target sample of 600 observations and a random direction. The target samples of sizes 150 and 300 are the first rows of the target sample of size 600, which permits paired comparison across target sizes. The target support is drawn uniformly, its entries are $\pm k^{-1/2}$ with independent signs, and the source signal is $(\theta^*_{\mathsf{T}} + \rho v)/(1+\rho^2)^{1/2}$ for a unit vector v on the target support orthogonal to θ^*_{T}, with ρ chosen to give the sign-invariant distance δ. The sparse spectral start selects the k coordinates with the largest values of $n^{-1}\sum_i Y_i X_{ij}^2$, takes the leading eigenvector of the response-weighted second-moment matrix on these coordinates, and scales it by the square root of the mean response. The source estimate applies (4) to the source sample from its sparse spectral start. The random target start keeps the k largest entries of a standard normal vector and is scaled by the square root of the mean target response. Pooling uses the sparse spectral start and the updates on the combined sample, which weights each observation equally. Every refinement uses at most 500 updates and stops when the relative change of the iterate is at most 10^{-12}. A fit succeeds when $d_{\pm}(z, \theta^*_{\mathsf{T}}) \le 10^{-6}$. The recorded source-to-target distances describe the initial errors; they do not verify the unspecified sufficient radius in Corollary 3.

For the truncated-power comparator, we use Algorithm 1 of Cai, Li and You (2023). The moment matrix weights observation i by Y_i, and its truncated version keeps amplitude ratios $Y_i^{1/2}/\{n^{-1}\sum_j Y_j\}^{1/2}$ between $1/2$ and 10. The initial support uses the largest absolute entries in the column indexed by the largest diagonal entry of the untruncated moment matrix. We use 500 power iterations for each sparsity in $\{k, 2k, 5k, 10k, 50k\}$, project each result back to k coordinates, and apply the same amplitude refinement to each candidate. We select by target amplitude training loss without using the true signal. This comparator uses five candidate fits and a larger optimization budget than one source-initialized fit. It compares initializers under common refinement; it does not use the hard thresholding pursuit refinement in the numerical study of Cai, Li and You (2023).

X.3 Validation-selector protocol

The production seed is 20260932. Each repetition generates independent source, target-fitting, and target-validation samples. Source initialization and pooling are fitted before the validation sample is used. The selected candidate has the smaller validation quartic loss. The oracle candidate has the smaller population-Hessian squared error. The reported regret equals selected risk minus oracle risk and is nonnegative on every repetition. Table 6 reports the mean risk and correct-choice rate.

For Table 7, the fitting and source samples, the random seeds, and all fitting settings are the same as in Table 6. The 100 validation observations are followed by 1,500 independent target observations. The first 50, 100, 400, or 1,600 observations define the four validation samples. Both candidates are fixed across these comparisons, and each validation observation is independent of the fitting data. The population losses are calculated analytically from the fitted vectors and target parameter. Paired differences use the same candidate pair and validation sample. Pointwise Monte Carlo intervals use

the Student distribution with 99 degrees of freedom and quantify variation across independent fitting and validation repetitions.

X.4 MATRIX-COMPLETION PROTOCOL

The production seed is 20260933. For each repetition, a Gaussian 80×10 matrix is orthonormalized. Its first five columns define the target eigenspace and its last five columns define the orthogonal direction of the source shift. Target and source entry indices are sampled uniformly with replacement. Write $(i_k, j_k, Y_k)_{k=1}^n$ for the row index, column index, and response in a sample of size n. For a factor $U \in \mathbb{R}^{80\times 5}$, let $U_i \in \mathbb{R}^5$ be the transpose of its ith row. The factorized criterion is

$$\frac{80^2}{2n} \sum_{k=1}^{n} \{U_{i_k}^\top U_{j_k} - Y_k\}^2.$$

Repeated entries are aggregated into count and response-sum matrices, with the sum of squared responses included. This preserves the loss and gradient exactly. The source estimator uses spectral initialization and 30 source iterations. The target spectral, source-initialized, and pooled runs receive 30 iterations. The random comparison uses three 30-iteration runs and selects the smallest target training loss. All gradient runs use backtracking from step 0.2, with at most 24 attempts. A step must decrease the loss by at least 10^{-4} times the step size times the squared gradient norm. We stop when the gradient norm is at most 10^{-8}, no step is accepted, or the absolute loss change is at most 10^{-8} times the larger of one and the absolute preceding loss. Table 8 reports marginal results, and Table 23 reports paired differences.

X.5 DIGITS NMF PROTOCOL

The digits analysis uses the scikit-learn `load_digits` data set. Each 8×8 image is represented by a row of 64 pixel intensities. Pixel intensities are divided by 16 and increased by 10^{-6} to make the matrix strictly positive. The source task is digits 0–4, and the target task is digits 5–9. Each split selects 30 target training images by stratified sampling and uses all remaining target images for testing. Within each split, 400 source images are sampled without replacement. The rank is 16.

The source dictionary is initialized by nonnegative double singular value decomposition, with zero entries replaced by the mean of the source data. The source fit uses tolerance 10^{-4} and an iteration cap of 500. Each target fit uses tolerance 10^{-4} and an iteration cap of 200. All fits use coordinate descent with squared Frobenius loss. Target-only NNDSVDa uses the same singular-value construction, with zero entries replaced by the target-training mean. The pooled baseline uses the pooled-data mean. The random comparisons use one start or eight starts; the eight-start estimate minimizes target training Frobenius loss across its candidates. The first random candidate uses the single-start seed. The other seven have independent recorded seeds. No target test image is used for this selection, and reported iteration counts include every candidate run. Before target fitting, we replace source dictionary entries below 10^{-4} by 10^{-4}. For the source-initialized fit, we draw target coefficient entries independently from the uniform distribution on $[0, 1)$ and replace values below 10^{-4} by 10^{-4}. Source images do not enter the target objective. Test coefficients are fitted with the estimated dictionary fixed. Table 21 reports the full summary.

Table 21: Digits NMF production summary with target-stage iterations.

Method	Mean RMSE	Median RMSE	90% q.	Target iter.
source dictionary initialization	0.1503	0.1505	0.1550	85.48
target random initialization	0.1523	0.1523	0.1570	200.00
Target NNDSVDa	0.1542	0.1547	0.1593	200.00
Target 8 random starts	0.1535	0.1536	0.1588	1599.37
pooled source-target dictionary	0.1592	0.1591	0.1644	200.00

The source dictionary initialization has the lowest mean and median test RMSE among the five methods under the stated fitting budgets. The pooled source-target dictionary is less accurate for the target task, which is consistent with the risk comparison in Section F: source data can increase

the target reconstruction error when the source-target difference is not modeled. The NNDSVDa, single random, and pooled fits reach the 200-iteration cap on every split. The eight-start method uses 1,599.37 target iterations on average. These comparisons do not establish the ordering of fully converged solutions.

Y Additional Numerical Diagnostics

The tables in Section T report marginal summaries by method. This section adds paired comparisons computed from the same repetitions. Pairing compares methods on the same generated sample or target split and accounts for their covariance in the Monte Carlo uncertainty. A negative difference means that source initialization has smaller error than the comparator on the same repetition.

Table 22: Paired phase-retrieval comparisons. Differences are source initialization error minus comparator error.

Design shift	Comparator	Mean diff.	Median diff.	Source win rate
0.35	pooled data	-0.0444	-0.0671	0.75
0.35	target random restart	-0.1230	-0.0022	0.75
0.35	target spectral	-0.6731	-0.8611	0.88
0.55	pooled data	-0.1907	-0.2064	0.96
0.55	target random restart	-0.1776	-0.0018	0.72
0.55	target spectral	-0.6228	-0.7363	0.88
0.75	pooled data	-0.2860	-0.3222	0.95
0.75	target random restart	-0.0946	-0.0009	0.65
0.75	target spectral	-0.6366	-0.7275	0.85

Table 22 shows that source initialization has smaller mean error than all three comparators at each reported shift. It has smaller error than pooling in 75, 96, and 95 percent of repetitions at $\rho = 0.35, 0.55, 0.75$, respectively. These paired comparisons use the same observations for each method.

Table 23: Paired matrix-completion comparisons.

ϕ	Comparator	Mean diff.	Median diff.	Source win rate
0.25	pooled data	-0.0469	-0.0472	1.00
0.25	target spectral	-0.0317	-0.0126	0.94
0.25	target random restart	-0.1193	-0.1133	1.00
0.45	pooled data	-0.2795	-0.2796	1.00
0.45	target spectral	-0.0257	-0.0116	0.92
0.45	target random restart	-0.1162	-0.1125	1.00

Table 23 shows that the larger source-target shift changes the pooled comparison and not the target criterion optimized after source initialization. Source initialization beats both target-only starts at every shift in at least 92 percent of repetitions. It also beats pooling in every repetition at both reported shifts.

Table 24: Paired digits NMF comparisons across 100 target splits. Differences are source dictionary initialization RMSE minus comparator RMSE.

Comparator	Mean diff.	Median diff.	Source win rate
target random initialization	-0.00197	-0.00243	0.69
Target NNDSVDa	-0.00395	-0.00405	0.91
Target 8 random starts	-0.00317	-0.00336	0.84

Comparator	Mean diff.	Median diff.	Source win rate
pooled source-target dictionary	-0.00888	-0.00884	0.97

Table 24 gives a target-split comparison for the real-data analysis. The source dictionary initialization improves on the target random start on 69 percent of splits and improves on the pooled dictionary on 97 percent of splits. Its win fractions against target NNDSVDa and eight random starts are 0.91 and 0.84. The pooling comparison is the empirical analogue of the source-data bias term in Section F: including the source images in the target dictionary fit changes the reconstruction problem for the target digits, whereas source initialization changes only the starting dictionary used by the target fit.

The precision of the paired means in Tables 22–24 depends on variation across repetitions. For each comparison, we compute the Monte Carlo standard error as the sample standard deviation of the 100 paired differences divided by $\sqrt{100}$. The approximate pointwise 95 percent Monte Carlo intervals use a Student t quantile with 99 degrees of freedom. For digits, these intervals condition on the fixed data set and describe variation from the randomized splits, source subsampling, and initialization.

Table 25: Monte Carlo uncertainty in paired mean error differences.

Design	Comparator	MCSE	Lower	Upper
Phase retrieval, design shift ρ				
0.35	pooled data	0.008697	-0.061643	-0.027130
0.35	target random restart	0.028927	-0.180365	-0.065569
0.35	target spectral	0.054822	-0.781898	-0.564341
0.55	pooled data	0.008937	-0.208474	-0.173008
0.55	target random restart	0.037973	-0.252956	-0.102263
0.55	target spectral	0.055470	-0.732825	-0.512696
0.75	pooled data	0.015172	-0.316140	-0.255933
0.75	target random restart	0.031522	-0.157173	-0.032081
0.75	target spectral	0.057025	-0.749742	-0.523441
Matrix completion, angle ϕ				
0.25	pooled data	0.001275	-0.049441	-0.044381
0.25	target spectral	0.004860	-0.041334	-0.022048
0.25	target random restart	0.005445	-0.130131	-0.108524
0.45	pooled data	0.001186	-0.281888	-0.277183
0.45	target spectral	0.004627	-0.034851	-0.016490
0.45	target random restart	0.005774	-0.127686	-0.104773
Digits				
–	target random	0.000362	-0.002689	-0.001251
–	pooled dictionary	0.000462	-0.009802	-0.007967

Table 25 shows that the intervals lie below zero for every target-only comparison and for pooling at all three phase shifts and both matrix shifts. Both digits intervals also lie below zero. The intervals quantify the precision of the mean improvements under the specified fitting budgets and error measures.

Table 26: Paired digits RMSE differences for the additional target-only baselines.

Comparator	Mean difference	95% interval	Source win rate
Target NNDSVDa	-0.00395	[-0.00465, -0.00325]	0.91
Target 8 random starts	-0.00317	[-0.00379, -0.00254]	0.84

Table 26 gives the paired comparisons with target NNDSVDa and eight random starts. Both intervals lie below zero. They describe variation across the same fixed-data splits and use the same pointwise Student t calculation as Table 25.

Z Discussion

The results support a precise claim. A source-trained estimator used only as a starting point is not a weak version of source-data pooling. It is a different statistical object. Its value is the probability of entering a target local region where target-only statistical theory is valid. Under the conditions of Corollary 4, entry with probability tending to one and sufficiently many target updates give the same first-order distribution as the target local estimator.

The effect of source size depends on the population shift. If the source parameter lies strictly inside the certified target ball, source consistency makes its entry probability tend to one. If it lies strictly outside, the entry probability tends to zero by Theorem 6. Under the same iteration condition, further source observations do not change this first-order distribution once entry has probability tending to one. Source data used in the estimating equation can continue to reduce variance, but only under a bias control that depends on the source-target gap. Proposition 9 quantifies the gain that would be available if one procedure achieved both target-basin entry and the local source-data risk. A concrete combined estimator needs its own basin analysis. When the regime is unknown, the validation-based selection rule of Proposition 10 attains the better mode up to the validation tolerance.

The numerical results distinguish computation, inference, and risk selection. The phase-inference experiments compare target coverage with coverage of the pooled population parameter and examine finite and infinite source-noise variances. The validation experiment shows that independent target risk can select between initialization and pooling with small empirical regret. Matrix completion and the digit analysis show fixed-budget improvements in larger factorization problems. The last two results do not establish first-order theory for those factorization estimators.

Several extensions are natural but are not claimed as theorems here. With multiple sources, the relevant design problem is sequential selection of initializers based on certified basin-entry probability and validation performance. For regularization paths, the source start can be used at a penalty level for which local target convergence is verified, followed by a target-only path. For modern fine-tuning, the present results suggest reporting basin-entry proxies, validation separation, and restart behavior; realistic architectures need model-specific basin analysis. The global convergence of certain infinite-width gradient flows in Chizat and Bach (2018) uses different initialization and limiting assumptions.

Supplementary References

Ash, J. T. and Adams, R. P. (2020). On warm-starting neural network training. *Advances in Neural Information Processing Systems* **33**, 3884–3894.

Balcan, M.-F., Khodak, M., Sharma, D. and Talwalkar, A. (2021). Learning-to-learn non-convex piecewise-Lipschitz functions. *Advances in Neural Information Processing Systems* **34**, 15056–15069.

Bastani, H. (2021). Predicting with proxies: transfer learning in high dimension. *Management Science* **67**, 2964–2984.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research* **12**, 149–198.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning* **79**, 151–175.

Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: minimax rate and adaptive classifier. *Annals of Statistics* **49**, 100–128.

Chen, Q., Zhou, J. and Al Kontar, R. (2026). Language-induced priors for domain adaptation. *arXiv:2605.14301*.

Chijiwa, D. (2024). Transferring learning trajectories of neural networks. *International Conference on Learning Representations*.

Fallah, A., Mokhtari, A. and Ozdaglar, A. (2020). On the convergence theory of gradient-based model-agnostic meta-learning algorithms. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 1082–1092.

Fan, J., Gao, C. and Klusowski, J. M. (2025). Robust transfer learning with unreliable source data. *Annals of Statistics* **53**, 1728–1752.

Finn, C., Abbeel, P. and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

Ghane, R., Akhtiamov, D. and Hassibi, B. (2024). Universality in transfer learning for linear models. *Advances in Neural Information Processing Systems* **37**, 125729–125779.

Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. *Advances in Neural Information Processing Systems* **32**, 9871–9881.

Hedges, C. E. (2026). Source-optimal training is transfer-suboptimal. *Transactions on Machine Learning Research*.

Ho, N., Khamaru, K., Dwivedi, R., Wainwright, M. J., Jordan, M. I. and Yu, B. (2025). Instability, computational efficiency and statistical accuracy. *Journal of Machine Learning Research* **26**(65), 1–68.

Kumar, A., Ma, T. and Liang, P. (2020). Understanding self-training for gradual domain adaptation. *Proceedings of the 37th International Conference on Machine Learning*, 5468–5479.

Kumar, A., Raghunathan, A., Jones, R., Ma, T. and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. *International Conference on Learning Representations*.

Kuzborskij, I. and Lampert, C. H. (2018). Data-dependent stability of stochastic gradient descent. *Proceedings of the 35th International Conference on Machine Learning*, 2815–2824.

Lauditi, C., Bordelon, B. and Pehlevan, C. (2026). Transfer learning in infinite width feature learning networks. *International Conference on Learning Representations*.

Liu, K., Wang, Z. and Wu, L. (2024). The local landscape of phase retrieval under limited samples. *IEEE Transactions on Information Theory* **70**, 9012–9035.

McNamara, D. and Balcan, M.-F. (2017). Risk bounds for transferring representations with and without fine-tuning. *Proceedings of the 34th International Conference on Machine Learning*, 2373–2381.

Mousavi Kalan, M., Fabian, Z., Avestimehr, S. and Soltanolkotabi, M. (2020). Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems* **33**, 1959–1969.

Neyshabur, B., Sedghi, H. and Zhang, C. (2020). What is being transferred in transfer learning? *Advances in Neural Information Processing Systems* **33**, 512–523.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345–1359.

Reeve, H. W. J., Cannings, T. I. and Samworth, R. J. (2021). Adaptive transfer learning. *Annals of Statistics* **49**, 3618–3649.

Sadrtdinov, I., Pozdeev, D., Vetrov, D. and Lobacheva, E. (2023). To stay or not to stay in the pre-train basin: insights on ensembling in transfer learning. *Advances in Neural Information Processing Systems* **36**, 15936–15964.

Shachaf, G., Brutzkus, A. and Globerson, A. (2021). A theoretical analysis of fine-tuning with linear teachers. *Advances in Neural Information Processing Systems* **34**, 15382–15394.

Shcherbinin, Y., Redina, A., Kalpin, M. and Kochetov, V. (2026). Equilibrium selection in multi-agent policy gradients via opponent-aware basin entry. *arXiv:2605.18078*.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90**, 227–244.

Shin, B., Oh, J., Cho, H. and Yun, C. (2024). DASH: warm-starting neural network training in stationary settings without loss of plasticity. *Advances in Neural Information Processing Systems* **37**, 43300–43340.

Sugiyama, M., Krauledat, M. and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* **8**, 985–1005.

Tahir, J., Ganguli, S. and Rotskoff, G. M. (2025). Features are fate: a theory of transfer learning in high-dimensional regression. *Proceedings of the 42nd International Conference on Machine Learning*, 58142–58168.

Tam, N. V. K., My, C. H. and Le Duy, V. N. (2025). Post-transfer learning statistical inference in high-dimensional regression. *Statistics and Computing* **35**, 205.

Tian, Y., Weng, H., Xia, L. and Feng, Y. (2026). Robust unsupervised multi-task and transfer learning on Gaussian mixture models. *Journal of the American Statistical Association*. `https://doi.org/10.1080/01621459.2026.2670031`.

Tinati, M. and Tu, S. (2026). On the asymptotics of self-supervised pre-training: two-stage M-estimation and representation symmetry. *Proceedings of the 39th Conference on Learning Theory*, 6197–6309.

Wu, J., Zou, D., Braverman, V., Gu, Q. and Kakade, S. (2022). The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems* **35**, 33041–33053.

Xu, Y., Nakada, R., Zhang, L. and Li, L. (2026). Residual feature integration is sufficient to prevent negative transfer. *International Conference on Learning Representations*.

Yao, Y. (2022). Constructing confidence intervals for the signals in sparse phase retrieval. *Electronic Journal of Statistics* **16**, 785–813.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems* **31**, 3036–3046.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7**, 1–46.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* **28**, 1302–1338.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16**, 559–616.

Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Annals of Statistics* **46**, 2747–2774.

Sun, R. and Luo, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory* **62**, 6535–6579.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* **12**, 389–434.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge.

Xu, J., Hsu, D. J. and Maleki, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. *Advances in Neural Information Processing Systems* **29**, 2676–2684.

Yang, S., Gao, C., Zeng, D. and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society, Series B* **85**, 575–596.