LIKELIHOOD LANDSCAPE AND MAXIMUM LIKELIHOOD ESTIMATION FOR THE DISCRETE ORBIT RECOVERY MODEL

ZHOU FAN, YI SUN, TIANHAO WANG, AND YIHONG WU

Abstract. We study the non-convex optimization landscape for maximum likelihood estimation in the discrete orbit recovery model with Gaussian noise. This is a statistical model motivated by applications in molecular microscopy and image processing, where each measurement of an unknown object is subject to an independent random rotation from a known rotational group. Equivalently, it is a Gaussian mixture model where the mixture centers belong to a group orbit.

We show that fundamental properties of the likelihood landscape depend on the signal-to-noise ratio and the group structure. At low noise, this landscape is "benign" for any discrete group, possessing no spurious local optima and only strict saddle points. At high noise, this landscape may develop spurious local optima, depending on the specific group. We discuss several positive and negative examples, and provide a general condition that ensures a globally benign landscape at high noise. For cyclic permutations of coordinates on $\mathbb{R}^d$ (multi-reference alignment), there may be spurious local optima when $d \geq 6$, and we establish a correspondence between these local optima and those of a surrogate function of the phase variables in the Fourier domain.

We show that the Fisher information matrix transitions from resembling that of a single Gaussian distribution in low noise to having a graded eigenvalue structure in high noise, which is determined by the graded algebra of invariant polynomials under the group action. In a local neighborhood of the true object, where the neighborhood size is independent of the signal-to-noise ratio, the landscape is strongly convex in a reparametrized system of variables given by a transcendence basis of this polynomial algebra. We discuss implications for optimization algorithms, including slow convergence of expectation-maximization, and possible advantages of momentum-based acceleration and variable reparametrization for first- and second-order descent methods.

Contents

Date: March 30, 2020.

1. Introduction

We study statistical estimation of a vector $\theta_* \in \mathbb{R}^d$ from noisy observations, where each observation is subject to a random and unknown rotation. Letting $G \subseteq \mathrm{O}(d)$ be a known subgroup of orthogonal rotations in dimension d, we consider the observation model

$$Y = g \cdot \theta_* + \sigma\varepsilon. \tag{1.1}$$

Here, $g \sim \mathrm{Unif}(G)$ is an unobserved uniform random element of this group, $\sigma > 0$ is the noise level, and $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$ is observation noise that is independent of g. This model is sometimes referred to as multi-reference alignment, the group action channel, or the orbit recovery problem [BRW17, BBSK$^+$17, BBM$^+$17, ABL$^+$18, APS18, BBLS18, PWB$^+$19, Bru19].

Study of this model has largely been motivated by its relevance to the structure recovery problem arising in single-particle cryo-electron microscopy (cryo-EM) [DAC$^+$88, HBC$^+$90, Fra06]. Cryo-EM is an experimental method of determining the 3D structure of a molecule by imaging many cryogenic samples of the molecule from different and unknown viewing angles. Due to limitations of electron dose, the individual images are subject to high levels of measurement noise, and they must be aligned and averaged to obtain a high-resolution reconstruction of the molecule. There is extensive literature on computational methods for this problem, and we refer readers to the recent survey [BBS19]. In our work, we study the simpler model (1.1), which omits many complications in cryo-EM such as a tomographic projection, the contrast-transfer function, and structural heterogeneity. We do this so as to focus our attention on some of the fundamental features of this reconstruction problem that may arise due to the latent rotation g.

It has been observed since [Sig98] that the difficulty of estimation in the model (1.1) has an atypically strong dependence on the noise level σ, and this is a common theme in subsequent study [BRW17, BBSK$^+$17, APS18, PWB$^+$19]. Figure 1.1 contrasts a low-noise and high-noise setting in a simple example, where G is the group of three-fold rotations on the plane $\mathbb{R}^2$. Three distinct clusters corresponding to the orbit points $\{g\theta_* : g \in G\}$ are observed in low noise, whereas only a single large cluster is apparent in high noise. The number of samples needed to recover θ_* and the dependence of this sample complexity on σ were studied in [BBSK$^+$17, APS18]. In particular, [BBSK$^+$17] showed that method-of-moments estimators can achieve rate-optimal sample complexity in σ, and connected this complexity to properties of the algebra of G-invariant polynomials.

The focus of our current work is, instead, on maximum likelihood estimation for θ_*. Maximum likelihood is a widely used approach in practice, for either *ab initio* estimation of θ_* or for iterative refinement of a pilot estimate obtained by other means [Sig98, SVN$^+$05, SNRGL$^+$07, SGV$^+$07]. Letting $Y_1, \dots, Y_n$ be i.i.d. observations from the model (1.1), the maximum likelihood estimate

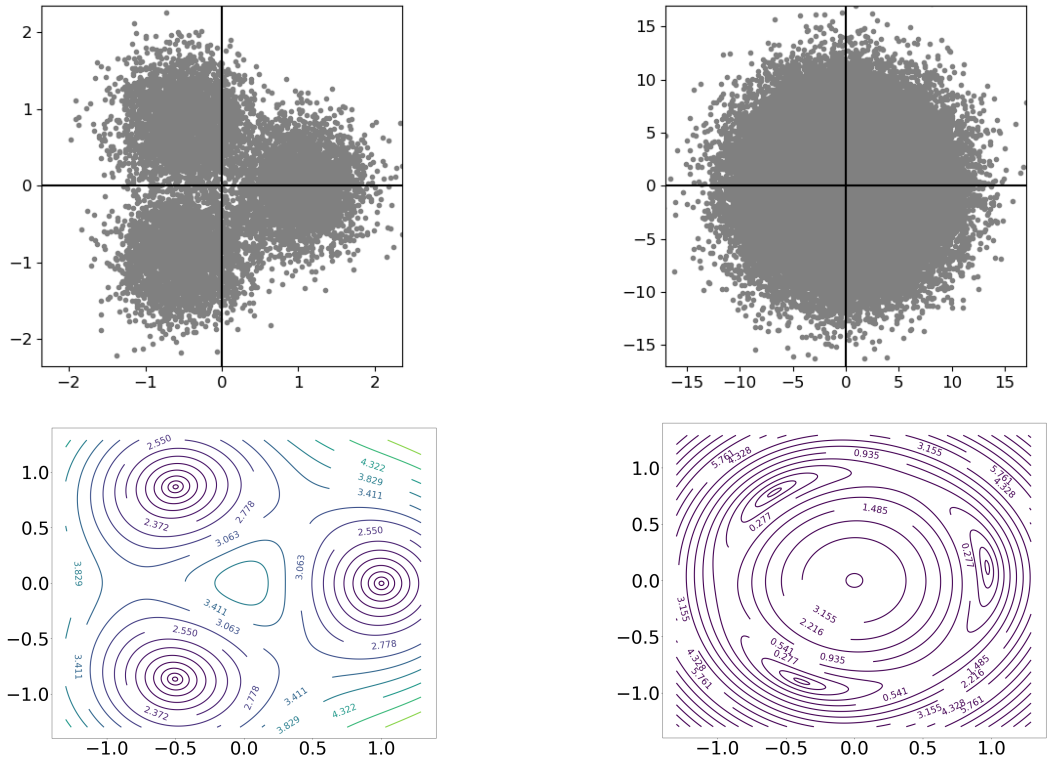


FIGURE 1.1. Data samples and contours of negative log-likelihood $R_n(\theta)$ for the group of three-fold rotations on $\mathbb{R}^2$, where $\theta_* = (1, 0)$. Left: 10,000 samples at noise level $\sigma = 0.4$. Right: 100,000 samples at noise level $\sigma = 4$. (Note the difference in axis limits between the data plots and contour plots.) Values in the contour plots are displayed under an affine transformation for better visualization.

(MLE) is a vector $\hat{\theta} \in \mathbb{R}^d$ which maximizes the log-likelihood function

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(Y_i),$$

where p_θ is the probability density of Y marginalizing over the latent rotation $g \sim \mathrm{Unif}(G)$. We denote the negative log-likelihood function by $R_n(\theta)$; this function is also depicted in Figure 1.1 for low and high noise. The success of optimization algorithms for computing the MLE for ab initio estimation and for iterative refinement depends, respectively, on the global function landscape of $R_n(\theta)$ and on its local landscape in a neighborhood of θ_*.

In this work, we study the function landscape of $R_n(\theta)$, assuming that the true vector $\theta_* \in \mathbb{R}^d$ is suitably generic. We restrict attention to discrete groups G, so that $R_n(\theta)$ has isolated critical points, and we derive several results. First, we show that the global landscape is "benign" for sufficiently low noise, having no spurious local minimizers for any discrete group. Second, we show that the local landscape in a σ-independent neighborhood of θ_* is also benign at any noise level $\sigma > 0$, and that $R_n(\theta)$ is strongly convex in this neighborhood after suitable reparametrization. Third, we relate the critical points of the global landscape in high noise to a sequence of simpler optimization problems defined by the symmetric moment tensors under G. We show that for discrete rotations in $\mathbb{R}^2$ as in Figure 1.1, and for the symmetric group that permutes the coordinates

of $\mathbb{R}^d$, the global landscape is benign also at high noise. In contrast, for the group of cyclic permutations in $\mathbb{R}^d$, the global landscape may not be benign once the dimension reaches $d = 6$.

Our motivations for studying the MLE and the likelihood landscape are two-fold. First, classical statistical theory indicates that in the limit $n \to \infty$ for fixed dimension d, the MLE achieves asymptotic efficiency, meaning that $\hat{\theta}$ converges to θ_* at an $O(1/\sqrt{n})$ rate, with asymptotically optimal covariance $I(\theta_*)^{-1}$ (the inverse of the Fisher information matrix) matching the Cramer-Rao lower bound (see [LC06, Sec. 2.5]). This need not hold for method-of-moments estimators as studied in [BBSK+17]. Our results connect one aspect of [BBSK+17] regarding the sample complexity for "list-recovery of generic signals" to the MLE, by showing that the eigenstructure of the Fisher information matrix $I(\theta_*)$ corresponds to a sequence of transcendence degrees in the graded algebra of G-invariant polynomials.

Second, a body of empirical literature in cryo-EM suggests that $R_n(\theta)$ may have spurious local minimizers. For ab initio estimation, this has motivated the development of a variety of optimization algorithms including stochastic hill climbing [EEB13], stochastic gradient descent [PRFB17], and "frequency marching" [BGPS17]. However, at present, the function landscape of $R_n(\theta)$ is not theoretically well-understood, even in simple examples of group actions. For instance, it is unclear how this landscape depends on properties of the group, and whether the roughness of this landscape is due to insufficient sample size or is a fundamental aspect of the model even in the $n \to \infty$ limit. Our work takes a step towards understanding these questions, and our results have concrete implications for descent-based optimization algorithms in this problem. We discuss these implications in Section 1.3 below.

1.1. **The orbit recovery model.** We study the orbit recovery model (1.1) in the setting of a discrete group. Let $G \subset \mathrm{O}(d) \subset \mathbb{R}^{d\times d}$ be a discrete subgroup of the orthogonal group in dimension d, with finite cardinality

$$|G| = K.$$

Each observation is modeled as

$$Y = g \cdot \theta_* + \sigma\varepsilon$$

where $g \sim \mathrm{Unif}(G)$, $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$, and these are independent. Here, $\sigma > 0$ is the noise level, which we will assume is known. This is a K-component Gaussian mixture model with equal weights, where the centers of the mixture components are the points of the *orbit* of θ_* under G, given by

$$\mathcal{O}_{\theta_*} = \{g\theta_* : g \in G\}.$$

The marginal density of Y in this model is the Gaussian mixture density

$$p_{\theta_*}(Y) = \frac{1}{K}\sum_{g\in G}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^d \exp\left(-\frac{\|Y - g\theta_*\|^2}{2\sigma^2}\right). \tag{1.2}$$

For $\theta, \theta' \in \mathbb{R}^d$, note that $p_\theta = p_{\theta'}$ if and only if the K mixture components have the same centers, i.e. $\mathcal{O}_{\theta'} = \mathcal{O}_\theta$. This means the parameter θ_* is statistically identifiable in this model up to its orbit.

Given n independent samples $Y_1, \dots, Y_n$ distributed according to (1.1), we study the landscape of the negative log-likelihood *empirical risk*

$$R_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(Y_i) + \text{const.} \tag{1.3}$$

Here, const denotes a θ-independent value that we introduce to simplify the expression for this risk; see (2.1) for details. This function $R_n(\theta)$ is non-convex for any non-trivial group G.

A maximum likelihood estimator $\hat{\theta} \in \mathbb{R}^d$ is any global minimizer of $R_n(\theta)$. Note that if $\hat{\theta}$ minimizes $R_n(\theta)$, then all points in its orbit $\mathcal{O}_{\hat{\theta}}$ also minimize $R_n(\theta)$, so the MLE is also only defined up to its orbit.

Fixing the true parameter $\theta_* \in \mathbb{R}^d$, we denote the mean of $R_n(\theta)$ by

$$R(\theta) = -\mathbb{E}\big[\log p_\theta(Y)\big] + \text{const}, \tag{1.4}$$

where $\mathbb{E}$ is the expectation over both g and ε in the model $Y = g \cdot \theta_* + \sigma\varepsilon$. This function $R(\theta)$ depends implicitly on the true parameter θ_*. We call $R(\theta)$ the *population risk*, and this may be understood as the $n \to \infty$ limit of $R_n(\theta)$. Note that

$$R(\theta) = D_{\mathrm{KL}}(p_{\theta_*} \| p_\theta) - \mathbb{E}[\log p_{\theta_*}(Y)] + \text{const} \tag{1.5}$$

where $D_{\mathrm{KL}}(p\|q) = \int p(y) \log \frac{p(y)}{q(y)} dy$ is the Kullback-Liebler divergence between densities p and q, and the remaining two terms do not depend on θ. Thus, a point $\theta \in \mathbb{R}^d$ is a global minimizer of $R(\theta)$ if and only if $p_{\theta_*} = p_\theta$, i.e. $\theta \in \mathcal{O}_{\theta_*}$.

It was established in [MBM18] that under mild conditions for empirical risks such as (1.3), due to concentration of the gradient and Hessian of $R_n(\theta)$ around those of $R(\theta)$, various properties of the function landscape of $R(\theta)$ translate to those of $R_n(\theta)$ for sufficiently large n—these properties include the number of critical points and the number of negative Hessian eigenvalues at each critical point. Versions of this argument were also used in the analyses of dictionary learning and phase retrieval in [SQW16, SQW18]. Our analysis will follow a similar approach, and the core of our arguments will pertain to the population risk (1.4) rather than its finite-n counterpart (1.3).

We will also study properties of the Fisher information matrix in this model. This is given by

$$I(\theta_*) = -\mathbb{E}\big[\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta_*}\big] = \nabla_\theta^2 R(\theta_*), \tag{1.6}$$

which is the Hessian of the population risk $R(\theta)$ evaluated at its global minimizer $\theta = \theta_*$. It was shown in [Bru19] that $I(\theta_*)$ is invertible if and only if all K points of the orbit $\mathcal{O}_{\theta_*}$ are distinct. We assume this condition in all of our results, and some of our results will further restrict θ_* to satisfy additional *generic* properties that hold outside the zero set of an analytic function on $\mathbb{R}^d$. Identifying the MLE $\hat{\theta}$ as the point in its orbit closest to θ_*, [APS18] verified that $\hat{\theta}$ is an asymptotically consistent estimate for θ_* as $n \to \infty$. By the classical theory of maximum likelihood estimation in parametric models (see [VdV00, Chapter 5]), we then have the convergence in law

$$\sqrt{n}(\hat{\theta} - \theta_*) \to \mathcal{N}\big(0, I(\theta_*)^{-1}\big). \tag{1.7}$$

Thus the eigenvalues of the Fisher information matrix determine the coordinate-wise asymptotic variances of the MLE in an orthogonal basis for $\mathbb{R}^d$.

1.2. **Overview of results.** We will be interested in the geometric properties of the function landscapes of $R_n(\theta)$ and $R(\theta)$. The most ideal setting for non-convex optimization is when these landscapes are benign in the following sense.

Definition 1.1. The landscape of a twice continuously-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is **globally benign** if the only local minimizers of f are global minimizers, f is strongly convex at each such local minimizer, and each saddle point of f is a strict saddle point.

This is equivalent to saying that the only points $\theta \in \mathbb{R}^d$ where $\nabla f(\theta) = 0$ and $\lambda_{\min}(\nabla^2 f(\theta)) \geq 0$ are the global minimizers of f, and $\lambda_{\min}(\nabla^2 f(\theta)) > 0$ strictly at all such points. This condition has been discussed in [GHJY15, LSJR16, JGN+17], which show that randomly-initialized gradient descent converges to a global minimizer almost surely under this condition, and that gradient descent perturbed with additive noise can furthermore converge in polynomial time under a quantitative version of this condition.

In our results, we will fix a generic true parameter $\theta_* \in \mathbb{R}^d$. We study low-noise and high-noise regimes, where the low-noise regime is defined by $\sigma < \sigma_0$ for a sufficiently small (θ_*, d, G)-dependent constant $\sigma_0 > 0$, and the high-noise regime by $\sigma > \sigma_0$ for a (different) sufficiently large (θ_*, d, G)-dependent constant $\sigma_0 > 0$. It is the high-noise regime that is of primary interest in applications such as cryo-EM. We provide results also for low noise, to contrast with the high-noise behavior,

and because these results may be of separate interest in other applications.

Global landscape and Fisher information at low noise. We show in Section 3 that both $R(\theta)$ and $R_n(\theta)$ are globally benign in the low noise regime, for any discrete group G and any θ_* whose orbit points are distinct under G. That is, there exists $\sigma_0 \equiv \sigma_0(\theta_*, d, G)$ for which $R(\theta)$ and $R_n(\theta)$ do not have any spurious local minimizers when $\sigma < \sigma_0$.

We also show that the Fisher information satisfies $I(\theta_*) \approx \sigma^{-2}\,\mathrm{Id}$, where the error of this approximation is exponentially small in σ^{-2}. Here, $\sigma^{-2}\,\mathrm{Id}$ is the Fisher information of the single Gaussian distribution $\mathcal{N}(\theta_*, \sigma^2\,\mathrm{Id})$. Thus the local geometries of $R(\theta)$ and $R_n(\theta)$ near θ_* resemble those of a single Gaussian, and they do not "feel" the effects of the other mixture components.

We remark that the group structure plays an important role in our proof of this global landscape result, and such a result is not true for general Gaussian mixture models: For the three-component Gaussian mixture model

$$\frac{1}{3}\mathcal{N}(\theta_1, \sigma^2\,\mathrm{Id}) + \frac{1}{3}\mathcal{N}(\theta_2, \sigma^2\,\mathrm{Id}) + \frac{1}{3}\mathcal{N}(\theta_3, \sigma^2\,\mathrm{Id}),$$

it is known that the negative log-likelihood population risk as a function of $(\theta_1, \theta_2, \theta_3) \in \mathbb{R}^{3d}$ can have spurious local minimizers, even in the $\sigma \to 0$ limit. Similar examples may be constructed for any number of mixture components $K \geq 3$ [JZB+16].

Fisher information at high noise. As the noise level σ increases, a transition occurs in the structure of the Fisher information matrix $I(\theta_*)$. We show in Section 4.4 that in the high-noise regime, for any generic $\theta_* \in \mathbb{R}^d$, there is a decomposition $d = d_1 + d_2 + \ldots + d_L$ where

$$I(\theta_*) \text{ has } d_\ell \text{ eigenvalues on the order of } \sigma^{-2\ell} \text{ for each } \ell = 1, \ldots, L. \tag{1.8}$$

The number d_ℓ is $\mathrm{trdeg}(\mathcal{R}^G_{\leq \ell}) - \mathrm{trdeg}(\mathcal{R}^G_{\leq \ell - 1})$, where $\mathrm{trdeg}(\mathcal{R}^G_{\leq \ell})$ is the transcendence degree over $\mathbb{R}$ of the space of G-invariant polynomials having degree $\leq \ell$. The number L is the smallest integer for which $\mathrm{trdeg}(\mathcal{R}^G_{\leq L}) = d$.

For the group of K-fold discrete rotations in $\mathbb{R}^2$, as in Figure 1.1, we have $L = K$, $d_2 = 1$, $d_K = 1$, and $d_\ell = 0$ for each other ℓ. Thus $I(\theta_*)$ has one eigenvalue of magnitude σ^{-4}, corresponding to the curvature of $R(\theta)$ in the radial direction, and one eigenvalue of magnitude σ^{-2K}, corresponding to the direction tangent to the circle $\{\theta \in \mathbb{R}^2 : \|\theta\| = \|\theta_*\|\}$. For the symmetric group of all permutations in $\mathbb{R}^d$, we have $L = d$ and $d_\ell = 1$ for each $\ell = 1, \ldots, d$. For cyclic permutations in $\mathbb{R}^d$, we have $L = 3$, $d_1 = 1$, $d_2 = \lceil \frac{d-1}{2} \rceil$, and $d_3 = \lfloor \frac{d-1}{2} \rfloor$. Here d_1 corresponds to the sum $\theta_1 + \ldots + \theta_d$, d_2 to the magnitudes of the remaining Fourier coefficients of θ, and d_3 to the phases.

Applying (1.8) to the classical efficiency result (1.7) for the MLE, this shows that $\hat{\theta}$ estimates θ_* with an asymptotic covariance of $O(\sigma^{2L}/n)$. This rate agrees with the results of [BBSK+17] on list-recovery of generic signals θ_* by a method-of-moments estimator. More precisely, (1.8) exhibits a decomposition of $\mathbb{R}^d$ into orthogonal subspaces of dimensions $d_1, \ldots, d_L$, such that the MLE $\hat{\theta}$ estimates θ_* with an asymptotic covariance of $O(\sigma^{2\ell}/n)$ in its component belonging to the ℓ^{th} subspace. For any continuously differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, a Taylor expansion of ψ (i.e. the statistical delta method) yields also the convergence in law

$$\sqrt{n}\Big(\psi(\hat{\theta}) - \psi(\theta_*)\Big) \to \mathcal{N}\Big(0, \nabla\psi(\theta_*)^\top I(\theta_*)^{-1} \nabla\psi(\theta_*)\Big) \tag{1.9}$$

as $n \to \infty$. We show that when ψ is any G-invariant polynomial of degree ℓ, the gradient $\nabla\psi(\theta_*)$ belongs to the span of the first ℓ subspaces, so that $\psi(\hat{\theta})$ estimates $\psi(\theta_*)$ with variance $O(\sigma^{2\ell}/n)$.

Global landscape at high noise. Denote by

$$T_\ell(\theta) = \mathbb{E}_g[(g\theta)^{\otimes \ell}] \in (\mathbb{R}^d)^{\otimes \ell} \tag{1.10}$$

the ℓ^{th} moment tensor of $g\theta$, where $\mathbb{E}_g$ is the expectation over the uniform law $g \sim \text{Unif}(G)$. The entries of $T_\ell(\theta)$ consist of all order-ℓ mixed moments of entries of the random vector $g\theta \in \mathbb{R}^d$. Let $\|\cdot\|_{\text{HS}}$ be the Euclidean norm of the vectorization of such a tensor in $\mathbb{R}^{d^\ell}$. We relate the local minimizers of $R(\theta)$ and $R_n(\theta)$ in the high-noise regime to a sequence of simpler optimization problems, given by successively minimizing

$$P_\ell(\theta) = \|T_\ell(\theta) - T_\ell(\theta_*)\|_{\text{HS}}^2 \tag{1.11}$$

over the variety

$$\mathcal{V}_{\ell-1} = \Big\{\theta \in \mathbb{R}^d : T_k(\theta) = T_k(\theta_*) \text{ for } k = 1, \dots, \ell-1\Big\}, \tag{1.12}$$

for $\ell = 1, \dots, L$. This sequence of optimization problems is related to the method-of-moments, in that (1.11) may be interpreted as matching the order-ℓ moments $T_\ell(\theta)$ to $T_\ell(\theta_*)$, subject to the constraint (1.12) that the moments of lower order have already been matched.

We show in Section 4.5 that for generic θ_*, if $\mathcal{V}_L = \mathcal{O}_{\theta_*}$, each variety $\mathcal{V}_\ell$ is non-singular with constant dimension, each restriction $P_\ell|_{\mathcal{V}_{\ell-1}}$ satisfies a strict saddle condition, and the only local minimizers of each restriction $P_\ell|_{\mathcal{V}_{\ell-1}}$ are the points $\theta \in \mathcal{V}_\ell$, then the global landscape of $R(\theta)$ is also benign in the high-noise regime. We analyze the two concrete examples of K-fold rotations in $\mathbb{R}^2$ and the symmetric group of all permutations in $\mathbb{R}^d$, showing that the global landscape is benign at high noise in these examples.

The first condition $\mathcal{V}_L = \mathcal{O}_{\theta_*}$ means that θ_* is uniquely specified, up to its orbit, by its first L moment tensors $T_1(\theta_*), \dots, T_L(\theta_*)$. These are the examples in [BBSK+17] where the notions of "generic list recovery" and "generic unique recovery" coincide. We note that this condition alone is not sufficient to guarantee a benign landscape. For instance, in the cyclic permutations example below, we have $L = 3$ and $\mathcal{V}_3 = \mathcal{O}_{\theta_*}$ for generic points $\theta_* \in \mathbb{R}^d$ in any dimension d, but spurious local minima can exist in dimensions $d \geq 6$.

Spurious local minimizers for cyclic permutations. The complexity of the sequence of optimization problems in (1.11–1.12) depends on the structure of the G-invariant polynomial algebra. As a more complex example, we study in Section 4.6 the group G of cyclic permutations in $\mathbb{R}^d$. Some authors refer to this specific action as the multi-reference alignment model, and the invariant polynomial algebra for this group bears some similarities to the continuous action of SO(3) that is relevant for cryo-EM applications [BRW17, BBSK+17, PWB+19].

For this group, we have $L = 3$, and $P_\ell(\theta)$ does not have spurious local minimizers over $\mathcal{V}_{\ell-1}$ for $\ell = 1$ and 2. For $\ell = 3$ and odd d, denoting $\mathcal{I} = \{1, 2, \dots, \frac{d-1}{2}\}$, minimizing $P_3(\theta)$ over $\mathcal{V}_2$ is equivalent to minimizing

$$F^+(t_1, \dots, t_{|\mathcal{I}|}) = -\frac{1}{6} \sum_{\substack{i,j,k \in \mathcal{I} \cup -\mathcal{I} \\ i+j+k \equiv 0 \bmod d}} r_{i,*}^2 r_{j,*}^2 r_{k,*}^2 \cos(t_i + t_j + t_k)$$

over phase variables $t_1, \dots, t_{|\mathcal{I}|} \in [0, 2\pi)$, where we identify $t_{-i} = -t_i$ and set $r_{i,*}$ as the modulus of the i^{th} Fourier coefficient of θ_*. When d is even, there is an additional term to this function as well as a second function $F^-(t_1, \dots, t_{|\mathcal{I}|})$, and we refer to Section 4.6 for details.

We show that for high noise and generic $\theta_* \in \mathbb{R}^d$, local minimizers of $R(\theta)$ are in correspondence with local minimizers of $F^\pm(t_1, \dots, t_{|\mathcal{I}|})$, where the magnitudes of the Fourier coefficients of any such local minimizer $\theta \in \mathbb{R}^d$ are close to those of θ_*, and the differences in phases between the Fourier coefficients of θ and those of θ_* are close to the corresponding local minimizer of $F^\pm$. In dimensions $d \leq 5$, there are no spurious local minimizers, and the landscapes of $R(\theta)$ and $R_n(\theta)$ are globally benign. In dimension $d = 6$, we exhibit an open set $U \subset \mathbb{R}^d$ such that $R(\theta)$ and $R_n(\theta)$ *do* have spurious local minimizers, for all $\theta_* \in U$. This is a phenomenon of the population risk $R(\theta)$ and is not caused by finite-n behavior, so descent procedures may converge to these spurious

local minimizers even in the limit of infinite sample size.

Local landscape at high noise. Motivated by the possibility that $R(\theta)$ and $R_n(\theta)$ are not globally benign, we study also their local landscapes restricted to a smaller neighborhood of θ_* in Section 4.4. We show that there is a *σ-independent* neighborhood U of θ_*, and a local reparametrization by an analytic map $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ that is 1-to-1 on U, such that R and R_n are strongly convex as functions of $\varphi \in \varphi(U)$, with unique local minimizers in U. The coordinates of this map φ may be taken to be d polynomials that form a transcendence basis of the G-invariant polynomial algebra.

We remark that this result does not automatically follow from the invertibility of the Fisher information $I(\theta_*)$ established in [Bru19], as this invertibility does not preclude the possibility that the size of this neighborhood U shrinks as $\sigma \to \infty$. In fact, it is not true that $R(\theta)$ must be convex over $\theta \in U$ for a σ-independent neighborhood U, and the reparametrization by φ is important to ensure convexity. For instance, in the high-noise picture of Figure 1.1, it is evident from the non-convex level sets that $R_n(\theta)$ is convex only in a small neighborhood of θ_*. However, it is convex in a much larger neighborhood of θ_* when reparametrized by two coordinates that represent the radius and angle.

High-noise expansion of the population risk. Our results in the high-noise regime are enabled by a series expansion of the population risk function in σ^{-2}, given by

$$R(\theta) = \sum_{\ell=1}^{\infty} \sigma^{-2\ell} S_\ell(\theta)$$

for certain G-invariant polynomial functions $S_\ell(\theta)$. For fixed $\theta_* \in \mathbb{R}^d$, each polynomial $S_\ell(\theta)$ takes the form

$$S_\ell(\theta) = \frac{1}{2(\ell!)} \|T_\ell(\theta) - T_\ell(\theta_*)\|_{\mathrm{HS}}^2 + Q_\ell(\theta)$$

where $Q_\ell(\theta)$ is in the algebra generated by G-invariant polynomials of degree $\le \ell - 1$. We derive these results and provide a rigorous interpretation of this expansion in Section 4.2.

By the relation (1.5), this is equivalent to a series expansion of the KL-divergence $D_{\mathrm{KL}}(p_{\theta_*} \| p_\theta)$ in σ^{-2}. In the works [BRW17, BBSK+17, APS18], analogous expansions were performed instead for upper and lower bounds to the KL-divergence, and these were then used to study the sample complexity of estimating θ_*. To study the log-likelihood landscape, we must perform this expansion for $R(\theta)$ itself. Our proof of this series expansion does not require G to be discrete (or θ_* to be generic), and we believe that this result may also be an important step in understanding the log-likelihood landscape for continuous group actions.

1.3. **Implications for optimization.** In this section, we discuss some implications of our results for descent-based optimization algorithms in high-noise settings.

Slow convergence of expectation-maximization. One of the most widely used optimization algorithms for minimizing $R_n(\theta)$ is expectation-maximization (EM) (see [DLR77], and [Sig98, SDCS10, BBS20] for applications in cryo-EM). Starting from an initialization $\theta^{(0)} \in \mathbb{R}^d$, the EM algorithm iteratively computes

$$\theta^{(t+1)} = \arg\min_{\theta \in \mathbb{R}^d} Q(\theta \mid \theta^{(t)})$$

where

$$Q(\theta \mid \theta^{(t)}) = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{g|Y_i,\theta^{(t)}} \left[\log \left(\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^d \exp\left(-\frac{\|Y_i - g\theta\|^2}{2\sigma^2} \right) \right) \right]$$

is the expectation of the full-data negative log-likelihood over the posterior law of $g \in G$. For each sample Y_i, the density of this posterior law is

$$p(g \mid Y_i, \theta^{(t)}) = \exp\left(-\frac{\|Y_i - g\theta^{(t)}\|^2}{2\sigma^2}\right) \Big/ \sum_{h\in G} \exp\left(-\frac{\|Y_i - h\theta^{(t)}\|^2}{2\sigma^2}\right),$$

leading to the following explicit form of the EM iteration:

$$\theta^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{g|Y_i,\theta^{(t)}}[g^\top Y_i].$$

It is straightforward to verify that this is equivalent to the gradient descent (GD) update

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla R_n(\theta^{(t)})$$

with a fixed step size $\eta = \sigma^2$.

Our results indicate that in the high-noise regime, this step size $\eta = \sigma^2$ corresponding to EM may not be correctly tuned for optimal convergence. For applying GD to a smooth and strongly convex function $f(\theta)$ where

$$\alpha \operatorname{Id} \preceq \nabla^2 f(\theta) \preceq \beta \operatorname{Id},$$

the optimal step size is $\eta \asymp 1/\beta$, and GD with this step size achieves a convergence rate

$$\|\theta^{(t)} - \theta^{(0)}\|^2 \le O\Big((1 - c\alpha/\beta)^t\Big) \tag{1.13}$$

for a constant $c > 0$ (see [Nes13, Theorem 2.1.14]). For any mean-zero group G, we have (by Lemma 4.9) that $d_1 = 0$ in the decomposition $d = d_1 + \ldots + d_L$ in (1.8), so that $\lambda_{\max}(\nabla^2 R_n(\theta)) \asymp \sigma^{-4}$ locally near θ_*. Thus there is a flattening of the landscape near θ_*, and GD should instead be tuned with the larger step size $\eta \asymp \sigma^4$ after reaching a small enough neighborhood of θ_*.

Figure 1.2 illustrates this for three-fold rotations in $\mathbb{R}^2$, comparing 250 iterations of EM versus GD with step size $\eta = \sigma^4$ on the high-noise example of Figure 1.1. EM converges quite slowly after reaching a vicinity of the circle $\{\theta \in \mathbb{R}^2 : \|\theta\| = \|\theta_*\|\}$, and the improved convergence rate for step size $\eta = \sigma^4$ is apparent.

Nesterov acceleration for gradient descent. The structure (1.8) for the eigenvalues of $I(\theta_*)$ also indicates that the Hessians of the risk functions $R_n(\theta)$ and $R(\theta)$ may be highly anisotropic and ill-conditioned near θ_* in high-noise settings. This poses a known problem for the convergence of gradient descent with any fixed step size, including EM, as evident from the factor α/β in (1.13).

This also suggests that substantial improvements in convergence may be obtained by using momentum or acceleration methods [Pol64, Nes13]. For example, using the Nesterov acceleration scheme

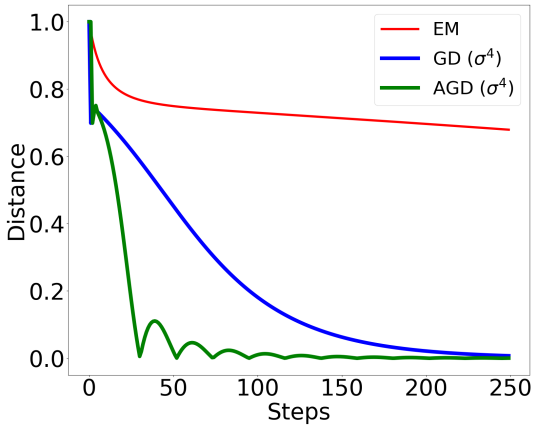
$$\begin{aligned}\mu^{(t+1)} &= \theta^{(t)} - \eta \cdot \nabla R_n(\theta^{(t)}) \\ \theta^{(t+1)} &= (1+\tau)\mu^{(t+1)} - \tau\mu^{(t)},\end{aligned}$$

accelerated gradient descent (AGD) can achieve the improved convergence rate

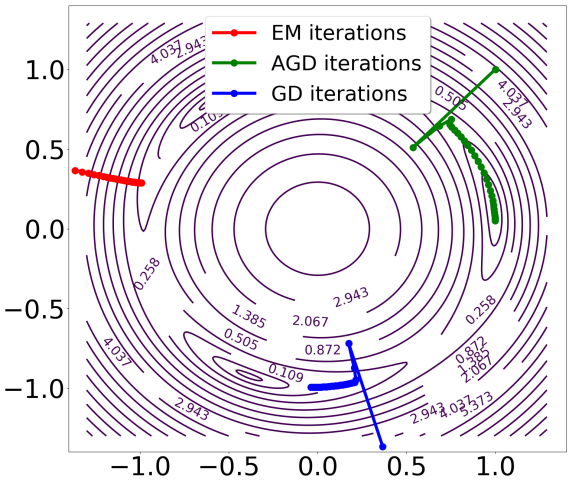
$$\|\theta^{(t)} - \theta^{(0)}\|^2 \le O\Big((1 - c\sqrt{\alpha/\beta})^t\Big), \tag{1.14}$$

see [Nes13, Theorem 2.2.3]. Figure 1.2 also illustrates the convergence of AGD on the same three-fold rotations example, with step size $\eta = \sigma^4$ and momentum parameters $\tau \equiv \tau_t$ defined as (see [Bub15, Section 3.7.2])

$$\lambda_0 = 0, \quad \lambda_t = \Big(1 + \sqrt{1 + 4\lambda_{t-1}^2}\Big)\Big/2, \quad \tau_t = (\lambda_t - 1)/\lambda_{t+1}.$$



(A) Distances $\mathrm{dist}(\theta^{(t)}, \mathcal{O}_{\theta_*})$ to the orbit of the true parameter $\theta_* = (1, 0)$, for 250 iterates $\theta^{(1)}, \ldots, \theta^{(250)}$ of each algorithm.



(B) First 30 iterates for each algorithm, depicted on the contour plot of the negative log-likelihood function $R_n(\theta)$. Iterates for EM and GD are rotated by angles of $2\pi/3$ and $4\pi/3$ for easier visualization.

FIGURE 1.2. Convergence of expectation-maximization (EM), gradient descent (GD) with step size $\eta = \sigma^4$, and Nesterov-accelerated gradient descent (AGD) with step size $\eta = \sigma^4$ on the three-fold rotations example with $n = 100{,}000$ samples and noise level $\sigma = 4$. All three algorithms are initialized at $\theta^{(0)} = (1, 1)$.

The iterates $\theta^{(t)}$ reach the orbit $\mathcal{O}_{\theta_*}$ within 30 iterations of AGD, when neither EM nor standard GD with $\eta = \sigma^4$ is close to having converged.

Reparametrization for second-order trust region methods. Second-order descent procedures may also be applied to minimize $R_n(\theta)$. Since R_n is non-convex, it is possible for its second-order approximation at an iterate $\theta^{(t)}$ to have a direction of negative curvature. When this occurs, it is common to apply a trust-region approach, where the next update $\theta^{(t+1)}$ is constrained to lie within a fixed-radius ball around $\theta^{(t)}$ [SQW15, SQW16, SQW18, MBM18]. This trust region is used until the iterates $\theta^{(t)}$ reach a neighborhood of strong convexity around a local minimizer of $R_n(\theta)$, after which the algorithm naturally transitions to a standard second-order Newton method for minimizing convex objectives.

At high noise, the region of convexity for $R(\theta)$ and $R_n(\theta)$ around θ_* may be vanishingly small in σ, requiring more careful tuning of this trust-region algorithm and a large number of iterations before reaching this convex region. However, as mentioned in Section 1.2, our results indicate that the region of convexity is much larger, and is σ-independent, upon reparametrizing by G-invariant coordinates $\varphi \equiv \varphi(\theta)$. This suggests that second-order methods may be more effective and stable when applied in the parametrization by φ, rather than the original parametrization by θ.

1.4. **Notation.** We write $\mathbb{E}_\varepsilon$ for the expectation over $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$. We write

$$\mathbb{E}_g[f(g)] = \frac{1}{K} \sum_{g \in G} f(g)$$

for the expectation over the uniform law $g \sim \mathrm{Unif}(G)$, and Var_g and Cov_g for the associated variance and covariance. Similarly $\mathbb{E}_h$ is the expectation over $h \sim \mathrm{Unif}(G)$, and $\mathbb{E}_{g_1, g_2}$ is the expectation over independent elements $g_1, g_2 \sim \mathrm{Unif}(G)$ unless stated otherwise.

We consider θ_*, d, G as constant throughout the paper. We write $C, C', c, c' > 0$ for constants that may depend on θ_*, d, G and change from instance to instance. These *do not* depend on the noise level σ, and we will be explicit about the dependence of our results on σ.

For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its gradient and Hessian by $\nabla f \in \mathbb{R}^d$ and $\nabla^2 f \in \mathbb{R}^{d\times d}$. More generally, we denote by $\nabla^k f \in (\mathbb{R}^d)^{\otimes k}$ the symmetric tensor of its k^{th} order partial derivatives. For a coordinate θ_i of θ, $\partial_{\theta_i} f$ is the partial derivative in θ_i. For $f : \mathbb{R}^d \to \mathbb{R}^k$, $\mathsf{d}f \in \mathbb{R}^{k\times d}$ is its full derivative (i.e. Jacobian matrix). When $k = 1$, we take the convention that ∇f is a column vector, so $\nabla f = \mathsf{d}f^\top$. We write ∇_θ, ∇^ℓ_θ, and d_θ to clarify that these are taken with respect to θ, and we write $\nabla_\theta f(\theta_*)$, $\nabla^\ell_\theta f(\theta_*)$, and $\mathsf{d}_\theta f(\theta_*)$ for their evaluations at $\theta = \theta_*$.

For a symmetric matrix $M \in \mathbb{R}^{d\times d}$, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ are its largest and smallest eigenvalues, and $\succeq$ and $\succ$ denote the positive-semidefinite and positive-definite ordering. For $\mu \in \mathbb{R}^d$ and $\rho > 0$, $B_\rho(\mu)$ is the open ℓ_2 ball of radius ρ around μ. $\|\cdot\|$ is the ℓ_2 norm for vectors and $\ell_2 \to \ell_2$ operator norm (largest singular value) for matrices, $\langle\cdot,\cdot\rangle$ is the ℓ_2 inner product, and $\|\cdot\|_{\text{HS}}$ is the vectorized ℓ_2 norm for higher-order tensors. $\operatorname{dist}(x, S) = \inf_{y\in S}\|x - y\|$ is the ℓ_2-distance from x to a set S. Id is the identity matrix, $\mathcal{N}(\cdot,\cdot)$ denotes the Gaussian distribution parametrized by mean and variance/covariance, and $[\ell] = \{1, \dots, \ell\}$.

For $\alpha = 1, 2$, denote by $\|W\|_{\psi_\alpha} = \inf\{t > 0 : \mathbb{E}_\varepsilon[\exp((|W|/t)^\alpha) \le 2\}$ the sub-exponential and sub-Gaussian norms of the random variable W. (See [Ver18, Chapter 2].)

Acknowledgments. We would like to thank Roy Lederman for helpful conversations at the onset of this work. Z. F. was supported in part by NSF Grant DMS-1916198. Y. S. was supported in part by a Junior Fellow award from the Simons Foundation and NSF Grant DMS-1701654. Y. W. was supported in part by NSF Grant CCF-1900507, NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship.

2. Preliminaries

This section collects several more basic results about the population risk $R(\theta)$ and its empirical counterpart $R_n(\theta)$, including expressions for their derivatives, bounds on critical points, and the concentration of $R_n(\theta)$ around $R(\theta)$.

2.1. The risk, gradient, and Hessian.

Let us first derive some simpler expressions for the risks $R_n(\theta)$ and $R(\theta)$. We represent each sample Y as

$$Y = h(\theta_* + \sigma\varepsilon)$$

where $h \sim \text{Unif}(G)$, $\varepsilon \sim \mathcal{N}(0, \text{Id})$, and they are independent. This is equivalent to the model (1.1), by the rotational invariance of the law of ε. Then the marginal log-likelihood (1.2) is given by

$$-\log p_\theta(Y) = -\log \mathbb{E}_g\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^d \exp\left(-\frac{\|h(\theta_* + \sigma\varepsilon) - g\theta\|^2}{2\sigma^2}\right)\right].$$

Applying $\|h(\theta_* + \sigma\varepsilon) - g\theta\| = \|\theta_* + \sigma\varepsilon - h^\top g\theta\|$ and the equality in law $h^\top g \overset{L}{=} g$ for any fixed $h \in G$, we have

$$\begin{aligned}-\log p_\theta(Y) &= -\log \mathbb{E}_g\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^d \exp\left(-\frac{\|\theta_* + \sigma\varepsilon - g\theta\|^2}{2\sigma^2}\right)\right] \\ &= \frac{d}{2}\log(2\pi\sigma^2) + \frac{\|\theta_* + \sigma\varepsilon\|^2}{2\sigma^2} + \frac{\|\theta\|^2}{2\sigma^2} - \log \mathbb{E}_g\left[\exp\left(\frac{\langle\theta_* + \sigma\varepsilon, g\theta\rangle}{\sigma^2}\right)\right].\end{aligned}$$

The first two terms above do not depend on θ, and we omit them in the sequel. We define the empirical risk as

$$R_n(\theta) = \frac{\|\theta\|^2}{2\sigma^2} - \frac{1}{n}\sum_{i=1}^n \log \mathbb{E}_g\left[\exp\left(\frac{\langle \theta_* + \sigma\varepsilon_i, g\theta\rangle}{\sigma^2}\right)\right]. \tag{2.1}$$

Then $R_n(\theta)$ is a constant shift of the negative log-likelihood for independent samples $Y_1, \dots, Y_n$, as stated in (1.3). We define the corresponding population risk $R(\theta) = \mathbb{E}[R_n(\theta)]$ by

$$R(\theta) = \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_\varepsilon\left[\log \mathbb{E}_g\left[\exp\left(\frac{\langle \theta_* + \sigma\varepsilon, g\theta\rangle}{\sigma^2}\right)\right]\right]. \tag{2.2}$$

Next, let us express the gradients, Hessians, and higher-order derivatives of these risk functions in terms of a reweighted law for $g \in G$. Given θ and ε, we introduce the reweighted probability law on G defined by

$$p(g \mid \varepsilon, \theta) = \exp\left(\frac{\langle \theta_* + \sigma\varepsilon, g\theta\rangle}{\sigma^2}\right) \Bigg/ \sum_{h\in G} \exp\left(\frac{\langle \theta_* + \sigma\varepsilon, h\theta\rangle}{\sigma^2}\right). \tag{2.3}$$

We write $\mathbb{E}_g[\cdot \mid \varepsilon, \theta]$, $\mathrm{Var}_g[\cdot \mid \varepsilon, \theta]$, and $\mathrm{Cov}_g[\cdot \mid \varepsilon, \theta]$ for the expectation, variance, and covariance with respect to this reweighted law of g. We also write $\kappa_g^\ell[\cdot \mid \varepsilon, \theta]$ for the ℓ^{th} cumulant tensor with respect to this law; see Appendix A.1 for the definition.

Lemma 2.1. *The derivatives of $R_n(\theta)$ take the forms*

$$\nabla R_n(\theta) = \frac{1}{\sigma^2}\left(\theta - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_g\left[g^\top(\theta_* + \sigma\varepsilon_i)\Big|\varepsilon_i, \theta\right]\right) \tag{2.4}$$

$$\nabla^2 R_n(\theta) = \frac{1}{\sigma^2}\left(\mathrm{Id} - \frac{1}{\sigma^2}\cdot\frac{1}{n}\sum_{i=1}^n \mathrm{Cov}_g\left[g^\top(\theta_* + \sigma\varepsilon_i)\Big|\varepsilon_i, \theta\right]\right) \tag{2.5}$$

$$\nabla^\ell R_n(\theta) = -\frac{1}{\sigma^{2\ell}}\cdot\frac{1}{n}\sum_{i=1}^n \kappa_g^\ell\left[g^\top(\theta_* + \sigma\varepsilon_i)\Big|\varepsilon_i, \theta\right] \quad \text{for } \ell \ge 3. \tag{2.6}$$

Proof. For any random vector $u \in \mathbb{R}^d$, the derivatives of its cumulant generating function are given by

$$\nabla_\theta^\ell \log \mathbb{E}[e^{\langle u, \theta\rangle}] = \kappa^\ell[u \mid \theta]$$

where $\kappa^\ell[u \mid \theta] \in (\mathbb{R}^d)^{\otimes \ell}$ is the ℓ^{th} cumulant tensor of u under its reweighted law defined by $\mathbb{E}[f(u) \mid \theta] = \mathbb{E}[f(u)e^{\langle u,\theta\rangle}]/\mathbb{E}[e^{\langle u,\theta\rangle}]$. (See Appendix A.1.) In particular, for $\ell = 1, 2$, these are the mean and covariance with respect to this law. Then (2.4–2.6) follow from differentiating (2.1) in θ, and applying this to the random vector $u = g^\top(\theta_* + \sigma\varepsilon_i)/\sigma^2$ conditional on ε_i. □

Lemma 2.2. *The derivatives of $R(\theta)$ take the forms*

$$\nabla R(\theta) = \frac{1}{\sigma^2}\left(\theta - \mathbb{E}_\varepsilon\left[\mathbb{E}_g\left[g^\top(\theta_* + \sigma\varepsilon)\Big|\varepsilon, \theta\right]\right]\right) \tag{2.7}$$

$$= \frac{1}{\sigma^2}\left(\mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g \mid \varepsilon, \theta]^\top \mathbb{E}_g[g \mid \varepsilon, \theta]\Big]\,\theta - \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g \mid \varepsilon, \theta]\Big]^\top \theta_*\right) \tag{2.8}$$

$$\nabla^2 R(\theta) = \frac{1}{\sigma^2}\left(\mathrm{Id} - \frac{1}{\sigma^2}\,\mathbb{E}_\varepsilon\Big[\mathrm{Cov}_g\left[g^\top(\theta_* + \sigma\varepsilon)\Big|\varepsilon, \theta\right]\Big]\right) \tag{2.9}$$

$$\nabla^\ell R(\theta) = -\frac{1}{\sigma^{2\ell}}\,\mathbb{E}_\varepsilon\Big[\kappa_g^\ell\Big[g^\top(\theta_* + \sigma\varepsilon)\Big|\varepsilon, \theta\Big]\Big] \quad \text{for } \ell \ge 3 \tag{2.10}$$

Proof. The identities (2.7), (2.9), and (2.10) are obtained by taking the expectations of (2.4–2.6) over $\varepsilon_1, \dots, \varepsilon_n$. (The derivatives of $R(\theta)$ in θ may be taken inside $\mathbb{E}_\varepsilon$ by a standard application of the dominated convergence theorem.)

For (2.8), we apply Gaussian integration by parts to rewrite the $\mathbb{E}_\varepsilon[\mathbb{E}_g[g^\top \varepsilon \mid \varepsilon, \theta]]$ term in (2.7): Denote by $g_{\cdot j}$ the j^{th} column of a matrix $g \in G$, and by g_{ij} the (i,j) entry. Then recalling the density (2.3) and applying the integration-by-parts identity $\mathbb{E}[f(\xi)\xi] = \mathbb{E}[f'(\xi)]$ for $\xi \sim \mathcal{N}(0,1)$, we get

$$\mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g_{\cdot j}^\top \varepsilon \mid \varepsilon, \theta]\Big] = \sum_{i=1}^d \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[p(g \mid \varepsilon, \theta) g_{ij}]\varepsilon_i\Big] = \sum_{i=1}^d \mathbb{E}_\varepsilon\Big[\partial_{\varepsilon_i} \mathbb{E}_g[p(g \mid \varepsilon, \theta) g_{ij}]\Big].$$

Write $(g\theta)_i$ as the i^{th} coordinate of $g\theta$, and note that differentiating (2.3) in ε_i gives

$$\partial_{\varepsilon_i} p(g \mid \varepsilon, \theta) = \frac{1}{\sigma}\Big(p(g \mid \varepsilon, \theta)(g\theta)_i - p(g \mid \varepsilon, \theta)\mathbb{E}_h[p(h \mid \varepsilon, \theta)(h\theta)_i]\Big)$$

where $h \sim \mathrm{Unif}(G)$ is independent of g. Then

$$\begin{aligned}
\sigma\, \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g_{\cdot j}^\top \varepsilon \mid \varepsilon, \theta]\Big] &= \sum_{i=1}^d \mathbb{E}_\varepsilon\Big[\operatorname{Cov}_g[g_{ij}, (g\theta)_i \mid \varepsilon, \theta]\Big] \\
&= \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g_{\cdot j}^\top g\theta \mid \varepsilon, \theta] - \mathbb{E}_g[g_{\cdot j} \mid \varepsilon, \theta]^\top \mathbb{E}_g[g\theta \mid \varepsilon, \theta]\Big] \\
&= \theta_j - \mathbb{E}_g[g_{\cdot j} \mid \varepsilon, \theta]^\top \mathbb{E}_g[g \mid \varepsilon, \theta]\theta,
\end{aligned}$$

the last line using $g_{\cdot j}^\top g\theta = \theta_j$ for any fixed orthogonal matrix $g \in G$. Combining this for $j = 1, \dots, d$,

$$\sigma\, \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g^\top \varepsilon \mid \varepsilon, \theta]\Big] = \theta - \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g \mid \varepsilon, \theta]^\top \mathbb{E}_g[g \mid \varepsilon, \theta]\Big]\theta.$$

Substituting into (2.7) yields (2.8). □

2.2. **Subgroup decompositions.** If the group G is the product of two groups G_1 and G_2 acting on orthogonal subspaces of $\mathbb{R}^d$, then both the empirical and population risks decompose as a sum corresponding to these two components. This is stated formally in the following lemma.

Lemma 2.3. *Let $V = [V_1 \mid V_2]$ be an orthogonal matrix, where $V_1 \in \mathbb{R}^{d \times d_1}$, $V_2 \in \mathbb{R}^{d \times d_2}$, and $d_1 + d_2 = d$. Suppose that $G \subset \mathrm{O}(d)$ decomposes as*

$$G = \left\{ V^\top \begin{pmatrix} g_1 & 0 \\ 0 & g_2 \end{pmatrix} V : g_1 \in G_1,\, g_2 \in G_2 \right\}$$

for subgroups $G_1 \subset \mathrm{O}(d_1)$ and $G_2 \subset \mathrm{O}(d_2)$, and write the corresponding decompositions $\theta_1 = V_1^\top \theta$, $\theta_2 = V_2^\top \theta$, $\theta_{1,*} = V_1^\top \theta_*$, $\theta_{2,*} = V_2^\top \theta_*$. Then

$$R_n(\theta) = R_n^{G_1}(\theta_1) + R_n^{G_2}(\theta_2) \qquad \text{and} \qquad R(\theta) = R^{G_1}(\theta_1) + R^{G_2}(\theta_2),$$

where $R_n^{G_1}$ and R^{G_1} denote the empirical and population risks (1.3) and (1.4) defined by G_1 and $\theta_{1,*}$ in dimension d_1, and similarly for G_2.

Proof. Note that $\|\theta\|^2 = \|\theta_1\|^2 + \|\theta_2\|^2$. Writing $g \in G$ as $g = V_1 g_1 V_1^\top + V_2 g_2 V_2^\top$, we have

$$\langle \theta_* + \sigma\varepsilon_i, g\theta \rangle = \langle \theta_{1,*} + \sigma V_1^\top \varepsilon_i, g_1\theta_1 \rangle + \langle \theta_{2,*} + \sigma V_2^\top \varepsilon_i, g_2\theta_2 \rangle.$$

The expectation $\mathbb{E}_g$ may be written as independent expectations over $g_1 \sim \mathrm{Unif}(G_1)$ and $g_2 \sim \mathrm{Unif}(G_2)$. Furthermore, $V_1^\top \varepsilon_i$ and $V_2^\top \varepsilon_i$ are independent Gaussian vectors of dimensions d_1 and d_2. Applying these to (2.1) yields $R_n(\theta) = R_n^{G_1}(\theta_1) + R_n^{G_2}(\theta_2)$. Taking the expectation yields $R(\theta) = R^{G_1}(\theta_1) + R^{G_2}(\theta_2)$. □

In particular, we may always reduce our study to a group G where $\mathbb{E}_g[g] = 0$, because of the following result. (Here $\mathbb{E}_g[g]$ is the expectation in $\mathbb{R}^{d \times d}$ when we consider $G \subset \mathrm{O}(d)$.)

Lemma 2.4. *Suppose $\mathbb{E}_g[g]$ has rank d_1 where $0 < d_1 \le d$, and set $d_2 = d - d_1$. Let $V = [V_1 \mid V_2]$ be an orthogonal matrix where the columns of $V_2 \in \mathbb{R}^{d\times d_2}$ span the kernel of $\mathbb{E}_g[g]$. Then*

$$G = \left\{ V^\top \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & g_2 \end{pmatrix} V : g_2 \in G_2 \right\} \tag{2.11}$$

where $G_2 \subset \mathrm{O}(d_2)$ is a subgroup that is group-isomorphic to G, and $\mathbb{E}_{g_2}[g_2] = 0$ for $g_2 \sim \mathrm{Unif}(G_2)$.

Proof. Observe that if $g \sim \mathrm{Unif}(G)$, then $g^\top = g^{-1} \sim \mathrm{Unif}(G)$, so $\mathbb{E}_g[g] = \mathbb{E}_g[g^\top] = \mathbb{E}_g[g]^\top$. Furthermore, if $g, h \sim \mathrm{Unif}(G)$ are independent, then $gh \sim \mathrm{Unif}(G)$, so $\mathbb{E}_g[g] = \mathbb{E}_{g,h}[gh] = \mathbb{E}_g[g]\mathbb{E}_h[h] = \mathbb{E}_g[g]^2$. Hence $\mathbb{E}_g[g]$ is symmetric and idempotent, so it is an orthogonal projection. For any θ in the range of this projection, $\theta = \mathbb{E}_g[g]\theta = \mathbb{E}_g[g\theta]$, so $\|\theta\|^2 = \theta^\top \mathbb{E}_g[g\theta]$. As each $g\theta$ is also a vector on the sphere of radius $\|\theta\|$, we have $\theta^\top g\theta < \|\theta\|^2$ unless $\theta = g\theta$. Thus, $\theta = g\theta$ for every $g \in G$, so G acts as the identity on the column span of V_1. This shows that each $g \in G$ has the form (2.11) for some matrix $g_2 \in \mathrm{O}(d_2)$, and this 1-to-1 mapping from g to g_2 must be a group isomorphism between G and G_2. Since G_2 represents the action of G on the column span of V_2, which is the kernel of $\mathbb{E}_g[g]$, we have $\mathbb{E}_{g_2}[g_2] = 0$. □

Combining Lemmas 2.3 and 2.4, we may always decompose $R_n(\theta) = R_n^{\mathrm{Id}}(\theta_1) + R_n^{G_2}(\theta_2)$ and $R(\theta) = R^{\mathrm{Id}}(\theta_1) + R^{G_2}(\theta_2)$, where θ_2 is the component of θ in the kernel of $\mathbb{E}_g[g]$. For θ_1, the risks $R_n^{\mathrm{Id}}(\theta_1)$ and $R^{\mathrm{Id}}(\theta_1)$ correspond to the single Gaussian model $\mathcal{N}(\theta_{1,*}, \sigma^2 \,\mathrm{Id})$. Then $R^{\mathrm{Id}}(\theta_1)$ and $R_n^{\mathrm{Id}}(\theta_1)$ are strongly convex, and our study of the landscapes of $R(\theta)$ and $R_n(\theta)$ reduces to studying $R^{G_2}(\theta_2)$ and $R_n^{G_2}(\theta_2)$ for the mean-zero group G_2.

2.3. **Generic parameters and critical points.** Throughout, we will assume that the true parameter $\theta_* \in \mathbb{R}^d$ is generic in the following sense.

Definition 2.5. For a connected open set $U \subseteq \mathbb{R}^d$, a statement holds for **generic** $\theta_* \in U$ if it holds for all θ_* outside the zero set of an analytic function $f : U \to \mathbb{R}^k$ that is not identically zero on U.

The zero set of any such analytic function has measure zero (see [Mit20]), so in particular, a statement that holds for generic $\theta_* \in \mathbb{R}^d$ holds everywhere outside a measure-zero subset of $\mathbb{R}^d$.

At a minimum, we will require that the points of the orbit $\mathcal{O}_{\theta_*}$ are distinct, so $|\mathcal{O}_{\theta_*}| = |G| = K$. This holds for generic θ_* because for any $g \ne h \in G$, the condition $(g-h)\theta_* = 0$ defines a subspace of dimension at most $d-1$.

Definition 2.6. For an open domain $U \subseteq \mathbb{R}^d$ and $f : U \to \mathbb{R}$ twice continuously differentiable, a point $\theta \in U$ is a **critical point** of f if $\nabla f(\theta) = 0$. The critical point is **non-degenerate** if $\nabla^2 f(\theta)$ is non-singular. The function f is **Morse** if all critical points are non-degenerate. The same definitions apply to $f : M \to \mathbb{R}$ for any manifold M, upon parametrizing M by a local chart.

A correspondence between non-degenerate critical points of a function $f_1 : U \to \mathbb{R}$ and those of a function f_2 uniformly close to f_1 was shown in [MBM18]. We will apply the following version of this result for only the local minimizers, which has a more elementary proof.

Lemma 2.7. *Let $\theta_0 \in \mathbb{R}^d$, and let $f_1, f_2 : B_\varepsilon(\theta_0) \to \mathbb{R}$ be two functions which are twice continuously differentiable. Suppose θ_0 is a critical point of f_1, and $\lambda_{\min}(\nabla^2 f_1(\theta)) \ge c_0$ for some $c_0 > 0$ and all $\theta \in B_\varepsilon(\theta_0)$. If*

$$|f_1(\theta) - f_2(\theta)| \le \delta \qquad \text{and} \qquad \|\nabla^2 f_1(\theta) - \nabla^2 f_2(\theta)\| \le \delta$$

for some $\delta < \min(c_0, c_0\varepsilon^2/4)$ and all $\theta \in B_\varepsilon(\theta_0)$, then f_2 has a unique critical point in $B_\varepsilon(\theta_0)$, which is a local minimizer of f_2.

Proof. The given conditions imply $\lambda_{\min}(\nabla^2 f_2(\theta)) > 0$ for all $\theta \in B_\varepsilon(\theta_0)$, so f_2 is strongly convex and has at most one critical point. They also imply that for each $\theta \in B_\varepsilon(\theta_0)$ with $\|\theta - \theta_0\| = r$,

$$f_2(\theta) - f_2(\theta_0) \ge f_1(\theta) - f_1(\theta_0) - 2\delta \ge \frac{c_0 r^2}{2} - 2\delta.$$

For r sufficiently close to ε, we have $c_0 r^2/2 - 2\delta > 0$. Then f_2 must have a local minimizer in $B_r(\theta_0)$. □

2.4. **Bounds for critical points.** A consequence of (2.4) and (2.7) is the following simple bound for critical points of $R(\theta)$ and $R_n(\theta)$.

Lemma 2.8. *For d-dependent constants $C, C', c > 0$, we have $\sigma^2\|\nabla R(\theta)\| \geq \|\theta\| - \|\theta_*\| - C\sigma$, and $\sigma^2\|\nabla R_n(\theta)\| \geq \|\theta\| - \|\theta_*\| - C\sigma$ with probability at least $1 - C'e^{-cn}$. In particular, any critical point θ of $R(\theta)$ satisfies $\|\theta\| \leq \|\theta_*\| + C\sigma$, and the same holds for $R_n(\theta)$ with probability $1 - C'e^{-cn}$.*

Proof. The bound for $\|\nabla R(\theta)\|$ follows from (2.7) and

$$\Big\|\mathbb{E}_\varepsilon\big[\mathbb{E}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \sigma]\big]\Big\| \leq \mathbb{E}_\varepsilon[\|\theta_* + \sigma\varepsilon\|] \leq \|\theta_*\| + \sigma\,\mathbb{E}_\varepsilon[\|\varepsilon\|] \leq \|\theta_*\| + \sigma\sqrt{d}.$$

The bound for $\|\nabla R_n(\theta)\|$ follows similarly from (2.1), on the event $n^{-1}\sum_{i=1}^n \|\varepsilon_i\| \leq C$ which has probability at least $1 - C'e^{-cn}$ by Hoeffding's inequality for sub-Gaussian random variables (see [Ver18, Theorem 2.6.2]). Since $\nabla R(\theta) = 0$ at a critical point θ, and similarly for $R_n(\theta)$, the statements for critical points follow. □

When σ is large, this bound is not sharp in its dependence on σ. We will in fact show that any critical point θ of $R(\theta)$ satisfies $\|\theta\| \leq C$ for a σ-independent constant $C > 0$. The following strengthening of Lemma 2.8 first provides the a-priori bound $\|\theta\| \leq C\sigma^{2/3}$. Then, combined with a series expansion of $R(\theta)$ in σ^{-2}, we will improve this to $\|\theta\| \leq C$ in Lemma 4.17 of Section 4.

Lemma 2.9. *For some (θ_*, d, G)-dependent constants $C, c, \sigma_0 > 0$ and all $\sigma > \sigma_0$,*

$$\sigma^2\|\nabla R(\theta)\| > c\min\left(\frac{\|\theta\|^3}{\sigma^2}, \frac{\|\theta\|}{\sigma^{2/3}}\right) - \|\theta_*\|, \tag{2.12}$$

and every critical point θ of $R(\theta)$ satisfies $\|\theta\| < C\sigma^{2/3}$.

Proof. We apply the form of $\nabla R(\theta)$ given in (2.8). Denote $\bar{Y} = (\theta_* + \sigma\varepsilon)/\|\theta_* + \sigma\varepsilon\|$ and $\bar{\theta} = \theta/\|\theta\|$. Then

$$\begin{aligned}
\sigma^2\|\nabla R(\theta)\| &\geq \langle \bar{\theta}, \sigma^2\nabla R(\theta)\rangle \\
&\geq \bar{\theta}^\top\mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g \mid \varepsilon, \theta]^\top\mathbb{E}_g[g \mid \varepsilon, \theta]\Big]\theta - \bar{\theta}^\top\mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g \mid \varepsilon, \theta]^\top\Big]\theta_* \\
&= \|\theta\| \cdot \mathbb{E}_\varepsilon\Big[\|\mathbb{E}_g[g\bar{\theta} \mid \varepsilon, \theta]\|^2\Big] - \bar{\theta}^\top\mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g \mid \varepsilon, \theta]^\top\Big]\theta_* \\
&\geq \|\theta\| \cdot \mathbb{E}_\varepsilon\Big[(\bar{Y}^\top\mathbb{E}_g[g\bar{\theta} \mid \varepsilon, \theta])^2\Big] - \|\theta_*\| \\
&= \|\theta\| \cdot \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[\bar{Y}^\top g\bar{\theta} \mid \varepsilon, \theta]^2\Big] - \|\theta_*\|.
\end{aligned} \tag{2.13}$$

We analyze the quantity $\mathbb{E}_g[\bar{Y}^\top g\bar{\theta} \mid \varepsilon, \theta]$ for fixed ε (and hence fixed $\bar{Y}$): Note that $|\bar{Y}^\top g\bar{\theta}| \leq 1$. Let $K(s)$ be the cumulant generating function of $\bar{Y}^\top g\bar{\theta}$ over the uniform law $g \sim \mathrm{Unif}(G)$, and let $K'(s)$ be its derivative. Denote

$$t \equiv t(\varepsilon, \theta) = \frac{\|\theta_* + \sigma\varepsilon\|\|\theta\|}{\sigma^2}.$$

Then

$$\mathbb{E}_g[\bar{Y}^\top g\bar{\theta} \mid \varepsilon, \theta] = \mathbb{E}_g[p(g \mid \varepsilon, \theta)\bar{Y}^\top g\bar{\theta}] = \frac{\mathbb{E}_g[\bar{Y}^\top g\bar{\theta} \cdot e^{t\bar{Y}^\top g\bar{\theta}}]}{\mathbb{E}_g[e^{t\bar{Y}^\top g\bar{\theta}}]} = \frac{d}{ds}\log\mathbb{E}_g[e^{s\bar{Y}^\top g\bar{\theta}}]\Big|_{s=t} = K'(t). \tag{2.14}$$

Writing κ_ℓ as the ℓ^{th} cumulant of this law, we have

$$K(s) = \sum_{\ell=1}^\infty \kappa_\ell\frac{s^\ell}{\ell!}, \tag{2.15}$$

where this series is absolutely convergent for $|s| < 1/e$ by Lemma A.1. Set

$$t_\sigma \equiv t_\sigma(\varepsilon, \theta) = \min(t(\varepsilon, \theta), \sigma^{-1/3}),$$

where $t_\sigma < 1/e$ for $\sigma > \sigma_0$ and large enough σ_0. Since $K(0) = 0$, using the convexity of the cumulant generating function K we can bound its derivative from below by

$$K'(t) \geq K'(t_\sigma) \geq \frac{K(t_\sigma)}{t_\sigma} = \sum_{\ell=1}^{\infty} \kappa_\ell \frac{t_\sigma^{\ell-1}}{\ell!}.$$

Applying $|\kappa_\ell| \leq \ell^\ell$ from Lemma A.1 and $\ell! \geq \ell^\ell/e^\ell$,

$$K'(t) \geq \kappa_1 + \frac{t_\sigma}{2}\kappa_2 - \sum_{\ell=3}^{\infty} e^\ell t_\sigma^{\ell-1} \geq \kappa_1 + \frac{t_\sigma}{2}\kappa_2 - 30t_\sigma^2$$

for $\sigma > \sigma_0$ and large enough σ_0. Here, $\kappa_1 = \mathbb{E}_g[\bar{Y}^\top g\bar{\theta}]$ and $\kappa_2 = \mathrm{Var}_g[\bar{Y}^\top g\bar{\theta}]$.

Now observe that there exists a constant $c_0 \equiv c_0(d) > 0$, such that if v is any random vector on the unit sphere in $\mathbb{R}^d$, then there is a deterministic vector u_0 on the unit sphere for which

$$\min(\mathbb{E}[u_0^\top v], \mathrm{Var}[u_0^\top v]) > 2c_0.$$

This is because if the mean of v is near 0 and v lies on the sphere, then the variance of v must be bounded below by a constant in some direction. Then also for some $\delta_0 > 0$ depending only on c_0, we have

$$\min(\mathbb{E}[u^\top v], \mathrm{Var}[u^\top v]) > c_0 \text{ for all } u \in B_{\delta_0}(u_0).$$

Let us apply this to the random vector $v = g\bar{\theta}$ under the uniform law of g. (So u_0 depends on G and θ.) Then for $\sigma > \sigma_0$, on the event $\bar{Y} \in B_{\delta_0}(u_0)$, we get

$$K'(t) \geq \frac{c_0}{2}t_\sigma - 30t_\sigma^2 \geq \frac{c_0}{3}t_\sigma.$$

Recalling (2.14) and applying this to (2.13),

$$\begin{aligned}\sigma^2\|\nabla R(\theta)\| &\geq \|\theta\| \cdot \mathbb{E}_\varepsilon\left[\left(\frac{c_0}{3}t_\sigma(\varepsilon,\theta)\right)^2 \mathbf{1}\{\bar{Y} \in B_{\delta_0}(u_0)\}\right] - \|\theta_*\| \\ &\geq \|\theta\| \cdot \mathbb{E}_\varepsilon\left[\left(\frac{c_0}{3}t_\sigma(\varepsilon,\theta)\right)^2 \mathbf{1}\{\bar{Y} \in B_{\delta_0}(u_0), \|\theta_* + \sigma\varepsilon\| \geq \sigma\}\right] - \|\theta_*\|.\end{aligned}$$

On the event $\|\theta_* + \sigma\varepsilon\| \geq \sigma$, we have $t(\varepsilon, \theta) \geq \|\theta\|/\sigma$, so $t_\sigma(\varepsilon, \theta) \geq \min(\|\theta\|/\sigma, \sigma^{-1/3})$. Then

$$\sigma^2\|\nabla R(\theta)\| \geq \frac{c_0^2}{9}\min\left(\frac{\|\theta\|^3}{\sigma^2}, \frac{\|\theta\|}{\sigma^{2/3}}\right)\mathbb{P}\Big[\bar{Y} \in B_{\delta_0}(u_0),\, \|\theta_* + \sigma\varepsilon\| \geq \sigma\Big] - \|\theta_*\|.$$

Recalling the definition $\bar{Y} = (\theta_* + \sigma\varepsilon)/\|\theta_* + \sigma\varepsilon\|$, as $\sigma \to \infty$, we have

$$\mathbb{P}\Big[\bar{Y} \in B_{\delta_0}(u_0),\, \|\theta_* + \sigma\varepsilon\| \geq \sigma\Big] \to \mathbb{P}\Big[\varepsilon/\|\varepsilon\| \in B_{\delta_0}(u_0),\, \|\varepsilon\| \geq 1\Big].$$

Since $\varepsilon/\|\varepsilon\|$ is uniformly distributed on the sphere, the limit is a positive constant depending only on the dimension d and δ_0. Furthermore, for fixed θ_*, this convergence is uniform over u_0 on the unit sphere. Thus we obtain

$$\mathbb{P}\Big[\bar{Y} \in B_{\delta_0}(u_0),\, \|\theta_* + \sigma\varepsilon\| \geq \sigma\Big] \geq c$$

for a constant $c \equiv c(d)$ and all $\sigma > \sigma_0(\theta_*, d, G)$. This yields (2.12). For a large enough constant $C \equiv C(\theta_*, d, G) > 0$, this implies $\|\nabla R(\theta)\| > 0$ when $\|\theta\| \geq C\sigma^{2/3}$, so any critical point satisfies $\|\theta\| < C\sigma^{2/3}$. □

2.5. **Concentration of the empirical risk.** We establish uniform concentration of $R_n(\theta)$, $\nabla R_n(\theta)$, and $\nabla^2 R_n(\theta)$ around their expectations. This will allow us to translate results about the population landscape of $R(\theta)$ to the empirical landscape of $R_n(\theta)$.

Lemma 2.10. *There exist (θ_*, d, G)-dependent constants $C, c > 0$ such that for any $r, t > 0$, denoting $B_r \equiv B_r(0) = \{\theta \in \mathbb{R}^d : \|\theta\| < r\}$,*

$$\mathbb{P}\Big[\sup_{\theta\in B_r} |R_n(\theta) - R(\theta)| \ge t\Big] \le \Big(\tfrac{Cr(1+\sigma)}{\sigma^2 t}\Big)^d \exp\Big(-cn\,\tfrac{\sigma^2 t^2}{r^2}\Big) + Ce^{-cn} \tag{2.16}$$

$$\mathbb{P}\Big[\sup_{\theta\in B_r} \|\nabla R_n(\theta) - \nabla R(\theta)\| \ge t\Big] \le \Big(\tfrac{Cr(1+\sigma^2)}{\sigma^4 t}\Big)^d \exp\Big(-cn\,\tfrac{\sigma^4 t^2}{1+\sigma^2}\Big) + Ce^{-cn} \tag{2.17}$$

$$\mathbb{P}\Big[\sup_{\theta\in B_r} \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \ge t\Big] \le \Big(\tfrac{Cr(1+\sigma^3)}{\sigma^6 t}\Big)^d \exp\Big(-cn\min\Big(\tfrac{\sigma^8 t^2}{1+\sigma^4}, \tfrac{\sigma^4 t}{1+\sigma^2}\Big)\Big) + Ce^{-cn^{2/3}} \tag{2.18}$$

We prove this by first showing pointwise concentration in Lemma 2.11, then establishing Lipschitz continuity of these risks, gradients, and Hessians in Lemma 2.12, and finally applying a covering net argument.

Lemma 2.11. *For some (θ_*, d, G)-dependent constants $C, c > 0$, any $\theta \in \mathbb{R}^d$, and any $t > 0$,*

$$\mathbb{P}\left[|R_n(\theta) - R(\theta)| \ge t\right] \le C\exp\Big(-cn\,\tfrac{\sigma^2 t^2}{\|\theta\|^2}\Big) \tag{2.19}$$

$$\mathbb{P}\left[\|\nabla R_n(\theta) - \nabla R(\theta)\| \ge t\right] \le C\exp\Big(-cn\,\tfrac{\sigma^4 t^2}{1+\sigma^2}\Big) \tag{2.20}$$

$$\mathbb{P}\left[\|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \ge t\right] \le C\exp\Big(-cn\min\Big(\tfrac{\sigma^8 t^2}{1+\sigma^4}, \tfrac{\sigma^4 t}{1+\sigma^2}\Big)\Big). \tag{2.21}$$

Proof. We apply the Bernstein and Hoeffding inequalities. Recall that for $\alpha = 1$ or 2, $\|f(\varepsilon)\|_{\psi_\alpha}$ denotes the sub-exponential or sub-Gaussian norm of the random variable $f(\varepsilon)$ over the law $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$.

For $R_n(\theta)$, recall the form (2.1). Set

$$f_1(\varepsilon) = \log \mathbb{E}_g\left[\exp\left(\frac{\langle \theta_* + \sigma\varepsilon, g\theta\rangle}{\sigma^2}\right)\right].$$

Then $\nabla_\varepsilon f_1(\varepsilon) = \mathbb{E}_g[g\theta \mid \varepsilon, \theta]/\sigma$, so $\|\nabla_\varepsilon f_1(\varepsilon)\| \le \mathbb{E}_g[\|g\theta\| \mid \varepsilon, \theta]/\sigma \le \|\theta\|/\sigma$ and f_1 is $\|\theta\|/\sigma$-Lipschitz. By Gaussian concentration of measure and Hoeffding's inequality (see [Ver18, Theorems 2.6.2, 5.2.2]), for constants $C, c > 0$ and any $t > 0$,

$$\|f_1(\varepsilon) - \mathbb{E}_\varepsilon f_1(\varepsilon)\|_{\psi_2} \le \frac{C\|\theta\|}{\sigma}, \quad \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n f_1(\varepsilon_i) - \mathbb{E}_\varepsilon[f_1(\varepsilon)]\right| \ge t\right] \le 2\exp\left(-cn\,\frac{\sigma^2 t^2}{\|\theta\|^2}\right).$$

Applying this to (2.1) yields (2.19).

For $\nabla R_n(\theta)$, recall (2.4). Denote by $g_{\cdot j}$ the jth column of g. Momentarily fixing j, denote

$$f_2(\varepsilon) = \mathbb{E}_g\left[g_{\cdot j}^\top(\theta_* + \sigma\varepsilon)\Big|\varepsilon, \theta\right], \qquad f_{2,g}(\varepsilon) = g_{\cdot j}^\top(\theta_* + \sigma\varepsilon)$$

where $f_{2,g}$ is defined for each fixed $g \in G$. Then

$$\|f_2(\varepsilon)\|_{\psi_2} = \left\|\sum_{g\in G} p(g \mid \varepsilon, \theta) f_{2,g}(\varepsilon)\right\|_{\psi_2} \le K \cdot \max_{g\in G}\left\|p(g \mid \varepsilon, \theta) f_{2,g}(\varepsilon)\right\|_{\psi_2} \le K \cdot \max_{g\in G}\|f_{2,g}(\varepsilon)\|_{\psi_2},$$

the last inequality applying $|p(g \mid \varepsilon, \theta)| \le 1$ and the definition of the sub-Gaussian norm. For each fixed $g \in G$, we have $\|f_{2,g}(\varepsilon)\|_{\psi_2} \le C(1+\sigma)$. Then by Hoeffding's inequality,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n f_2(\varepsilon_i) - \mathbb{E}_\varepsilon[f_2(\varepsilon)]\right| > t\right] \le 2\exp\left(-cn\,\frac{t^2}{(1+\sigma)^2}\right).$$

This establishes concentration of the jth coordinate of $R_n(\theta)$. Applying a union bound over indices $j = 1, \dots, d$ and replacing t by $\sigma^2 t$ yields (2.20).

For $\nabla^2 R_n(\theta)$, recall (2.5). Momentarily fixing the indices j and k, denote

$$\begin{aligned} f_3(\varepsilon) &= \mathrm{Cov}_g\left[g_{\cdot j}^\top(\theta_* + \sigma\varepsilon),\ g_{\cdot k}^\top(\theta_* + \sigma\varepsilon)\middle|\varepsilon, \theta\right] \\ &= \mathbb{E}_g\left[g_{\cdot j}^\top(\theta_* + \sigma\varepsilon) \cdot g_{\cdot k}^\top(\theta_* + \sigma\varepsilon)\middle|\varepsilon, \theta\right] - \mathbb{E}_g\left[g_{\cdot j}^\top(\theta_* + \sigma\varepsilon)\middle|\varepsilon, \theta\right] \cdot \mathbb{E}_g\left[g_{\cdot k}^\top(\theta_* + \sigma\varepsilon)\middle|\varepsilon, \theta\right] \end{aligned}$$

Using the same argument as above, we have the bounds

$$\left\|\mathbb{E}_g\left[g_{\cdot j}^\top(\theta_* + \sigma\varepsilon) \cdot g_{\cdot k}^\top(\theta_* + \sigma\varepsilon)\middle|\varepsilon, \theta\right]\right\|_{\psi_1} \le C(1+\sigma^2), \quad \left\|\mathbb{E}_g\left[g_{\cdot j}^\top(\theta_* + \sigma\varepsilon)\middle|\varepsilon, \theta\right]\right\|_{\psi_2} \le C(1+\sigma).$$

Together with the inequality $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$, this yields $\|f_3(\varepsilon)\|_{\psi_1} \le C(1+\sigma^2)$. Then by Bernstein's inequality (see [Ver18, Theorem 2.8.1]),

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n f_3(\varepsilon_i) - \mathbb{E}_\varepsilon[f_3(\varepsilon)]\right| > t\right] \le 2\exp\left(-cn\min\left(\frac{t^2}{(1+\sigma^2)^2}, \frac{t}{1+\sigma^2}\right)\right).$$

This establishes concentration of the (j,k) entry of $\nabla^2 R_n(\theta)$. Taking a union bound over $j, k \in \{1, \dots, d\}$ and replacing t by $\sigma^4 t$ yields (2.21). □

Lemma 2.12. *For a (θ_*, d, G)-dependent constant $C' > 0$, as functions over $\theta \in \mathbb{R}^d$,*

(a) $R(\theta) - \|\theta\|^2/(2\sigma^2)$ is $C'(1+\sigma)/\sigma^2$-Lipschitz.
(b) Each entry of $\nabla R(\theta) - \theta/\sigma^2$ is $C'(1+\sigma^2)/\sigma^4$-Lipschitz.
(c) Each entry of $\nabla^2 R(\theta) - \mathrm{Id}/\sigma^2$ is $C'(1+\sigma^3)/\sigma^6$-Lipschitz.

For d-dependent constants $C, c > 0$, statements (a) and (b) also hold for $R_n(\theta) - \|\theta\|^2/(2\sigma^2)$ and $\nabla R_n(\theta) - \theta/\sigma^2$ with probability at least $1 - Ce^{-cn}$, and (c) holds for $\nabla^2 R_n(\theta) - \mathrm{Id}/\sigma^2$ with probability at least $1 - Ce^{-cn^{2/3}}$.

Proof. To prove the desired Lipschitz property, it suffices to bound the first three derivatives of $R(\theta)$. Recall the expressions (2.7), (2.9), and (2.10) for $\nabla^\ell R(\theta)$. Note that $\|g^\top(\theta_* + \sigma\varepsilon)\| = \|\theta_* + \sigma\varepsilon\|$. Thus, under the law (2.3), each entry of $g^\top(\theta_* + \sigma\varepsilon)$ has magnitude at most $\|\theta_* + \sigma\varepsilon\|$. Invoking Lemma A.1(b), we conclude that for each $\ell \ge 1$ and some constant $C \equiv C(\ell, d, \|\theta_*\|)$,

$$\|\kappa_g^\ell[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta]\|_{\mathrm{HS}} \le C(1 + \sigma^\ell\|\varepsilon\|^\ell)$$

where $\ell = 1, 2$ for the mean and covariance.

Applying these bounds to (2.7), (2.9), and (2.10) and taking the expectation over $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$ yields the Lipschitz properties for the population risk $R(\theta)$. Recalling the forms (2.4–2.6), this also shows the Lipschitz properties for the empirical risk $R_n(\theta)$ on the events

$$\mathcal{E}^\alpha = \left\{\frac{1}{n}\sum_{i=1}^n \|\varepsilon_i\|^\alpha \le C_0\right\}$$

for $\alpha = 1, 2, 3$ respectively, where $C_0 > 0$ is any fixed constant. For $\alpha = 1, 2$ and a sufficiently large constant $C_0 > 0$, we have $\mathbb{P}[\mathcal{E}^\alpha] \ge 1 - Ce^{-cn}$ by the Hoeffding and Bernstein inequalities. For $\alpha = 3$, we show in Appendix A.3 using the result of [AW15] that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \|\varepsilon_i\|^3 \le C_0\right] \ge 1 - Ce^{-cn^{2/3}} \tag{2.22}$$

for a sufficiently large constant $C_0 > 0$. (Note that this bound is optimal, by considering the deviation of a single summand $n^{-1}\|\varepsilon_i\|^3$.) This concludes the proof. □

Proof of Lemma 2.10. Denote $\bar{R}_n(\theta) = R_n(\theta) - \|\theta\|^2/(2\sigma^2)$ and $\bar{R}(\theta) = R(\theta) - \|\theta\|^2/(2\sigma^2)$. Note that concentration of $R_n(\theta), \nabla R_n(\theta), \nabla^2 R_n(\theta)$ is equivalent to that of $\bar{R}_n(\theta), \nabla \bar{R}_n(\theta), \nabla^2 \bar{R}_n(\theta)$.

For $\bar{R}_n(\theta)$, we take a δ-net N of B_r having cardinality $|N| \le (Cr/\delta)^d$. Applying (2.19) and a union bound over N,

$$\mathbb{P}\left[\sup_{\mu\in N} \left|\bar{R}_n(\mu) - \bar{R}(\mu)\right| \ge t/3\right] \le \left(\frac{Cr}{\delta}\right)^d \exp\left(-cn\,\frac{\sigma^2 t^2}{r^2}\right).$$

By the Lipschitz bounds for $\bar{R}(\theta)$ and $\bar{R}_n(\theta)$ in Lemma 2.12, picking $\delta = c\sigma^2 t/(1+\sigma)$ for a small enough constant $c > 0$ ensures on an event of probability $1 - Ce^{-cn}$ that $|\bar{R}(\theta) - \bar{R}(\mu)| \le t/3$ and $|\bar{R}_n(\theta) - \bar{R}_n(\mu)| \le t/3$ for each point $\theta \in B_r$ and the closest point $\mu \in N$. Combining these shows (2.16). The bounds (2.17) and (2.18) are obtained similarly. □

3. Landscape analysis for low noise

In this section, we analyze the function landscapes of $R(\theta)$ and $R_n(\theta)$ in the low-noise regime $\sigma < \sigma_0(\theta_*, d, G)$. Section 3.1 analyzes the local landscapes in a neighborhood of θ_*, as well as the Fisher information $I(\theta_*) = \nabla^2_\theta R(\theta_*)$, and Theorem 3.1 shows that these behave similarly to a single-component Gaussian model $\mathcal{N}(\theta_*, \sigma^2 \operatorname{Id})$. Section 3.2 analyzes the global landscapes, and Theorem 3.3 and Corollary 3.5 show that these are globally benign for small σ and large n.

3.1. **Local landscape and Fisher information.**

Theorem 3.1. *For any $\theta_* \in \mathbb{R}^d$ where $|\mathcal{O}_{\theta_*}| = |G| = K$, there exist (θ_*, d, G)-dependent constants $\sigma_0, c, \rho > 0$ such that as long as $\sigma < \sigma_0$, every $\theta \in B_\rho(\theta_*)$ satisfies*

$$\|\nabla^2 R(\theta) - \sigma^{-2} \operatorname{Id}\| < e^{-c/\sigma^2}. \tag{3.1}$$

In particular, the Fisher information satisfies $\|I(\theta_*) - \sigma^{-2}\operatorname{Id}\| < e^{-c/\sigma^2}$.

Note that by rotational symmetry of $R(\theta)$, the same statements hold for $B_\rho(\mu)$ and each $\mu \in \mathcal{O}_{\theta_*}$.

Proof. Since the K points of $\mathcal{O}_{\theta_*}$ are distinct and have the same norm, we must have $\|\theta_*\|^2 > \theta_*^\top \mu$ for each $\mu \in \mathcal{O}_{\theta_*}$ different from θ_*. Pick (θ_*-dependent) constants $c_0, \rho > 0$ such that $(\theta_* - \mu)^\top \theta_* > 3c_0$ and $\|\theta_* - \mu\|\rho < c_0$ for all such μ, and also $\rho < \|\theta_*\|/2$. Define

$$\mathcal{E} = \{\varepsilon \in \mathbb{R}^d : 2\sigma\|\varepsilon\|\|\theta\| \le c_0\}. \tag{3.2}$$

Consider $\theta \in B_\rho(\theta_*)$, and recall the form (2.9) for $\nabla^2 R(\theta)$. For any unit vector $v \in \mathbb{R}^d$, we have

$$\begin{aligned} v^\top \mathbb{E}_\varepsilon\Big[\operatorname{Cov}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta]\Big] v &= \mathbb{E}_\varepsilon\Big[\operatorname{Var}_g[\langle v, g^\top(\theta_* + \sigma\varepsilon)\rangle \mid \varepsilon, \theta]\Big] \\ &= \mathbb{E}_\varepsilon\Big[\operatorname{Var}_g[\langle gv, \theta_* + \sigma\varepsilon\rangle \mid \varepsilon, \theta]\Big] \le \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[\langle gv - v,\, \theta_* + \sigma\varepsilon\rangle^2 \mid \varepsilon, \theta]\Big]. \end{aligned}$$

Let us decompose the last line as $\mathrm{I} + \mathrm{II}$ where

$$\begin{aligned} \mathrm{I} &= \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\varepsilon \notin \mathcal{E}\}\mathbb{E}_g[\langle gv - v,\, \theta_* + \sigma\varepsilon\rangle^2 \mid \varepsilon, \theta]\Big], \\ \mathrm{II} &= \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\varepsilon \in \mathcal{E}\}\mathbb{E}_g[\langle gv - v,\, \theta_* + \sigma\varepsilon\rangle^2 \mid \varepsilon, \theta]\Big]. \end{aligned}$$

For I, we have $\|\theta\| \le \|\theta_*\| + \rho$. Applying the chi-squared tail bound $\mathbb{P}[\|\varepsilon\|^2 > t] < e^{-ct}$ for all $t > C$, we get $\mathbb{P}[\varepsilon \notin \mathcal{E}] < e^{-c/\sigma^2}$. Then by Cauchy-Schwarz,

$$\begin{aligned} \mathrm{I} &\le \mathbb{P}[\varepsilon \notin \mathcal{E}]^{1/2}\mathbb{E}_\varepsilon[\mathbb{E}_g[\langle gv - v,\, \theta_* + \sigma\varepsilon\rangle^2 \mid \varepsilon, \theta]^2]^{1/2} \\ &\le \mathbb{P}[\varepsilon \notin \mathcal{E}]^{1/2}\mathbb{E}_\varepsilon[(2\|\theta_* + \sigma\varepsilon\|)^4]^{1/2} < e^{-c'/\sigma^2} \end{aligned}$$

for constants $c', \sigma_0 > 0$ and all $\sigma < \sigma_0$. For II, let us bound $\mathbb{P}[g \neq \mathrm{Id} \mid \varepsilon, \theta]$ when $\varepsilon \in \mathcal{E}$: For any $g \neq \mathrm{Id}$, letting $\mu = g^\top \theta_*$,

$$\langle \theta_* + \sigma\varepsilon,\, \theta - g\theta \rangle \geq (\theta_* - \mu)^\top \theta - 2\sigma\|\varepsilon\|\|\theta\| \geq (\theta_* - \mu)^\top \theta_* - 2\sigma\|\varepsilon\|\|\theta\| - \|\theta_* - \mu\|\rho > c_0.$$

Then recalling (2.3), $p(\mathrm{Id} \mid \varepsilon, \theta)/p(g \mid \varepsilon, \theta) > e^{c_0/\sigma^2}$ and so

$$p(\mathrm{Id} \mid \varepsilon, \theta) > e^{c_0/\sigma^2}/(e^{c_0/\sigma^2} + K - 1) > 1 - e^{-c/\sigma^2} \tag{3.3}$$

for constants $c, \sigma_0 > 0$ and all $\sigma < \sigma_0$. Thus $\mathbb{P}[g \neq \mathrm{Id} \mid \varepsilon, \theta] = 1 - p(\mathrm{Id} \mid \varepsilon, \theta) < e^{-c/\sigma^2}$, so

$$\mathrm{II} \leq \mathbb{E}_\varepsilon[\mathbf{1}\{\varepsilon \in \mathcal{E}\}\mathbb{P}[g \neq \mathrm{Id} \mid \varepsilon, \theta] \cdot (2\|\theta_* + \sigma\varepsilon\|)^2] < e^{-c'/\sigma^2}.$$

Combining these, we get $v^\top \mathbb{E}_\varepsilon[\mathrm{Cov}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta]]v < e^{-c/\sigma^2}$ for any unit vector $v \in \mathbb{R}^d$. Then (3.1) follows from (2.9). Specializing to $\theta = \theta_*$ yields the statement for $I(\theta_*)$. □

The following corollary then shows that with high probability when $n \gg \sigma^{-1}\log\sigma^{-1}$, the empirical risk $R_n(\theta)$ is strongly convex with a unique local minimizer in $B_\rho(\theta_*)$. By rotational symmetry, the same statement holds for $B_\rho(\mu)$ and each $\mu \in \mathcal{O}_{\theta_*}$.

Corollary 3.2. *For some (θ_*, d, G)-dependent constants $C, c, \sigma_0 > 0$, if $\sigma < \sigma_0$, then with probability at least $1 - Ce^{-cn^{2/3}} - \sigma^{-C}e^{-c\sigma n}$, $\lambda_{\min}(\nabla^2 R_n(\theta)) \geq 1/(2\sigma^2)$ for all $\theta \in B_\rho(\theta_*)$, and $R_n(\theta)$ has a unique local minimizer and critical point in $B_\rho(\theta_*)$.*

Proof. This follows from Lemma 2.7 and Theorem 3.1 if we can show that

$$\sup_{\theta \in B_\rho(\theta_*)} \|R_n(\theta) - R(\theta)\| \leq c_1/\sigma^2 \qquad \text{and} \qquad \sup_{\theta \in B_\rho(\theta_*)} \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \leq c_1/\sigma^2$$

for a small enough constant $c_1 > 0$. Applying (2.16) with $r = \|\theta_*\| + \rho$ and $t = c_1/\sigma^2$, we obtain $\sup_{\theta \in B_\rho(\theta_*)} \|R_n(\theta) - R(\theta)\| \leq c_1/\sigma^2$ with probability $1 - Ce^{-cn}$. Applying (2.18), we also obtain $\sup_{\theta \in B_\rho(\theta_*)} \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \leq c_1/\sigma^2$ with probability $1 - \sigma^{-C}e^{-c\sigma^4 n} - Ce^{-cn^{2/3}}$. To reduce σ^4 to σ in this probability bound, let us derive a sharper concentration inequality for $\nabla^2 R_n(\theta)$ than the general result provided by (2.21), when $\theta \in B_\rho(\theta_*)$ and $\sigma < \sigma_0$.

Recall the set $\mathcal{E}$ in (3.2) and the form for $\nabla^2 R_n(\theta)$ in (2.5). Let us write this as

$$\nabla^2 R_n(\theta) = \frac{1}{\sigma^2}\,\mathrm{Id} - \frac{1}{\sigma^4} \cdot \frac{1}{n}\sum_{i=1}^n (X_i + Y_i) - \frac{1}{\sigma^2} \cdot \frac{1}{n}\sum_{i=1}^n Z_i \tag{3.4}$$

where $X_i, Y_i, Z_i \in \mathbb{R}^{d\times d}$ are given by

$$\begin{aligned}
X_i &= \left(\mathrm{Cov}_g\left[g^\top(\theta_* + \sigma\varepsilon_i)\middle|\varepsilon_i, \theta\right] - \sigma^2\,\mathrm{Cov}_g\left[g^\top\varepsilon_i\middle|\varepsilon_i, \theta\right]\right)\mathbf{1}\{\varepsilon_i \in \mathcal{E}\} \\
Y_i &= \left(\mathrm{Cov}_g\left[g^\top(\theta_* + \sigma\varepsilon_i)\middle|\varepsilon_i, \theta\right] - \sigma^2\,\mathrm{Cov}_g\left[g^\top\varepsilon_i\middle|\varepsilon_i, \theta\right]\right)\mathbf{1}\{\varepsilon_i \notin \mathcal{E}\} \\
Z_i &= \mathrm{Cov}_g\left[g^\top\varepsilon_i\middle|\varepsilon_i, \theta\right].
\end{aligned}$$

Observe that since $\|Z_i\| \leq \|\mathbb{E}_g[g^\top\varepsilon_i\varepsilon_i^\top g \mid \varepsilon_i, \theta]\| \leq \|\varepsilon_i\|^2$, and $\|\varepsilon_i\|^2$ has constant sub-exponential norm, each entry of Z_i also has constant sub-exponential norm (where constants may depend on d). Applying Bernstein's inequality entrywise and taking a union bound over all entries, for constants $C, c > 0$ and any $t > 0$,

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\right\| \geq t\right] \leq Ce^{-cn\min(t,t^2)}. \tag{3.5}$$

For X_i, note that $p(\mathrm{Id} \mid \varepsilon, \theta) > 1 - e^{-c/\sigma^2}$ when $\varepsilon \in \mathcal{E}$, as shown in (3.3). Then for any unit vector $v \in \mathbb{R}^d$,

$$\begin{aligned}|v^\top X_i v| &= \left|\mathrm{Var}_g\left[\langle gv, \theta_* + \sigma\varepsilon_i\rangle \middle| \varepsilon_i, \theta\right] - \sigma^2 \mathrm{Var}_g\left[\langle gv, \varepsilon_i\rangle \middle| \varepsilon_i, \theta\right]\right| \mathbf{1}\{\varepsilon_i \in \mathcal{E}\} \\ &\le \left(\mathbb{E}_g\left[\langle gv - v, \theta_* + \sigma\varepsilon_i\rangle^2 \middle| \varepsilon_i, \theta\right] + \sigma^2 \mathbb{E}_g\left[\langle gv - v, \varepsilon_i\rangle^2 \middle| \varepsilon_i, \theta\right]\right) \mathbf{1}\{\varepsilon_i \in \mathcal{E}\} \\ &\le \mathbb{P}[g \ne \mathrm{Id} \mid \varepsilon_i, \theta]\left(4\|\theta_* + \sigma\varepsilon_i\|^2 + 4\sigma^2\|\varepsilon_i\|^2\right)\mathbf{1}\{\varepsilon_i \in \mathcal{E}\} \le Ce^{-c/\sigma^2}.\end{aligned}$$

Thus $\|X_i\| \le Ce^{-c/\sigma^2}$ for each $i = 1, \dots, n$. Applying Hoeffding's inequality entrywise to X_i and taking a union bound over all entries,

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i]\right\| \ge t\sigma^2\right] \le C\exp\left(-ne^{c'/\sigma^2}t^2\right). \tag{3.6}$$

For Y_i, let us fix indices $j, k \in \{1, \dots, d\}$ and consider $\sum_i (Y_i)_{jk}$. Let $W_1, \dots, W_m$ be i.i.d. random variables whose law is that of $(Y_i)_{jk}$ conditional on $\varepsilon_i \notin \mathcal{E}$. We apply Hoeffding's inequality for $W_1, \dots, W_m$: Observe that since the two quadratic terms in ε_i cancel in the definition of Y_i, we have $|(Y_i)_{jk}| \le C(1 + \sigma\|\varepsilon_i\|)$ for a constant $C = C(\|\theta_*\|) > 0$. Then

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{W_i^2}{t^2}\right)\right] &\le \mathbb{E}_\varepsilon\left[\exp\left(\frac{C(1+\sigma^2\|\varepsilon\|^2)}{t^2}\right)\middle|\varepsilon \notin \mathcal{E}\right] \\ &= e^{C/t^2} \cdot \mathbb{E}_\varepsilon\left[\exp\left(\frac{C\sigma^2\|\varepsilon\|^2}{t^2}\right)\middle|\|\varepsilon\|^2 > \left(\frac{c_0}{2\sigma\|\theta\|}\right)^2\right].\end{aligned}$$

Specializing [CM00, Eq. (2.9)] to the chi-squared distribution, we obtain

$$\mathbb{E}\left[\exp(s\|\varepsilon\|^2) \mid \|\varepsilon\|^2 > x\right] = \frac{\mathbb{P}[\|\varepsilon\|^2 > x(1-2s)]}{\mathbb{P}[\|\varepsilon\|^2 > x]}(1-2s)^{-d/2}$$

for $s < 1/2$. Here $\mathbb{P}[\|\varepsilon\|^2 > x] = \Gamma(d/2, x/2)/\Gamma(d/2)$ where $\Gamma(a, y)$ is the upper-incomplete Gamma function which satisfies $\Gamma(a, y)/y^{a-1}e^{-y} \to 1$ as $y \to \infty$, for fixed a (see [AS48, Eq. (6.5.32)]). Then

$$\frac{\mathbb{P}[\|\varepsilon\|^2 > x(1-2s)]}{\mathbb{P}[\|\varepsilon\|^2 > x]} \cdot (1-2s)^{-d/2+1}e^{-xs} \to 1$$

as $x \to \infty$, uniformly over $s \in (0, 1/2)$. Setting $x = c_0^2/(2\sigma\|\theta\|)^2$ and $t = C_1$ for a large enough constant $C_1 > 0$, we obtain that $C/t^2 < 0.05$, $s \equiv C\sigma^2/t^2 < 0.05/x$, and hence $\mathbb{E}[\exp(W_i^2/t^2)] \le 2$ when $\sigma < \sigma_0$ for small enough $\sigma_0 > 0$. Thus $\|W_i\|_{\psi_2} \le C_1$, and Hoeffding's inequality yields, for a constant $c > 0$ and any $s \ge 0$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^m W_i - \mathbb{E}[W_i]\right| \ge s\right] \le 2e^{-cms^2}.$$

Returning to $(Y_i)_{jk}$, let $S = \{i \in [n] : \varepsilon_i \notin \mathcal{E}\}$. The above shows that, conditional on S,

$$\mathbb{P}\left[\left|\frac{1}{|S|}\sum_{i\in S}(Y_i)_{jk}\right| \ge s + |\mathbb{E}W_i| \;\middle|\; S\right] \le 2e^{-c|S|s^2}.$$

Noting that $(Y_i)_{jk} = 0$ when $i \notin S$, this implies

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n (Y_i)_{jk} - \mathbb{E}[(Y_i)_{jk}]\right| \ge \left(s + |\mathbb{E}W_i|\right)\frac{|S|}{n} + \left|\mathbb{E}(Y_i)_{jk}\right| \;\middle|\; S\right] \le 2e^{-c|S|s^2}.$$

We have $\mathbb{P}[\varepsilon_i \notin \mathcal{E}] \le e^{-c/\sigma^2}$, by a chi-squared tail bound. From the bound $\|W_i\|_{\psi_2} \le C_1$, we have $|\mathbb{E}W_i| \le C$. Then also $\mathbb{E}(Y_i)_{jk} = (\mathbb{E}W_i) \cdot \mathbb{P}[\varepsilon_i \notin \varepsilon] \le Ce^{-c/\sigma^2}$. Setting $t\sigma^2 = s|S|/n$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n (Y_i)_{jk} - \mathbb{E}[(Y_i)_{jk}]\right| \ge t\sigma^2 + Ce^{-c'/\sigma^2} \;\middle|\; S\right] \le 2e^{-cn^2\sigma^4 t^2/|S|}$$

for some constants $C, c, c' > 0$. On the event $|S| \le n\sigma^3$, we obtain the bound $2e^{-cn\sigma t^2}$. By a Chernoff bound, $\mathbb{P}[|S| > n\sigma^3] \le \exp(-n\, D_{\mathrm{KL}}(\sigma^3 || e^{-c/\sigma^2}))$ for the Bernoulli relative entropy

$$D_{\mathrm{KL}}(\sigma^3||e^{-c/\sigma^2}) = \sigma^3 \log\frac{\sigma^3}{e^{-c/\sigma^2}} + (1-\sigma^3)\log\frac{1-\sigma^3}{1-e^{-c/\sigma^2}} \ge c'\sigma.$$

Combining these, we obtain unconditionally that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n (Y_i)_{jk} - \mathbb{E}[(Y_i)_{jk}]\right| \ge t\sigma^2 + Ce^{-c'/\sigma^2}\right] \le Ce^{-cn\sigma t^2}. \tag{3.7}$$

Picking a sufficiently small constant t in (3.5), (3.6), and (3.7) and applying this to (3.4), we obtain $\|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \le c_1/(2\sigma^2)$ with probability at least $1 - Ce^{-c\sigma n}$. This is a pointwise bound for each $\theta \in B_\rho(\theta_*)$. Taking a union bound over a δ-net of this ball for $\delta = c\sigma^4$, and applying the Lipschitz continuity of $\nabla^2 R(\theta)$ and $\nabla^2 R_n(\theta)$ from Lemma 2.12, we get the uniform bound $\sup_{\theta\in B_\rho(\theta_*)} \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \le c_1/\sigma^2$ with probability $1 - Ce^{-cn^{2/3}} - \sigma^{-C}e^{-c\sigma n}$ as desired. □

3.2. **Global landscape.**

Theorem 3.3. *Let $\theta_* \in \mathbb{R}^d$ be such that $|\mathcal{O}_{\theta_*}| = |G| = K$. There exists a (θ_*, d, G)-dependent constant $\sigma_0 > 0$ such that as long as $\sigma < \sigma_0$, the landscape of $R(\theta)$ is globally benign.*

More quantitatively, let ρ be as in Theorem 3.1. Then there is a (θ_*, d, G)-dependent constant $c > 0$ and a decomposition $\mathbb{R}^d \setminus \bigcup_{\mu\in\mathcal{O}_{\theta_*}} B_\rho(\mu) \equiv \mathcal{A} \sqcup \mathcal{B}$, where for $\theta \in \mathcal{A}$

$$\lambda_{\min}(\nabla^2 R(\theta)) < -c/\sigma^3, \tag{3.8}$$

and for $\theta \in \mathcal{B}$

$$\|\nabla R(\theta)\| > c/\sigma^2. \tag{3.9}$$

Let us provide some intuition for the proof: Recall the reweighted law (2.3) for $g \in G$. We enumerate

$$G = \{g_1, \dots, g_K\},$$

fix a small constant $\tau > 0$, and divide the space of $\varepsilon \in \mathbb{R}^d$ into the regions

$$\mathcal{E}_i(\theta,\tau) = \Big\{\varepsilon \in \mathbb{R}^d : p(g_k \mid \varepsilon, \theta) \le \tau \text{ for all } k \in \{1,\dots,K\} \setminus \{i\}\Big\}, \tag{3.10}$$

$$\mathcal{E}_{ij}(\theta,\tau) = \Big\{\varepsilon \in \mathbb{R}^d : p(g_i \mid \varepsilon, \theta) > \tau \text{ and } p(g_j \mid \varepsilon, \theta) > \tau\Big\}. \tag{3.11}$$

Here, for τ small enough, $\mathcal{E}_i(\theta,\tau)$ is the space of noise vectors ε for which the ε-dependent distribution (2.3) places nearly all of its weight on g_i, and $\mathcal{E}_{ij}(\theta,\tau)$ is the space of ε for which this distribution "straddles" its weight between at least two points $g_i \ne g_j \in G$.

We will choose the set $\mathcal{B}$ in Theorem 3.3 to be those vectors $\theta \in \mathbb{R}^d$ for which $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] \approx 1$ for some $i \in \{1,\dots,K\}$. Thus, for some fixed $g_i \in G$, with high probability over ε, the law (2.3) places nearly all of its weight on the single element g_i. Intuitively, from the form (2.3), these are the points $\theta \in \mathbb{R}^d$ which are closer to $g_i^\top\theta_*$ than to the other points $g_j^\top\theta_*$ for $j \ne i$.

The remaining points $\mathbb{R}^d \setminus \mathcal{B}$ will constitute $\mathcal{A}$. A key step of the proof is to show that if $\theta \notin \mathcal{B}$, then there must be a pair $i \ne j$ for which $\mathbb{P}[\varepsilon \in \mathcal{E}_{ij}(\theta,\tau)] \gtrsim \sigma$. That is, with some small probability of order σ, the law (2.3) straddles its weight between g_i and g_j. (Note that this is not

tautological from the definitions, as we must rule out the possibility, e.g., that $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] = 1/2$ and $\mathbb{P}[\varepsilon \in \mathcal{E}_j(\theta,\tau)] = 1/2$ for some $i \neq j$, but $\mathbb{P}[\varepsilon \in \mathcal{E}_{ij}(\theta,\tau)] = 0$. Indeed, from the form of (2.3), we see that even if θ is exactly equidistant from $g_i^\top \theta_*$ and $g_j^\top \theta_*$, the probability over ε is only $O(\sigma)$ that $p(g_i \mid \varepsilon, \theta)$ and $p(g_j \mid \varepsilon, \theta)$ are comparable.) We prove this claim using a Gaussian isoperimetric argument in Lemma 3.4 below.

Lemma 3.4. *Fix any $\theta \neq 0$ and $\tau \in (0, (K+9)^{-1})$, and define $\mathcal{E}_i, \mathcal{E}_{ij}$ by (3.10) and (3.11). Suppose, for some $i \in \{1,\dots,K\}$ and $p \in (0, 1/2]$, that*

$$p \le \mathbb{P}[\varepsilon \in \mathcal{E}_i] \le 1/2.$$

Then for some $j \in \{1,\dots,K\} \setminus \{i\}$,

$$\mathbb{P}[\varepsilon \in \mathcal{E}_{ij}] \ge \frac{p}{(K-1)\sqrt{2\pi}} \min\left(\frac{\sigma}{\|\theta\|}, 1\right).$$

Proof. Let $\mathcal{E}_i^t = \{\varepsilon \in \mathbb{R}^d : \mathrm{dist}(\varepsilon, \mathcal{E}_i) < t\}$. We first claim that if $\varepsilon \in \mathcal{E}_i^t \setminus \mathcal{E}_i$ for $t = \sigma/\|\theta\|$, then there exists some $j \neq i$ for which $\varepsilon \in \mathcal{E}_{ij}$. For this, note that

$$\nabla_\varepsilon [\log p(g \mid \varepsilon, \theta)] = \frac{1}{\sigma}\Big(g\theta - \mathbb{E}_h[h\theta \mid \varepsilon, \theta]\Big),$$

so $\varepsilon \mapsto \log p(g_i \mid \varepsilon, \theta)$ has the Lipschitz bound $\|\nabla_\varepsilon \log p(g_i \mid \varepsilon, \theta)\| \le 2\|\theta\|/\sigma$. Suppose that $\varepsilon \in \mathcal{E}_i^t \setminus \mathcal{E}_i$. Then there is $\varepsilon' \in \mathcal{E}_i$ with $\|\varepsilon - \varepsilon'\| < \sigma/\|\theta\|$, so $\log p(g_i \mid \varepsilon', \theta) - \log p(g_i \mid \varepsilon, \theta) \le 2$ and

$$p(g_i \mid \varepsilon', \theta)/p(g_i \mid \varepsilon, \theta) \le e^2 < 8.$$

Since $p(g_1 \mid \varepsilon', \theta) + \ldots + p(g_K \mid \varepsilon', \theta) = 1$ and $(K+9)\tau < 1$, when $\varepsilon' \in \mathcal{E}_i$ we must have $p(g_i \mid \varepsilon', \theta) \ge 1 - (K-1)\tau > 8\tau$. Then the above implies $p(g_i \mid \varepsilon, \theta) > \tau$. Since $\varepsilon \notin \mathcal{E}_i$, by definition of $\mathcal{E}_i$ we must also have $p(g_j \mid \varepsilon, \theta) > \tau$ for some $j \neq i$, so that $\varepsilon \in \mathcal{E}_{ij}$ as desired. Note that this index $j \in \{1,\dots,K\}$ may depend on ε. However, this shows that for at least one *fixed* index $j \in \{1,\dots,K\} \setminus \{i\}$,

$$\mathbb{P}[\varepsilon \in \mathcal{E}_{ij}] \ge \frac{\mathbb{P}[\varepsilon \in \mathcal{E}_i^t \setminus \mathcal{E}_i]}{K-1}. \tag{3.12}$$

We now apply the Gaussian isoperimetric inequality to lower bound the right side: For Φ the standard normal distribution function,

$$\Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i^t]) \ge \Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i]) + t,$$

see [BLM13, Theorem 10.15]. Then, denoting by ϕ the standard normal density,

$$\begin{aligned}\mathbb{P}[\varepsilon \in \mathcal{E}_i^t \setminus \mathcal{E}_i] = \mathbb{P}[\varepsilon \in \mathcal{E}_i^t] - \mathbb{P}[\varepsilon \in \mathcal{E}_i] &\ge \Phi(\Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i]) + t) - \Phi(\Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i])) \\ &= \int_{\Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i])}^{\Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i])+t} \phi(r) dr.\end{aligned}$$

Applying $\mathbb{P}[\varepsilon \in \mathcal{E}_i] \in [p, 1/2]$ by assumption, we get $\Phi^{-1}(\mathbb{P}[\varepsilon \in \mathcal{E}_i]) \in [\Phi^{-1}(p), 0]$. Then there is always an interval of values for r, having length $\min(t,1)$ and contained in the above range of integration, for which $\phi(r) \ge \min(\phi(\Phi^{-1}(p)), \phi(1))$ over this interval. Applying the tail bound $\Phi(x) \le e^{-x^2/2}$ for all $x \le 0$, we get $\Phi^{-1}(p) \ge -\sqrt{2\log 1/p}$ and $\phi(\Phi^{-1}(p)) \ge p/\sqrt{2\pi}$. For $p \le 1/2$ we have $p/\sqrt{2\pi} < \phi(1)$. Combining these observations gives

$$\mathbb{P}[\varepsilon \in \mathcal{E}_i^t \setminus \mathcal{E}_i] \ge \min(t,1) \cdot \frac{p}{\sqrt{2\pi}}.$$

Recalling $t = \sigma/\|\theta\|$ and combining with (3.12) yields the lemma. ∎

Proof of Theorem 3.3. Let us fix two positive constants

$$\tau < \min\left(\frac{1}{K+9},\ \frac{\rho}{8\|\theta_*\|K}\right) \tag{3.13}$$

and

$$p < \left(\frac{\rho}{12\|\theta_*\|}\right)^2 \Big/ K. \tag{3.14}$$

Define $\mathcal{E}_i(\theta,\tau)$ and $\mathcal{E}_{ij}(\theta,\tau)$ by (3.10) and (3.11) with this choice of τ, and set

$$\mathcal{A} = \left\{\theta \in \mathbb{R}^d \setminus \mathcal{C} : \mathbb{P}[\varepsilon \in \mathcal{E}_{ij}(\theta,\tau)] > \frac{p}{K\sqrt{2\pi}} \cdot \frac{\sigma}{3\|\theta_*\|} \text{ for some } i \neq j\right\},$$
$$\mathcal{B} = \left\{\theta \in \mathbb{R}^d \setminus \mathcal{C} : \mathbb{P}[\varepsilon \in \mathcal{E}_{ij}(\theta,\tau)] \leq \frac{p}{K\sqrt{2\pi}} \cdot \frac{\sigma}{3\|\theta_*\|} \text{ for all } i \neq j\right\}.$$

To check (3.8) when $\theta \in \mathcal{A}$, recall the form of $\nabla^2 R(\theta)$ in (2.9). We apply $\mathbb{P}[\varepsilon \in \mathcal{E}_{ij}(\theta,\tau)] > c\sigma$ for a constant $c > 0$ and some $i \neq j$, by the definition of $\mathcal{A}$. Choose a constant $c_0 > 0$ such that $\|g_i^\top \theta_* - g_j^\top \theta_*\| > 3c_0$. Then a chi-squared tail bound yields

$$\mathbb{P}\Big[\|\varepsilon\| \leq c_0/\sigma \text{ and } \varepsilon \in \mathcal{E}_{ij}(\theta,\tau)\Big] > c'\sigma \tag{3.15}$$

for a different constant $c' < c$ and all $\sigma < \sigma_0$. For ε satisfying (3.15), we have

$$\|g_i^\top(\theta_* + \sigma\varepsilon) - g_j^\top(\theta_* + \sigma\varepsilon)\| \geq \|g_i^\top \theta_* - g_j^\top \theta_*\| - 2\sigma\|\varepsilon\| \geq c_0,$$

and also $p(g_i \mid \varepsilon, \theta) > \tau$ and $p(g_j \mid \varepsilon, \theta) > \tau$. Then for such ε, denoting $\mu = \mathbb{E}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta]$, we have

$$\begin{aligned}\operatorname{Tr}\operatorname{Cov}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta] &= \mathbb{E}_g[\|g^\top(\theta_* + \sigma\varepsilon) - \mu\|^2 \mid \varepsilon, \theta] \\ &\geq \tau \cdot \|g_i^\top(\theta_* + \sigma\varepsilon) - \mu\|^2 + \tau \cdot \|g_j^\top(\theta_* + \sigma\varepsilon) - \mu\|^2 > c.\end{aligned}$$

Combining this with (3.15) implies that

$$\lambda_{\max}\left(\mathbb{E}_\varepsilon\Big[\operatorname{Cov}_g\Big[g^\top(\theta_* + \sigma\varepsilon)\ \Big|\ \varepsilon, \theta\Big]\Big]\right) > c\sigma.$$

Then (3.8) follows from (2.9).

To check (3.9) when $\theta \in \mathcal{B}$, note that if $\|\theta\| \geq 3\|\theta_*\|$, then (3.9) follows from Lemma 2.8. For $\theta \in \mathcal{B}$ such that $\|\theta\| < 3\|\theta_*\|$, the definition of $\mathcal{B}$ and Lemma 3.4 imply that either $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] < p$ or $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] > 1/2$ for every $i \in \{1,\dots,K\}$. Note that since $K\tau < 1$, we must have:

- $\mathcal{E}_1(\theta,\tau), \dots, \mathcal{E}_K(\theta,\tau)$ are disjoint.
- $\{\mathcal{E}_i(\theta,\tau)\}_{i=1}^K$ and $\{\mathcal{E}_{ij}(\theta,\tau)\}_{i\neq j}$ together cover all of $\mathbb{R}^d$.

The first observation implies that $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] > 1/2$ for at most one index $i \in \{1,\dots,K\}$, so we must have $\mathbb{P}[\varepsilon \in \mathcal{E}_j(\theta,\tau)] < p$ for all other $j \neq i$. Combining this with the second observation,

$$1 \leq \mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] + \sum_{j:j\neq i} \mathbb{P}[\varepsilon \in \mathcal{E}_j(\theta,\tau)] + \sum_{j\neq k} \mathbb{P}[\varepsilon \in \mathcal{E}_{jk}(\theta,\tau)] \leq \mathbb{P}[\varepsilon \in \mathcal{E}_i] + (K-1)p + \binom{K}{2} c\sigma.$$

For $\sigma < \sigma_0$ and sufficiently small σ_0, this implies $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta,\tau)] \geq 1 - Kp$.

Recall the form (2.7) for $\nabla R(\theta)$. For this index i, let us write

$$\mathbb{E}_\varepsilon\Big[\mathbb{E}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta]\Big] - g_i^\top \theta_* = \mathrm{I} + \mathrm{II} + \mathrm{III}$$

where

$$\mathrm{I} = \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\varepsilon \notin \mathcal{E}_i\}\Big(\mathbb{E}_g[g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta] - g_i^\top \theta_*\Big)\Big],$$
$$\mathrm{II} = \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\varepsilon \in \mathcal{E}_i\}\Big(\mathbb{E}_g[\mathbf{1}\{g \neq g_i\} g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta]\Big)\Big],$$

$$\mathrm{III} = \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\varepsilon \in \mathcal{E}_i\}\Big(\mathbb{E}_g[\mathbf{1}\{g = g_i\}g^\top(\theta_* + \sigma\varepsilon) \mid \varepsilon, \theta] - g_i^\top\theta_*\Big)\Big].$$

Applying Cauchy-Schwarz, the above bound $\mathbb{P}[\varepsilon \in \mathcal{E}_i(\theta, \tau)] \geq 1 - Kp$, and the condition (3.14) for p, we get for $\sigma < \sigma_0$ and small enough σ_0 that

$$\|\mathrm{I}\| \leq \mathbb{P}[\varepsilon \notin \mathcal{E}_i]^{1/2}\mathbb{E}_\varepsilon[(\|\theta_* + \sigma\varepsilon\| + \|\theta_*\|)^2]^{1/2} \leq (Kp)^{1/2} \cdot 3\|\theta_*\| < \rho/4.$$

When $\varepsilon \in \mathcal{E}_i$, we have $\mathbb{P}_g[g = g_i \mid \varepsilon, \theta] = p(g_i \mid \varepsilon, \theta) > 1 - K\tau$. Then by the condition (3.13) for τ, for $\sigma < \sigma_0$,

$$\|\mathrm{II}\| \leq \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\varepsilon \in \mathcal{E}_i\}\mathbb{P}_g[g \neq g_i \mid \varepsilon, \theta] \cdot \|\theta_* + \sigma\varepsilon\|\Big] \leq K\tau \cdot 2\|\theta_*\| < \rho/4.$$

For III, we cancel $g_i^\top\theta_*$ to get the bound

$$\|\mathrm{III}\| \leq \mathbb{E}_\varepsilon\Big[\mathbb{E}_g[\mathbf{1}\{g = g_i\}\|g^\top(\sigma\varepsilon)\| \mid \varepsilon, \theta]\Big] \leq \sigma\,\mathbb{E}_\varepsilon[\|\varepsilon\|] < \rho/4.$$

Combining these with (2.7) yields

$$\|\sigma^2\nabla R(\theta) - (\theta - g_i^\top\theta_*)\| < 3\rho/4,$$

and (3.9) follows since $\|\theta - g_i^\top\theta_*\| \geq \rho$ because $\theta \notin \bigcup_{\mu\in\mathcal{O}_{\theta_*}} B_\rho(\mu)$. These conditions (3.8), (3.9), and Theorem 3.1 together show that the landscape of $R(\theta)$ is globally benign. □

The following then shows that the landscape of $R_n(\theta)$ is also globally benign with high probability, when $n \gg \sigma^{-2}\log\sigma^{-1}$.

Corollary 3.5. *In the setting of Theorem 3.3, the same statements hold for the empirical risk $R_n(\theta)$ with probability at least $1 - \sigma^{-C}e^{-c\sigma^2 n} - Ce^{-cn^{2/3}}$.*

Proof. For $\sigma < \sigma_0$ and small enough σ_0, with probability $1 - Ce^{-cn}$, we have $\|\nabla R_n(\theta)\| \geq c/\sigma^2$ for all θ such that $\|\theta\| > 3\|\theta_*\|$ by Lemma 2.8. Applying the concentration result (2.17) with $t = c_0/\sigma^2$, and (2.18) with $t = c_0/\sigma^3$, over the ball B_r for $r = 3\|\theta_*\|$, for small enough c_0 we obtain (3.8) and (3.9) also for the empirical risk $R_n(\theta)$, with probability $1 - \sigma^{-C}e^{-c\sigma^2 n} - Ce^{-cn^{2/3}}$. The result then follows from combining with Corollary 3.2. □

4. Landscape analysis for high noise

In this section, we now analyze the function landscapes of $R(\theta)$ and $R_n(\theta)$ in the high-noise regime $\sigma > \sigma_0(\theta_*, d, G)$. Our results relate to the algebra of G-invariant polynomials and systems of reparametrized coordinates in local neighborhoods, which we first review in Section 4.1.

Our analysis for high noise is based on a large-σ series expansion of the population risk,

$$R(\theta) \text{ “} = \text{” } \sum_{\ell=1}^{\infty} \sigma^{-2\ell} S_\ell(\theta). \tag{4.1}$$

We derive this in Section 4.2 by using the series expansion of the cumulant generating function $\log \mathbb{E}_g \exp(\langle\theta_* + \sigma\varepsilon, g\theta\rangle/\sigma^2)$ in (2.2). We write “ = ” because we do not show convergence of this series for any finite value of σ, but rather quantify the accuracy of the approximation to $R(\theta)$ given by its first k terms, for any fixed k as $\sigma \to \infty$.

The functions $S_\ell(\theta)$ in (4.1) do not depend on σ, and we analyze the form of these terms also in Section 4.2. We show in Section 4.3 that the local landscape of $R(\theta)$ around any point $\widetilde{\theta} \in \mathbb{R}^d$ may be understood, for large σ, by analyzing the successive landscapes of these functions $S_\ell(\theta)$ in a reparametrized system of coordinates near $\widetilde{\theta}$.

In Section 4.4, we apply this at $\widetilde{\theta} = \theta_*$ to analyze the local landscape near θ_*. Theorem 4.14 and Corollary 4.16 show that $R(\theta)$ and $R_n(\theta)$ are strongly convex in a σ-independent neighborhood of θ_*, when reparametrized by a transcendence basis of the G-invariant polynomial algebra. Theorem

4.14 also shows that $I(\theta_*)$ has a certain graded structure, where the magnitudes of its eigenvalues correspond to a sequence of transcendence degrees in this algebra.

In Section 4.5, we patch together the local results of Section 4.3 to study the global landscapes of $R(\theta)$ and $R_n(\theta)$. Theorems 4.18, 4.21 and Corollaries 4.20, 4.24 establish globally benign landscapes for K-fold discrete rotations on $\mathbb{R}^2$ and the symmetric group of all permutations on $\mathbb{R}^d$, for large σ and large n. Theorem 4.25 then generalizes this to a more abstract condition, in terms of minimizing the sequence of polynomials $P_\ell(\theta)$ in (1.11) over the sequence of moment varieties $\mathcal{V}_{\ell-1}$ in (1.12).

Finally, in Section 4.6, we analyze the global landscape for cyclic permutations on $\mathbb{R}^d$ (i.e. multi-reference alignment). Theorem 4.26 and Corollary 4.29 show that the local minimizers of $R(\theta)$ and $R_n(\theta)$ are in correspondence with those of a minimization problem in phase space. Corollary 4.27 shows that their landscapes are benign in dimensions $d \le 5$ (for large σ and large n), but may not be benign even for generic θ_* when the dimension reaches $d = 6$.

4.1. **Invariant polynomials and local reparametrization.**

Definition 4.1. For a subgroup $G \subseteq \mathrm{O}(d)$, a polynomial function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is **G-invariant** if $\varphi(g\theta) = \varphi(\theta)$ for all $g \in G$. We denote by $\mathcal{R}^G$ the algebra (over $\mathbb{R}$) of all G-invariant polynomials on $\mathbb{R}^d$, and by $\mathcal{R}^G_{\le \ell} \subset \mathcal{R}^G$ the vector space of such polynomials having degree $\le \ell$.

Definition 4.2. Polynomials $\varphi_1, \dots, \varphi_k : \mathbb{R}^d \to \mathbb{R}$ are **algebraically independent** (over $\mathbb{R}$) if there is no non-zero polynomial $P : \mathbb{R}^k \to \mathbb{R}$ for which $P(\varphi_1(\theta), \dots, \varphi_k(\theta))$ is identically 0 over $\theta \in \mathbb{R}^d$. For a subset $A \subseteq \mathcal{R}^G$, its **transcendence degree** $\mathrm{trdeg}(A)$ is the maximum number of algebraically independent elements in A.

One may construct a transcendence basis of d such polynomials according to the following lemma; we provide a proof for convenience in Appendix A.2.

Lemma 4.3. *For any finite subgroup $G \subset \mathrm{O}(d)$, there exists a smallest integer $L \ge 1$ for which $\mathrm{trdeg}(\mathcal{R}_{\le L}) = d$. Writing $d = d_1 + \ldots + d_L$ where*

$$d_\ell = \mathrm{trdeg}(\mathcal{R}^G_{\le \ell}) - \mathrm{trdeg}(\mathcal{R}^G_{\le \ell-1}),$$

there also exist d algebraically independent G-invariant polynomials $\varphi = (\varphi^1, \dots, \varphi^L)$, where each subvector φ^ℓ consists of d_ℓ polynomials having degree exactly ℓ.

It was shown in [BBSK+17] that this number L is the highest-order moment needed for a moment-of-moments estimator to recover a generic signal θ_* in the model (1.1), up to a finite list of possibilities including (but not necessarily limited to) the orbit points $\mathcal{O}_{\theta_*}$, and that the number of samples required for this type of recovery scales as $O(\sigma^{2L})$.

In our local analysis around a point $\widetilde{\theta} \in \mathbb{R}^d$, we will switch to a system of reparametrized coordinates. Let us specify our notation for such a reparametrization.

Definition 4.4. A function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is a **local reparametrization** in an open neighborhood U of $\widetilde{\theta} \in \mathbb{R}^d$ if φ is 1-to-1 on U with inverse function $\theta(\varphi)$, and $\varphi(\theta)$ and $\theta(\varphi)$ are analytic respectively on U and $\varphi(U)$.

If φ is a local reparametrization, then $\mathsf{d}_\theta \varphi$ is non-singular and equal to $(\mathsf{d}_\varphi \theta)^{-1}$ at each $\theta \in U$. Conversely, by the inverse function theorem, if $\varphi(\theta)$ is analytic and $\mathsf{d}_\theta \varphi(\widetilde{\theta})$ is non-singular, then there is such an open neighborhood U of $\widetilde{\theta}$ on which φ defines a local reparametrization.

To ease notation, we write (with a slight abuse) $f(\varphi)$ for $f(\theta(\varphi))$ when the meaning is clear, and we write $\nabla_\varphi f(\varphi)$, $\nabla^2_\varphi f(\varphi)$, and $\partial_{\varphi_i} f(\varphi)$ for the gradient, Hessian, and partial derivatives of $f(\varphi)$ with respect to φ. For a decomposition $\varphi = (\varphi^1, \dots, \varphi^L)$ of dimensions $d_1, \dots, d_L$, we denote by $\nabla_{\varphi^\ell} f(\varphi) \in \mathbb{R}^{d_\ell}$ and $\nabla^2_{\varphi^\ell} f(\varphi) \in \mathbb{R}^{d_\ell \times d_\ell}$ the subvectors and submatrices of $\nabla_\varphi f(\varphi)$ and $\nabla^2_\varphi f(\varphi)$ corresponding to the coordinates in φ^ℓ.

Recalling $\nabla_\theta f(\theta) = \mathsf{d}_\theta f(\theta)^\top$, by the chain rule and product rule, we have

$$\nabla_\theta f(\theta) = (\mathsf{d}_\theta \varphi)^\top \nabla_\varphi f(\varphi) \tag{4.2}$$

$$\nabla^2_\theta f(\theta) = (\mathsf{d}_\theta \varphi)^\top \cdot \nabla^2_\varphi f(\varphi) \cdot \mathsf{d}_\theta \varphi + \sum_{i=1}^{d} \partial_{\varphi_i} f(\varphi) \cdot \nabla^2_\theta \varphi_i \tag{4.3}$$

Note that $\nabla_\theta f(\widetilde{\theta}) = 0$ if and only if $\nabla_\varphi f(\widetilde{\varphi}) = 0$ for $\widetilde{\varphi} = \varphi(\widetilde{\theta})$, i.e. critical points do not depend on the choice of parametrization. At a critical point $\widetilde{\theta}$ of $f(\theta)$, letting $\widetilde{\varphi} = \varphi(\widetilde{\theta})$, the identity (4.3) simplies to just the first term,

$$\nabla^2_\theta f(\widetilde{\theta}) = (\mathsf{d}_\theta \varphi(\widetilde{\theta}))^\top \cdot \nabla^2_\varphi f(\widetilde{\varphi}) \cdot \mathsf{d}_\theta \varphi(\widetilde{\theta}),$$

so that the rank and signs of the eigenvalues of $\nabla^2_\theta f(\widetilde{\theta})$ also do not depend on the choice of parametrization. This may be false when θ is not a critical point—in particular, strong convexity of $f(\varphi)$ as a function of $\varphi \in \varphi(U)$ does not imply strong convexity of $f(\theta)$ as a function of $\theta \in U$.

For analyzing specific groups, we will explicitly describe our reparametrization φ. For more general results, we will reparametrize by the transcendence basis of polynomials φ in Lemma 4.3. The following clarifies the relationship between algebraic independence of these polynomials and linear independence of their gradients, and implies in particular that φ is a local reparametrization at generic points of $\mathbb{R}^d$. We provide a proof also in Appendix A.2.

Lemma 4.5. *Let $G \subset \mathrm{O}(d)$ be a finite subgroup, and let $\varphi_1, \dots, \varphi_k$ be polynomials in $\mathcal{R}^G$.*

(a) If $\varphi_1, \dots, \varphi_k$ are algebraically independent, then $\nabla\varphi_1, \dots, \nabla\varphi_k$ are linearly independent at generic points $\theta \in \mathbb{R}^d$.

(b) If $\nabla\varphi_1, \dots, \nabla\varphi_k$ are linearly independent at any point $\theta \in \mathbb{R}^d$, then $\varphi_1, \dots, \varphi_k$ are algebraically independent.

(c) If $\nabla\varphi_1, \dots, \nabla\varphi_k$ are linearly independent at a point $\widetilde{\theta} \in \mathbb{R}^d$, and $\varphi_1, \dots, \varphi_k \in \mathcal{R}^G_{\le \ell}$ with $k = \mathrm{trdeg}(\mathcal{R}^G_{\le \ell})$, then there is an open neighborhood U of $\widetilde{\theta}$ such that for every polynomial $\psi \in \mathcal{R}^G_{\le \ell}$, there is an analytic function $f : \mathbb{R}^k \to \mathbb{R}$ for which $\psi(\theta) = f(\varphi_1(\theta), \dots, \varphi_k(\theta))$ for all $\theta \in U$.

4.2. **Series expansion of the population risk.** For any partition π of $[\ell + m] \equiv \{1, \dots, \ell + m\}$, denote by $|\pi|$ the number of sets in π, and label these sets as $1, \dots, |\pi|$. For each $i \in [\ell + m]$, denote by $\pi(i) \in \{1, \dots, |\pi|\}$ the index of the set containing element i. For $0 \le m \le \ell$, define

$$M_{\ell,m}(\pi \mid \theta, \theta_*) = \mathbb{E}_{g_1,\dots,g_{|\pi|}} \left[\prod_{j=1}^{m} \left\langle g_{\pi(2j-1)}\theta,\ g_{\pi(2j)}\theta \right\rangle \cdot \prod_{j=2m+1}^{\ell+m} \left\langle \theta_*,\ g_{\pi(j)}\theta \right\rangle \right] \tag{4.4}$$

where the expectation is over independent group elements $g_1, \dots, g_{|\pi|} \sim \mathrm{Unif}(G)$.

Example 4.6. Consider $\ell = 3$, $m = 1$, and $\pi = \{\{1, 2\}, \{3, 4\}\}$. For this partition π, we have $|\pi| = 2$ and $(\pi(1), \pi(2), \pi(3), \pi(4)) = (1, 1, 2, 2)$. Letting $g_1, g_2 \sim \mathrm{Unif}(G)$ be two independent and uniformly distributed group elements,

$$M_{3,1}(\pi \mid \theta, \theta_*) = \mathbb{E}_{g_1,g_2} \left[\langle g_1\theta, g_1\theta \rangle \langle \theta_*, g_2\theta \rangle^2 \right]. \tag{4.5}$$

For $\pi = \{\{1, 3\}, \{2\}, \{4\}\}$, we have $|\pi| = 3$ and $(\pi(1), \pi(2), \pi(3), \pi(4)) = (1, 2, 1, 3)$. Then

$$M_{3,1}(\pi \mid \theta, \theta_*) = \mathbb{E}_{g_1,g_2,g_3} \left[\langle g_1\theta, g_2\theta \rangle \langle \theta_*, g_1\theta \rangle \langle \theta_*, g_3\theta \rangle \right]. \tag{4.6}$$

Similarly, for $\pi = \{\{1, 3, 4\}, \{2\}\}$, we have

$$M_{3,1}(\pi \mid \theta, \theta_*) = \mathbb{E}_{g_1,g_2} \left[\langle g_1\theta, g_2\theta \rangle \langle \theta_*, g_1\theta \rangle^2 \right]. \tag{4.7}$$

□

Define the set

$$\mathcal{P}(\ell, m) = \Big\{ \text{ partitions } \pi \text{ of } [\ell + m] : \ \pi(2j-1) \neq \pi(2j) \text{ for all } j = 1, \dots, m \Big\}. \tag{4.8}$$

That is, partitions $\pi \in \mathcal{P}(\ell, m)$ separate each pair of elements $\{1,2\}, \{3,4\}, \dots, \{2m-1, 2m\}$. Then the terms of the series expansion (4.1) are given by

$$S_\ell(\theta) = \frac{1}{\ell!} \sum_{m=0}^{\ell} \frac{1}{2^m} \binom{\ell}{m} \sum_{\pi \in \mathcal{P}(\ell,m)} (|\pi| - 1)! (-1)^{|\pi|} M_{\ell,m}(\pi \mid \theta, \theta_*). \tag{4.9}$$

We denote the approximation given by the first k terms as

$$R^k(\theta) = \sum_{\ell=1}^{k} \sigma^{-2\ell} S_\ell(\theta). \tag{4.10}$$

The following is our rigorous result corresponding to (4.1), which bounds the approximation error of $R(\theta)$ by $R^k(\theta)$ for $\|\theta\| \ll \sigma / \log \sigma$ and fixed k, as $\sigma \to \infty$. We provide its proof at the end of this section.

Lemma 4.7. *Fix any function $r : (0, \infty) \to (0, \infty)$ such that $r(\sigma) \to \infty$ and $r(\sigma) \cdot (\log \sigma)/\sigma \to 0$ as $\sigma \to \infty$. For each $k \geq 1$, there exist (θ_*, d, G)-dependent constants $C, \sigma_0 > 0$ depending also on k, such that for all $\sigma > \sigma_0$ and all $\theta \in \mathbb{R}^d$ with $\|\theta\| < r(\sigma)$,*

$$\begin{aligned}
\Big| R(\theta) - R^k(\theta) \Big| &\leq \left(\frac{C \log \sigma}{\sigma} \right)^{2k+2} (\|\theta\| \vee 1)^{2k+2} \\
\Big\| \nabla R(\theta) - \nabla R^k(\theta) \Big\| &\leq \left(\frac{C \log \sigma}{\sigma} \right)^{2k+2} (\|\theta\| \vee 1)^{2k+1} \\
\Big\| \nabla^2 R(\theta) - \nabla^2 R^k(\theta) \Big\| &\leq \left(\frac{C \log \sigma}{\sigma} \right)^{2k+2} (\|\theta\| \vee 1)^{2k}.
\end{aligned}$$

From the definition in (4.4), we observe that for any fixed $\theta_* \in \mathbb{R}^d$, the term $M_{\ell,m}(\pi \mid \theta, \theta_*)$ is a G-invariant polynomial function of θ. Counting the number of occurrences of θ, $M_{\ell,m}(\pi \mid \theta, \theta_*)$ has degree $\ell + m$ in θ. Hence, $S_\ell(\theta)$ is a G-invariant polynomial of degree 2ℓ. The following shows that, in fact, $S_\ell(\theta)$ is in the algebra generated by the polynomials $\mathcal{R}^G_{\leq \ell}$ of degree at most ℓ. (That is, S_ℓ is a polynomial function of elements of $\mathcal{R}^G_{\leq \ell}$.) Furthermore, its dependence on the polynomials of degree ℓ has an explicit form in terms of the moment tensor $T_\ell(\theta) = \mathbb{E}_g[(g\theta)^{\otimes \ell}]$ from (1.10). These properties will allow us to understand the dependence of $S_\ell(\theta)$ on the transcendence basis for $\mathcal{R}^G_{\leq \ell}$ constructed in Lemma 4.3.

Lemma 4.8. *For each fixed $\theta_* \in \mathbb{R}^d$ and each $\ell \geq 1$, we have*

$$S_\ell(\theta) = \frac{1}{2(\ell!)} \big\| T_\ell(\theta) - T_\ell(\theta_*) \big\|^2_{\mathrm{HS}} + Q_\ell(\theta) \tag{4.11}$$

where $Q_\ell(\theta)$ is a polynomial (with coefficients depending on θ_*) in the algebra generated by $\mathcal{R}^G_{\leq \ell - 1}$. In particular, $S_\ell(\theta)$ is in the algebra generated by $\mathcal{R}^G_{\leq \ell}$.

Proof. We consider the terms $M_{\ell,m}(\pi \mid \theta, \theta_*)$ which constitute $S_\ell(\theta)$. For each $\pi \in \mathcal{P}(\ell, m)$, applying the constraint that $\pi(2j-1) \neq \pi(2j)$ for $j = 1, \dots, m$, we observe that each set in the partition π has cardinality at most ℓ, and hence each distinct group element g_i for $i = 1, \dots, |\pi|$ appears at most ℓ times inside the expectation in (4.4).

If each set in π has cardinality at most $\ell - 1$ (e.g. (4.5) and (4.6) in Example 4.6), then we claim that $M_{\ell,m}(\pi \mid \theta, \theta_*)$ is in the generated algebra of $\mathcal{R}^G_{\leq \ell - 1}$. To see this, observe that for any $k \leq \ell - 1$

and tensor $A \in (\mathbb{R}^d)^{\otimes k}$, we may write

$$\mathbb{E}_g\left[\sum_{i_1,\dots,i_k=1}^{d}\left(\prod_{j=1}^{k}(g\theta)_{i_j}\right)A_{i_1,\dots,i_k}\right] = \mathbb{E}_g\Big[\big\langle (g\theta)^{\otimes k},\, A\big\rangle\Big] = \Big\langle T_k(\theta),\, A\Big\rangle.$$

Each entry of the moment tensor $T_k(\theta)$ is a G-invariant polynomial of degree k, and hence belongs to $\mathcal{R}^G_{\le k}$. Applying this identity once for each distinct element $g_1,\dots,g_{|\pi|}$ in (4.4), and using that each such element appears $k \le \ell-1$ times, we get that $M_{\ell,m}(\pi \mid \theta,\theta_*)$ belongs to the algebra generated by $\mathcal{R}^G_{\le \ell-1}$. Absorbing the contributions of these terms $M_{\ell,m}(\pi \mid \theta,\theta_*)$ into $Q_\ell(\theta)$, it remains to consider those partitions $\pi \in \mathcal{P}(\ell,m)$ where some set in π has cardinality ℓ.

Without loss of generality, let us order the sets of π so that its first set has cardinality ℓ. Then g_1 appears ℓ times in (4.4), so exactly one of $\{\pi(2j-1),\pi(2j)\}$ must be 1 for each $j=1,\dots,m$, and every $\pi(j)$ must be 1 for $j = 2m+1,\dots,\ell+m$. For notational convenience, consider π such that $\pi(2j-1)=1$ for each $j=1,\dots,m$ (e.g. (4.7) in Example 4.6). For such π, we have

$$M_{\ell,m}(\pi \mid \theta,\theta_*) = \mathbb{E}_{g_1,\dots,g_{|\pi|}}\big[\langle g_1\theta, g_{\pi(2)}\theta\rangle \dots \langle g_1\theta, g_{\pi(2m)}\theta\rangle \langle g_1\theta,\theta_*\rangle^{\ell-m}\big]. \tag{4.12}$$

Suppose now that there is a second set of π which has cardinality at most $\ell-1$, corresponding to the element g_2. Then g_2 appears between 1 and $\ell-1$ times in $g_{\pi(2)}, g_{\pi(4)},\dots,g_{\pi(2m)}$. We may decouple the corresponding g_1's by introducing a new independent variable $\tilde{g}_1 \sim \mathrm{Unif}(G)$, setting $\tilde{g}_2 = \tilde{g}_1 g_1^{-1} g_2$, and writing

$$\langle g_1\theta, g_2\theta\rangle = \langle \theta, g_1^{-1}g_2\theta\rangle = \langle \tilde{g}_1\theta, \tilde{g}_2\theta\rangle.$$

The expectation over the uniform random pair (g_1,g_2) may be replaced by that over the uniform random triple $(g_1,\tilde{g}_1,\tilde{g}_2)$, reducing (4.12) into an expectation where each distinct group element now appears $\le \ell-1$ times. Then by the argument for the previous case, we also have that $M_{\ell,m}(\pi \mid \theta,\theta_*)$ belongs to the algebra generated by $\mathcal{R}^G_{\le \ell-1}$ in this case, and these terms may be absorbed into $Q_\ell(\theta)$.

The only partitions that remain are those where every set in π has cardinality ℓ. One such partition corresponds to $m=0$, where $\pi = \{\{1,2,\dots,\ell\}\}$. For this π, we have

$$M_{\ell,m}(\pi \mid \theta,\theta_*) = \mathbb{E}_g[\langle \theta_*, g\theta\rangle^\ell] = \mathbb{E}_{g_1,g_2}[\langle g_1\theta_*, g_2\theta\rangle^\ell] = \langle T_\ell(\theta_*), T_\ell(\theta)\rangle.$$

The remaining $2^{\ell-1}$ such partitions correspond to $m=\ell$ and $|\pi|=2$, where we may assume without loss of generality that $1\in\pi(1)$ and $2\in\pi(2)$, and take one element of each remaining pair $\{\pi(2j-1),\pi(2j)\}$ for $j=1,\dots,\ell$ to belong to $\pi(1)$ and the other to belong to $\pi(2)$. For these partitions π, we have

$$M_{\ell,m}(\pi \mid \theta,\theta_*) = \mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle^\ell] = \|T_\ell(\theta)\|^2_{\mathrm{HS}}.$$

Applying the above two displays to (4.9), we obtain

$$S_\ell(\theta) = -\frac{1}{\ell!}\langle T_\ell(\theta_*), T_\ell(\theta)\rangle + \frac{1}{2(\ell!)}\|T_\ell(\theta)\|^2_{\mathrm{HS}} + Q_\ell(\theta)$$

for some Q_ℓ in the algebra generated by $\mathcal{R}^G_{\le \ell-1}$. Completing the square yields $S_\ell(\theta) = \frac{1}{2(\ell!)}\|T_\ell(\theta) - T_\ell(\theta_*)\|^2_{\mathrm{HS}} - \frac{1}{2(\ell!)}\|T_\ell(\theta_*)\|^2_{\mathrm{HS}} + Q_\ell(\theta)$, where $\|T_\ell(\theta_*)\|^2_{\mathrm{HS}}$ does not depend on θ and can be absorbed into $Q_\ell(\theta)$. We thus arrive at the stated form of $S_\ell(\theta)$ in (4.11). Since the entries of $T_\ell(\theta)$ belong to $\mathcal{R}^G_{\le \ell}$, we obtain also that S_ℓ belongs to the algebra generated by $\mathcal{R}^G_{\le \ell}$. □

The following computation of the first three terms of (4.1) will be useful in our analysis of specific group actions. By Lemma 2.4, we assume without loss of generality that $\mathbb{E}_g[g]=0$.

Lemma 4.9. *If* $\mathbb{E}_g[g]=0$, *then*

$$S_1(\theta) = 0$$

$$S_2(\theta) = -\tfrac{1}{2}\mathbb{E}_g[\langle\theta_*, g\theta\rangle^2] + \tfrac{1}{4}\mathbb{E}_g[\langle\theta, g\theta\rangle^2]$$
$$S_3(\theta) = -\tfrac{1}{6}\mathbb{E}_g[\langle\theta_*, g\theta\rangle^3] + \tfrac{1}{12}\mathbb{E}_g[\langle\theta, g\theta\rangle^3] + \tfrac{1}{2}\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle\langle\theta_*, g_1\theta\rangle\langle\theta_*, g_2\theta\rangle] - \tfrac{1}{3}\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle\langle\theta, g_1\theta\rangle\langle\theta, g_2\theta\rangle].$$

Proof. If $\mathbb{E}_g[g] = 0$, then by (4.4), any $\pi \in \mathcal{P}(\ell, m)$ which has a singleton yields $M_{\ell,m}(\pi \mid \theta, \theta_*) = 0$.

For $\ell = 1$ and $m \in \{0, 1\}$, every $\pi \in \mathcal{P}(\ell, m)$ has a singleton, so $S_1(\theta) = 0$.

For $\ell = 2$ and $m \in \{0, 1, 2\}$, the only partitions $\pi \in \mathcal{P}(\ell, m)$ which do not have a singleton are $\{\{1, 2\}\}$ for $m = 0$ and $\{\{1, 3\}, \{2, 4\}\}$ and $\{\{1, 4\}, \{2, 3\}\}$ for $m = 2$. We get

$$\begin{aligned} S_2(\theta) &= -\tfrac{1}{2}M_{2,0}(\{\{1,2\}\}) + \tfrac{1}{8}M_{2,2}(\{\{1,3\},\{2,4\}\}) + \tfrac{1}{8}M_{2,2}(\{\{1,4\},\{2,3\}\}) \\ &= -\tfrac{1}{2}\mathbb{E}_g[\langle\theta_*, g\theta\rangle^2] + \tfrac{1}{4}\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle^2] \\ &= -\tfrac{1}{2}\mathbb{E}_g[\langle\theta_*, g\theta\rangle^2] + \tfrac{1}{4}\mathbb{E}_g[\langle\theta, g\theta\rangle^2], \end{aligned}$$

the last line applying the equality in law $g_1^\top g_2 \overset{L}{=} g_1$.

For $\ell = 3$, grouping together $\pi \in \mathcal{P}(\ell, m)$ that yield the same value of $M_{\ell,m}(\pi \mid \theta, \theta_*)$ by symmetry, we may check that

$$\begin{aligned} S_3(\theta) &= -\tfrac{1}{6}M_{3,0}(\{\{1,2,3\}\}) + 2\cdot\tfrac{1}{4}M_{3,1}(\{1,3\},\{2,4\}) + 4\cdot\tfrac{1}{8}M_{3,2}(\{1,3,5\},\{2,4\}) \\ &\quad + 4\cdot\tfrac{1}{48}M_{3,3}(\{1,3,5\},\{2,4,6\}) - 8\cdot\tfrac{1}{24}M_{3,3}(\{1,3\},\{2,5\},\{4,6\}) \\ &= -\tfrac{1}{6}\mathbb{E}_g[\langle\theta_*, g\theta\rangle^3] + \tfrac{1}{2}\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle\langle\theta_*, g_1\theta\rangle\langle\theta_*, g_2\theta\rangle] + \tfrac{1}{2}\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle^2\langle\theta_*, g_1\theta\rangle] \\ &\quad + \tfrac{1}{12}\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle^3] - \tfrac{1}{3}\mathbb{E}_{g_1,g_2,g_3}[\langle g_1\theta, g_2\theta\rangle\langle g_1\theta, g_3\theta\rangle\langle g_2\theta, g_3\theta\rangle]. \end{aligned}$$

By the equality in joint law $(g_1^\top g_2, g_1) \overset{L}{=} (g_2, g_1)$, the third term vanishes because

$$\mathbb{E}_{g_1,g_2}[\langle g_1\theta, g_2\theta\rangle^2\langle\theta_*, g_1\theta\rangle] = \mathbb{E}_{g_1,g_2}[\langle\theta, g_2\theta\rangle^2\langle\theta_*, g_1\theta\rangle] = \mathbb{E}_{g_2}[\langle\theta, g_2\theta\rangle^2]\mathbb{E}_{g_1}[\langle\theta_*, g_1\theta\rangle] = 0.$$

Applying $g_1^\top g_2 \overset{L}{=} g$ and $(g_1^\top g_2, g_1^\top g_3, g_2^\top g_3) \overset{L}{=} (g_1^\top g_2, g_1^\top, g_2^\top)$ to the remaining terms yields the form of S_3. □

Proof of Lemma 4.7. For notational convenience, set

$$z = \sigma^{-1}, \qquad s(z) = r(z^{-1}) = r(\sigma), \qquad q(z) = \log(z^{-1}) = \log\sigma.$$

The given conditions are $s(z) \to \infty$ and $zs(z)q(z) \to 0$ as $z \to 0$.

Recalling the form of $R(\theta)$ in (2.2), we consider the series expansion of the cumulant generating function

$$\log \mathbb{E}_g[e^{f(g)}] = \sum_{k=1}^{\infty} \frac{1}{k!}\,\kappa_k(f(g)) \tag{4.13}$$

for $f(g) = \langle z^2\theta_* + z\varepsilon, g\theta\rangle$, where $\kappa_k(f(g))$ is the k^{th} cumulant of $f(g)$ over the law $g \sim \mathrm{Unif}(G)$, conditional on ε. See Appendix A.1 for definitions.

We wish to take the expectation $\mathbb{E}_\varepsilon$ of this sum. To justify an exchange of $\mathbb{E}_\varepsilon$ and $\sum_k$ using Fubini's theorem, we consider the event $\|\varepsilon\| \le q(z)$ and set

$$Q(\theta) = \frac{\|\theta\|^2}{2}z^2 - \sum_{k=1}^{\infty}\frac{1}{k!}\mathbb{E}_\varepsilon\left[\kappa_k\Big(\langle z^2\theta_* + z\varepsilon,\ g\theta\rangle\Big)\mathbf{1}\{\|\varepsilon\| \le q(z)\}\right].$$

For $\|\theta\| < s(z)$ and on this event $\|\varepsilon\| \le q(z)$, observe that $\max_{g\in G}|f(g)| \le (z^2\|\theta_*\| + zq(z))s(z)$. By the given condition $zs(z)q(z) \to 0$ as $z \to 0$ (which also implies $z^2s(z) \to 0$), and by Lemma A.1(c), we observe that this series defining $Q(\theta)$ is absolutely convergent whenever $z < z_0$, for a small enough constant $z_0 > 0$. Then, writing (2.2) as

$$R(\theta) = \frac{\|\theta\|^2}{2}z^2 - \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\|\varepsilon\| \le q(z)\}\cdot\log\mathbb{E}_g[e^{f(g)}]\Big] - \mathbb{E}_\varepsilon\Big[\mathbf{1}\{\|\varepsilon\| > q(z)\}\cdot\log\mathbb{E}_g[e^{f(g)}]\Big]$$

and applying (4.13) and Fubini's theorem to exchange $\mathbb{E}_\varepsilon$ and $\sum_k$ in the second term, we arrive at

$$R(\theta) = Q(\theta) - \mathbb{E}_\varepsilon\left[\mathbf{1}\{\|\varepsilon\| > q(z)\} \cdot \log \mathbb{E}_g\left[e^{\langle z^2\theta_* + z\varepsilon, g\theta\rangle}\right]\right]. \tag{4.14}$$

It will be notationally convenient to rewrite $Q(\theta)$ using the cumulant tensors of g: Define the order-k moment tensor $\mathcal{T}_k(g)$ of g by

$$\mathcal{T}_k(g) = \mathbb{E}_g[g^{\otimes k}] \tag{4.15}$$

where $g^{\otimes k} \in (\mathbb{R}^{d\times d})^{\otimes k}$ is the k-fold tensor product of the linear map $g : \mathbb{R}^d \to \mathbb{R}^d$, acting on $(\mathbb{R}^d)^{\otimes k}$ via $g^{\otimes k}(v_1 \otimes \ldots \otimes v_k) = gv_1 \otimes \ldots \otimes gv_k$. Define the order-$k$ cumulant tensor $\mathcal{K}_k(g)$ by the moment-cumulant relation

$$\mathcal{K}_k(g) = \sum_{\text{partitions } \pi \text{ of } [k]} (|\pi|-1)!(-1)^{|\pi|-1} \bigotimes_{S\in\pi} \mathcal{T}_S(g), \tag{4.16}$$

which is analogous to the usual moment-cumulant relation for scalar random variables in (A.1). Here $\mathcal{T}_S(g)$ is the order-$|S|$ moment tensor of g acting on $(\mathbb{R}^d)^{\otimes S}$, corresponding to the $|S|$ coordinates belonging to S. For vectors $v_i, w_i \in \mathbb{R}^d$, we have the relation

$$\left\langle \bigotimes_{i\in S} v_i,\ \mathcal{T}_S(g)\left(\bigotimes_{i\in S} w_i\right)\right\rangle = \mathbb{E}_g\left[\left\langle \bigotimes_{i\in S} v_i,\ \bigotimes_{i\in S}(gw_i)\right\rangle\right] = \mathbb{E}_g\left[\prod_{i\in S}\langle v_i, gw_i\rangle\right].$$

Applying this, (4.16), and (A.1), we obtain

$$\left\langle \bigotimes_{i=1}^k v_i,\ \mathcal{K}_k(g)\left(\bigotimes_{i=1}^k w_i\right)\right\rangle = \kappa_k\Big(\langle v_1, gw_1\rangle, \ldots, \langle v_k, gw_k\rangle\Big). \tag{4.17}$$

Recall that $\kappa_k(f(g)) = \kappa_k(f(g), \ldots, f(g))$, where the latter mixed cumulant function is multi-linear and permutation invariant in its arguments. Applying (4.17) followed by a binomial expansion, we get

$$\kappa_k\Big(\langle z^2\theta_* + z\varepsilon,\ g\theta\rangle\Big) = \Big\langle (z^2\theta_* + z\varepsilon)^{\otimes k},\ \mathcal{K}_k(g)\theta^{\otimes k}\Big\rangle = \sum_{j=0}^k z^{2k-j}\binom{k}{j}\Big\langle \varepsilon^{\otimes j}\otimes\theta_*^{\otimes(k-j)},\ \mathcal{K}_k(g)\theta^{\otimes k}\Big\rangle.$$

So

$$Q(\theta) = \frac{\|\theta\|^2}{2}z^2 - \sum_{k=1}^\infty\sum_{j=0}^k \frac{z^{2k-j}}{j!(k-j)!}\Big\langle \mathbb{E}_\varepsilon[\varepsilon^{\otimes j}\mathbf{1}\{\|\varepsilon\|\le q(z)\}]\otimes\theta_*^{\otimes(k-j)},\ \mathcal{K}_k(g)\theta^{\otimes k}\Big\rangle.$$

Note that $\mathbb{E}_\varepsilon[\varepsilon^{\otimes j}\mathbf{1}\{\|\varepsilon\|\le q(z)\}] = 0$ if j is odd, by sign symmetry of the law of ε conditional on $\|\varepsilon\| \le q(z)$. Reparametrizing the terms for even j by $j = 2m$ and $\ell = k - m$, it may be checked that $\{(k,j) : k \ge 1, 0 \le j \le k\}$ is in bijection with $\{(\ell, m) : \ell \ge 1, 0 \le m \le \ell\}$. Thus, we obtain

$$Q(\theta) = \sum_{\ell=1}^\infty z^{2\ell} Q_\ell(\theta) \tag{4.18}$$

where

$$Q_\ell(\theta) = \mathbf{1}\{\ell = 1\}\frac{\|\theta\|^2}{2} - \sum_{m=0}^\ell \frac{1}{(2m)!(\ell-m)!}\Big\langle \mathbb{E}\big[\varepsilon^{\otimes 2m}\mathbf{1}\{\|\varepsilon\|\le q(z)\}\big]\otimes\theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)}\Big\rangle.$$

Applying the cumulant bound of Lemma A.1 together with (4.17) and $k! \ge k^k/e^k$, for $\ell \ge 2$,

$$|Q_\ell(\theta)| \le \sum_{m=0}^\ell \frac{1}{(2m)!(\ell-m)!}\mathbb{E}_\varepsilon\Big[\Big|\Big\langle \varepsilon^{\otimes 2m}\otimes\theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)}\Big\rangle\Big|\ \mathbf{1}\{\|\varepsilon\|\le q(z)\}\Big]$$

$$\le \sum_{m=0}^{\ell} \frac{1}{(\ell+m)!}\binom{\ell+m}{2m}(\ell+m)^{\ell+m} q(z)^{2m}\|\theta_*\|^{\ell-m}\|\theta\|^{\ell+m}$$
$$\le e^{2\ell}\|\theta\|^{2\ell}\sum_{m=0}^{\ell}\binom{\ell+m}{2m} q(z)^{2m}\|\theta_*\|^{\ell-m} \le e^{2\ell}(q(z)+\|\theta_*\|)^{2\ell}\|\theta\|^{2\ell}. \tag{4.19}$$

Then for $\|\theta\| < s(z)$ and $z < z_0$, the series in (4.18) is absolutely convergent. Differentiating each $Q_\ell(\theta)$ in θ using the product rule, a similar argument shows that for $\ell \ge 2$,

$$\|\nabla Q_\ell(\theta)\| \le 2\ell e^{2\ell}(q(z)+\|\theta_*\|)^{2\ell}\|\theta\|^{2\ell-1}, \tag{4.20}$$
$$\|\nabla^2 Q_\ell(\theta)\| \le 2\ell(2\ell-1)e^{2\ell}(q(z)+\|\theta_*\|)^{2\ell}\|\theta\|^{2\ell-2}. \tag{4.21}$$

Then both $\sum_\ell z^{2\ell}\nabla Q_\ell(\theta)$ and $\sum_\ell z^{2\ell}\nabla^2 Q_\ell(\theta)$ are also absolutely and uniformly convergent over $\|\theta\| < s(z)$, so

$$\nabla Q(\theta) = \sum_{\ell=1}^{\infty} z^{2\ell}\nabla Q_\ell(\theta), \qquad \nabla^2 Q(\theta) = \sum_{\ell=1}^{\infty} z^{2\ell}\nabla^2 Q_\ell(\theta).$$

We now fix an integer $k \ge 1$ and remove the truncation event $\|\varepsilon\| \le q(z)$. Note first that by Cauchy-Schwarz and a chi-squared tail bound, for all $z < z_0$ and some constants $C, c, z_0 > 0$, the second term in (4.14) is at most

$$\begin{aligned}\left|\mathbb{E}_\varepsilon\left[\mathbf{1}\{\|\varepsilon\| > q(z)\}\cdot \log \mathbb{E}_g\left[e^{\langle z^2\theta_* + z\varepsilon, g\theta\rangle}\right]\right]\right| &\le \mathbb{E}_\varepsilon\left[\mathbf{1}\{\|\varepsilon\| > q(z)\}\cdot\|z^2\theta_* + z\varepsilon\|\cdot\|\theta\|\right]\\ &\le \|\theta\|\cdot\mathbb{P}\big[\|\varepsilon\| > q(z)\big]^{1/2}\mathbb{E}_\varepsilon\big[\|z^2\theta_* + z\varepsilon\|^2\big]^{1/2}\\ &\le s(z)\cdot e^{-cq(z)^2}\cdot Cz.\end{aligned}$$

Recalling $zs(z) \to 0$ and $q(z) = \log(1/z)$, there exists z_0 (depending on k) such that $zs(z)e^{-cq(z)^2} \le z^{2k+2}$ for all $z < z_0$. Applying this to (4.14), and also using (4.19) to bound the sum over $\ell \ge k+1$ in (4.18), we obtain

$$\left|R(\theta) - \sum_{\ell=1}^{k} z^{2\ell} Q_\ell(\theta)\right| \le \big[Czq(z)(\|\theta\|\vee 1)\big]^{2k+2} \tag{4.22}$$

for $z < z_0$ and C, z_0 depending on k. For the gradient and Hessian, recall (2.3) and note that

$$\left\|\nabla_\theta \log \mathbb{E}_g\left[e^{\langle z^2\theta_* + z\varepsilon, g\theta\rangle}\right]\right\| = \left\|\mathbb{E}_g[g^\top(z^2\theta_* + z\varepsilon)\mid \varepsilon, \theta]\right\| \le \|z^2\theta_* + z\varepsilon\|,$$
$$\left\|\nabla^2_\theta \log \mathbb{E}_g\left[e^{\langle z^2\theta_* + z\varepsilon, g\theta\rangle}\right]\right\| = \left\|\operatorname{Cov}_g[g^\top(z^2\theta_* + z\varepsilon)\mid \varepsilon, \theta]\right\| \le \|z^2\theta_* + z\varepsilon\|^2.$$

Then applying a similar Cauchy-Schwarz argument together with (4.20) and (4.21), we get

$$\left\|\nabla R(\theta) - \sum_{\ell=1}^{k} z^{2\ell}\nabla Q_\ell(\theta)\right\| \le \big[Czq(z)\big]^{2k+2}(\|\theta\|\vee 1)^{2k+1}, \tag{4.23}$$
$$\left\|\nabla^2 R(\theta) - \sum_{\ell=1}^{k} z^{2\ell}\nabla^2 Q_\ell(\theta)\right\| \le \big[Czq(z)\big]^{2k+2}(\|\theta\|\vee 1)^{2k}. \tag{4.24}$$

Next, for each $\ell = 1, \dots, k$, define the version of Q_ℓ without the truncation event as

$$S_\ell(\theta) = \mathbf{1}\{\ell = 1\}\frac{\|\theta\|^2}{2} - \sum_{m=0}^{\ell}\frac{1}{(2m)!(\ell-m)!}\Big\langle \mathbb{E}_\varepsilon[\varepsilon^{\otimes 2m}]\otimes\theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)}\Big\rangle. \tag{4.25}$$

Then, for all $\ell \le k$ and some $C, z_0 > 0$ depending on k, the same Cauchy-Schwarz argument yields for $z < z_0$ that

$$|S_\ell(\theta) - Q_\ell(\theta)| \le \sum_{m=0}^{\ell} \frac{1}{(2m)!(\ell-m)!} \mathbb{P}\big[\|\varepsilon\| > q(z)\big]^{1/2} \mathbb{E}_\varepsilon\Big[\Big\langle \varepsilon^{\otimes 2m} \otimes \theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)}\Big\rangle^2\Big]^{1/2}$$
$$\le C \cdot \mathbb{P}[\|\varepsilon\| > q(z)]^{1/2} \cdot \|\theta\|^{2\ell} \le Ce^{-cq(z)^2}\|\theta\|^{2\ell} \le [Cz(\|\theta\| \vee 1)]^{2k+2}.$$

Applying this to each term $\ell = 1, \dots, k$ in (4.22), we get

$$\left| R(\theta) - \sum_{\ell=1}^{k} z^{2\ell} S_\ell(\theta) \right| \le [Czq(z)(\|\theta\| \vee 1)]^{2k+2}.$$

The differences $\|\nabla S_\ell(\theta) - \nabla Q_\ell(\theta)\|$ and $\|\nabla^2 S_\ell(\theta) - \nabla^2 Q_\ell(\theta)\|$ may be bounded similarly, and combined with (4.23) and (4.24) to show

$$\left\| \nabla R(\theta) - \sum_{\ell=1}^{k} z^{2\ell} \nabla S_\ell(\theta) \right\| \le [Czq(z)]^{2k+2}(\|\theta\| \vee 1)^{2k+1},$$
$$\left\| \nabla^2 R(\theta) - \sum_{\ell=1}^{k} z^{2\ell} \nabla^2 S_\ell(\theta) \right\| \le [Czq(z)]^{2k+2}(\|\theta\| \vee 1)^{2k}.$$

Recalling that $z = 1/\sigma$ and $q(z) = \log \sigma$, this concludes the proof upon verifying that $S_\ell(\theta)$ as defined in (4.25) is precisely the quantity defined in (4.9).

To show this, let us compute explicitly the expectation over $\varepsilon \sim \mathcal{N}(0, \mathrm{Id})$ in (4.25). Consider the identity matrix as an element of $(\mathbb{R}^d)^{\otimes 2}$,

$$\mathrm{Id} = \sum_{i=1}^{d} e_i \otimes e_i,$$

where e_i is the i^{th} standard basis vector in $\mathbb{R}^d$. For any pairing π of $[2m]$, denote $\bigotimes_{S\in\pi} \mathrm{Id} \in (\mathbb{R}^d)^{\otimes 2m}$ as the tensor product of m copies of Id that associates the two coordinates of each copy of Id with a pair $S \in \pi$. Using that the $2k^{\text{th}}$ moment of a standard Gaussian variable is the number of pairings of $[2k]$, we have for any basis vector $e_{i_1} \otimes \ldots \otimes e_{i_{2m}} \in (\mathbb{R}^d)^{2m}$ that

$$\langle \mathbb{E}_\varepsilon[\varepsilon^{\otimes 2m}], e_{i_1} \otimes \ldots e_{i_{2m}} \rangle = \mathbb{E}_\varepsilon \left[\prod_{j=1}^{2m} \varepsilon_{i_j} \right] = \sum_{\text{pairings } \pi \text{ of } [2m]} \prod_{(j_1,j_2)\in\pi} \mathbf{1}\{i_{j_1} = i_{j_2}\}$$
$$= \left\langle \sum_{\text{pairings } \pi \text{ of } [2m]} \left(\bigotimes_{S\in\pi} \mathrm{Id} \right),\ e_{i_1} \otimes \ldots \otimes e_{i_{2m}} \right\rangle.$$

Hence we see that

$$\mathbb{E}_\varepsilon[\varepsilon^{\otimes 2m}] = \sum_{\text{pairings } \pi \text{ of } [2m]} \bigotimes_{S\in\pi} \mathrm{Id}\,.$$

Applying (4.17) and the permutation invariance of $\kappa_{\ell+m}$ in its arguments, we get

$$\Big\langle \mathbb{E}_\varepsilon[\varepsilon^{\otimes 2m}] \otimes \theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)} \Big\rangle = (2m-1)!! \cdot \Big\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)} \Big\rangle, \quad (4.26)$$

since there are $(2m-1)!!$ total pairings, and by permutation invariance, the term corresponding to each pairing contributes equally to this inner product. (The right side of (4.26) corresponds to the consecutive pairing of $[2m]$.) Applying (4.26) and $(2m-1)!!/(2m)! = 1/(2^m m!)$ to (4.25),

$$S_\ell(\theta) = \mathbf{1}\{\ell = 1\}\frac{\|\theta\|^2}{2} - \frac{1}{\ell!} \sum_{m=0}^{\ell} \frac{1}{2^m} \binom{\ell}{m} \Big\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)} \Big\rangle. \quad (4.27)$$

Now we use (4.16) to write

$$\begin{aligned}
&\Big\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)} \Big\rangle \\
&= \sum_{\text{partitions } \pi \text{ of } [\ell+m]} (|\pi|-1)!(-1)^{|\pi|-1} \Big\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \Big(\bigotimes_{S\in\pi} \mathcal{T}_S(g)\Big)\theta^{\otimes(\ell+m)} \Big\rangle \\
&= \sum_{\text{partitions } \pi \text{ of } [\ell+m]} (|\pi|-1)!(-1)^{|\pi|-1} M_{\ell,m}(\pi) \qquad (4.28)
\end{aligned}$$

where we set

$$M_{\ell,m}(\pi) \equiv \Big\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \bigotimes_{S\in\pi} \mathbb{E}_g\big[(g\theta)^{\otimes S}\big] \Big\rangle.$$

We may move the expectations over g out of the inner product by writing this as an expectation over $|\pi|$ independent copies of g, one for each $S \in \pi$, so that

$$M_{\ell,m}(\pi) = \mathbb{E}_{g_1,\ldots,g_{|\pi|}} \left[\left\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \bigotimes_{i=1}^{\ell+m} (g_{\pi(i)}\theta) \right\rangle \right],$$

where for each $i \in [\ell+m]$, $\pi(i)$ denotes the index of the part in π containing i. Then using $\langle \mathrm{Id}, v\otimes w\rangle = \langle v, w\rangle$ and $\langle a\otimes b, c\otimes d\rangle = \langle a,c\rangle\langle b,d\rangle$, we see that this is exactly the quantity $M_{\ell,m}(\pi \mid \theta, \theta_*)$ defined previously in (4.4).

Finally, we combine (4.28) with (4.27) and describe a cancellation of terms that leads to (4.9): First, note that $\langle \mathrm{Id}, (g\theta)^{\otimes 2}\rangle = \langle g\theta, g\theta\rangle = \|\theta\|^2$, which does not depend on g. If $m \geq 1$ and $\{1,2\}$ belong to the same part in π, then

$$M_{\ell,m}(\pi) = \|\theta\|^2 M_{\ell-1,m-1}(\pi^-) \qquad (4.29)$$

where π^- is the partition of $\{3,\ldots,\ell+m\}$ obtained by removing 1 and 2. Suppose first that $\ell \geq 2$ and $m \geq 1$. Fix any partition π^- of $\{3,\ldots,\ell+m\}$. Let $\mathcal{S}$ be the collection of partitions of $[\ell+m]$ that do not separate $\{1,2\}$ and that reduce to π^- upon removing 1 and 2. There are two types of such partitions π: (a) π includes $1,2$ into a part of π^-. Then $|\pi| = |\pi^-|$ and there are $|\pi^-|$ such partitions; (b) π is the unique partition that adds $\{1,2\}$ as a new part to π^- so that $|\pi| = |\pi^-|+1$. Summing over both types and using (4.29), we get

$$\begin{aligned}
&\sum_{\pi\in\mathcal{S}} (|\pi|-1)!(-1)^{|\pi|-1} M_{\ell,m}(\pi) \\
&= \|\theta\|^2 M_{\ell-1,m-1}(\pi^-) \Big(|\pi^-| \cdot (|\pi^-|-1)!(-1)^{|\pi^-|-1} + 1\cdot(|\pi^-|)!(-1)^{|\pi^-|} \Big) = 0.
\end{aligned}$$

Summing over all π^-, the total contribution to (4.28) from partitions π that put $\{1,2\}$ in the same set is 0. Similarly, the total contribution to (4.28) from partitions π that put $\{3,4\}$ in the same set, but that *do not* put $\{1,2\}$ in the same set, is also 0, and so forth. Recalling the set of partitions $\mathcal{P}(\ell,m)$ defined in (4.8) which separate each pair $\{1,2\},\ldots,\{2m-1,2m\}$, we get in this case of $\ell \geq 2$ and $m \geq 1$ that only these partitions contribute to (4.28), i.e.

$$\Big\langle \mathrm{Id}^{\otimes m} \otimes \theta_*^{\otimes(\ell-m)},\ \mathcal{K}_{\ell+m}(g)\theta^{\otimes(\ell+m)} \Big\rangle = \sum_{\pi\in\mathcal{P}(\ell,m)} (|\pi|-1)!(-1)^{|\pi|-1} M_{\ell,m}(\pi).$$

Using that $\mathcal{P}(\ell,0)$ is simply the set of all partitions of $[\ell]$, and applying this to (4.27), we get that (4.27) is the same as (4.9) for $\ell \geq 2$. For $\ell = 1$, we have either $m = 0$ or $m = 1$. When $m = 1$, the only partition of $[\ell+m] = [2]$ not belonging to $\mathcal{P}(1,1)$ is $\{\{1,2\}\}$. Note that $M_{1,1}(\{\{1,2\}\}) = \mathbb{E}_g[\langle g\theta, g\theta\rangle] = \|\theta\|^2$, which cancels the leading term $\|\theta\|^2/2$ for $\ell = 1$ in (4.27). Thus (4.27) also concides with (4.9) for $\ell = 1$, concluding the proof. □

4.3. **Descent directions and pseudo-local-minimizers.** We now relate the series expansion result of Lemma 4.7 to the landscape of $R(\theta)$ around a fixed point $\widetilde{\theta} \in \mathbb{R}^d$, for large σ. The constants in this section may depend on this point $\widetilde{\theta}$.

The following lemma establishes a condition for $\widetilde{\theta} \in \mathbb{R}^d$ under which we will be able to show that $R(\theta)$ has either a first-order or second-order descent direction in a neighborhood $\widetilde{\theta}$.

Lemma 4.10. *Fix $\widetilde{\theta} \in \mathbb{R}^d$, let φ be a local reparametrization in an open neighborhood U of $\widetilde{\theta}$, and let $\widetilde{\varphi} = \varphi(\widetilde{\theta})$. Suppose there exists $\ell \geq 1$ and a partition of φ into subvectors $\varphi = (\varphi^1, \dots, \varphi^\ell)$ such that $S_1(\varphi), \dots, S_{\ell-1}(\varphi)$ are functions depending only on $\varphi^1, \dots, \varphi^{\ell-1}$ and not on φ^ℓ, and*

$$\textit{either} \quad \nabla_{\varphi^\ell} S_\ell(\widetilde{\varphi}) \neq 0 \quad \textit{or} \quad \lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi})) < 0.$$

Then there exist constants $c, \sigma_0 > 0$ and an open neighborhood U_0 of $\widetilde{\theta}$ (all depending on $\widetilde{\theta}, \theta_*, d, G$ but not on σ) such that for all $\sigma > \sigma_0$ and for every $\theta \in U_0$,

$$\textit{either} \quad \|\nabla_\theta R(\theta)\| \geq c\sigma^{-2\ell} \quad \textit{or} \quad \lambda_{\min}(\nabla^2_\theta R(\theta)) \leq -c\sigma^{-2\ell}.$$

Proof. First suppose that $\nabla_{\varphi^\ell} S_\ell(\widetilde{\varphi}) \neq 0$. Denote $c = \|\nabla_{\varphi^\ell} S_\ell(\widetilde{\varphi})\|$, and note that this constant c depends only on $\widetilde{\theta}, \theta_*, d, G$ and not on σ. By continuity of $\nabla_{\varphi^\ell} S_\ell$, this implies $\|\nabla_{\varphi^\ell} S_\ell(\varphi)\| > c/2$ for all φ in a neighborhood V_0 of $\widetilde{\varphi}$. Since $S_1, \dots, S_{\ell-1}$ do not depend on φ^ℓ, we have $\nabla_{\varphi^\ell} S_1 = \dots = \nabla_{\varphi^\ell} S_{\ell-1} = 0$. Then, recalling (4.10), we get $\|\nabla_\varphi R^\ell(\varphi)\| \geq \|\nabla_{\varphi^\ell} R^\ell(\varphi)\| > (c/2)\sigma^{-2\ell}$ for all $\varphi \in V_0$. Applying (4.2) and continuity and invertibility of $\mathsf{d}_\theta \varphi$ near $\widetilde{\theta}$, this implies that $\|\nabla_\theta R^\ell(\theta)\| > c'\sigma^{-2\ell}$ for a constant $c' > 0$ and all θ in a small enough neighborhood U_0 of $\widetilde{\theta}$. Then applying Lemma 4.7, for all $\sigma > \sigma_0$, large enough $\sigma > 0$, and all $\theta \in U_0$,

$$\|\nabla_\theta R(\theta)\| \geq (c'/2)\sigma^{-2\ell}.$$

Now suppose that $\lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi})) < 0$. The argument is similar: Denote $-c = \lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi}))$. Then $\lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\varphi)) < -c/2$ for all φ in a neighborhood V_0 of $\widetilde{\varphi}$ by continuity, so $\lambda_{\min}(\nabla^2_\varphi R^\ell(\varphi)) \leq \lambda_{\min}(\nabla^2_{\varphi^\ell} R^\ell(\varphi)) < -(c/2)\sigma^{-2\ell}$. Applying (4.3),

$$\nabla^2_\theta R^\ell(\theta) = (\mathsf{d}_\theta \varphi)^\top \cdot \nabla^2_\varphi R^\ell(\varphi) \cdot \mathsf{d}_\theta \varphi + \sum_{i=1}^d \partial_{\varphi_i} R^\ell(\varphi) \cdot \nabla^2_\theta \varphi_i.$$

Then by continuity and invertibility of $\mathsf{d}_\theta \varphi$ near $\widetilde{\theta}$, for the first term we have

$$\lambda_{\min}((\mathsf{d}_\theta \varphi)^\top \cdot \nabla^2_\varphi R^\ell(\varphi) \cdot \mathsf{d}_\theta \varphi) < -c'\sigma^{-2\ell}$$

for a constant $c' > 0$ and all θ in a neighborhood U_0 of $\widetilde{\theta}$. Then either $\lambda_{\min}(\nabla^2_\theta R^\ell(\theta)) < -(c'/2)\sigma^{-2\ell}$, or we must have for the second term and some $i \in \{1, \dots, d\}$ that $\|\nabla_\varphi R^\ell(\varphi)\| \geq |\partial_{\varphi_i} R^\ell(\varphi)| > c''\sigma^{-2\ell}$. Here, we may take $c'' = c'/(2d \max \|\nabla^2_\theta \varphi_j(\theta)\|)$, where this maximum is taken over all $j \in \{1, \dots, d\}$ and $\theta \in U_0$. Applying again Lemma 4.7, for all $\sigma > \sigma_0$ and large enough $\sigma_0 > 0$, this implies that for every $\theta \in U_0$, either $\lambda_{\min}(\nabla^2_\theta R^\ell(\theta)) \leq -(c'/4)\sigma^{-2\ell}$ or $\|\nabla_\theta R^\ell(\theta)\| \geq (c''/2)\sigma^{-2\ell}$. □

Conversely, the following is a condition for $\widetilde{\theta} \in \mathbb{R}^d$ under which we will show that $R(\theta)$ has a local minimizer in any fixed neighborhood of $\widetilde{\theta}$, for all sufficiently large σ. We call these points pseudo-local-minimizers, and these will be in correspondence with the true local minimizers of $R(\theta)$ for large σ. Note that pseudo-local-minimizers are determined by θ_*, d, G and do not depend on σ, but true local minimizers of $R(\theta)$ not belonging to $\mathcal{O}_{\theta_*}$ may (in general) depend on σ.

Definition 4.11. A point $\widetilde{\theta} \in \mathbb{R}^d$ is a **pseudo-local-minimizer** in a local reparametrization $\varphi = (\varphi^1, \dots, \varphi^L)$ around $\widetilde{\theta}$ if each function $S_\ell(\varphi)$ for $\ell = 1, \dots, L$ depends only on $\varphi^1, \dots, \varphi^\ell$ and

not on $\varphi^{\ell+1}, \dots, \varphi^L$, and for each $\ell \in \{1, \dots, L\}$ where φ^ℓ has non-zero dimension,

$$\nabla_{\varphi^\ell} S_\ell(\widetilde{\varphi}) = 0 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi})) > 0.$$

For each pseudo-local-minimizer $\widetilde{\theta}$, we will also show that the risk $R(\varphi)$ is strongly convex in a σ-independent neighborhood of $\widetilde{\varphi} = \varphi(\widetilde{\theta})$, and its Hessian $\nabla^2_\varphi R(\varphi)$ has the following graded block structure.

Definition 4.12. Consider a partition of coordinates $(\varphi^1, \dots, \varphi^L)$ for $\mathbb{R}^d$. Let $H \equiv H(\sigma) \in \mathbb{R}^{d \times d}$ be a symmetric matrix, and write its $L \times L$ block decomposition with respect to this partition as

$$H = \begin{pmatrix} H_{11} & \cdots & H_{1L} \\ \vdots & \ddots & \vdots \\ H_{L1} & \cdots & H_{LL} \end{pmatrix}.$$

The matrix $H(\sigma)$ has a **graded block structure** with respect to this partition if there are constants $C, c, \sigma_0 > 0$ such that for all $\sigma > \sigma_0$ and all $k, \ell \in \{1, \dots, L\}$ where φ^k and φ^ℓ have non-zero dimension,

$$C\sigma^{-2\ell} \ge \lambda_{\max}(H_{\ell\ell}) \ge \lambda_{\min}(H_{\ell\ell}) \ge c\sigma^{-2\ell} \qquad \text{and} \qquad \|H_{k\ell}\| \le C\sigma^{-2\max(k,\ell)}.$$

Thus the upper-left block of $H(\sigma)$ has magnitude σ^{-2}, the three blocks adjacent to this have magnitude σ^{-4}, and so forth. We allow φ^ℓ to have dimension 0, in which case the blocks $H_{k\ell}$ and $H_{\ell k}$ for $k = 1, \dots, L$ are empty.

Lemma 4.13. *Let $\widetilde{\theta} \in \mathbb{R}^d$ be a pseudo-local-minimizer in the reparametrization $\varphi = (\varphi^1, \dots, \varphi^L)$. Denote $\widetilde{\varphi} = \varphi(\widetilde{\theta})$. Then for any sufficiently small open neighborhood V_0 of φ, there exist constants $c, \sigma_0 > 0$ depending on $\widetilde{\theta}, V_0$ and θ_*, d, G but not on σ, such that for all $\sigma > \sigma_0$ and $\varphi \in V_0$,*
(a) $\nabla^2_\varphi R(\varphi)$ has a graded block structure with respect to the partition $\varphi = (\varphi^1, \dots, \varphi^L)$,
(b) $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \ge c\sigma^{-2L}$, and
(c) There is a unique critical point of $R(\varphi)$ in V_0, which is a local minimizer of $R(\varphi)$.

Proof of Lemma 4.13. For part (a), observe that the Hessian $\nabla^2_\varphi S_\ell(\varphi)$ is non-zero only in the upper-left $\ell \times \ell$ blocks of the decomposition corresponding to $(\varphi^1, \dots, \varphi^L)$. Since $\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi})$ is positive-definite by assumption, by continuity there is a neighborhood V_0 of $\widetilde{\varphi}$ and constants $C, c > 0$ for which

$$\lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\varphi)) \ge c \qquad \text{and} \qquad \|\nabla^2_\varphi S_\ell(\varphi)\| \le C \tag{4.30}$$

for all $\varphi \in V_0$. Applying this for each $\ell = 1, \dots, L$ and recalling (4.10), we see that $\nabla^2_\varphi R_L(\varphi)$ has a graded block structure. Then $\nabla^2_\varphi R(\varphi)$ also has a graded block structure, by Lemma 4.7. This shows (a). Part (b) will follow from (a) and Lemma 4.15 which we prove in the next section.

To show part (c), let us assume for expositional simplicity that each φ^ℓ has positive dimension—the same argument applies with minor modification to the setting where some of the vectors φ^ℓ have dimension 0. Let $\check{\varphi} = (\check{\varphi}^1, \dots, \check{\varphi}^L)$ be a point which minimizes $R(\varphi)$ over the compact set $\overline{V}_0$. Observe that the given condition implies $\widetilde{\varphi}$ is a local and global minimizer of S_1 over V_0, and that

$$S_1(\check{\varphi}) - S_1(\widetilde{\varphi}) \ge c\|\check{\varphi}^1 - \widetilde{\varphi}^1\|^2.$$

Then by Lemma 4.7, for all $\sigma > \sigma_0$ and large enough $\sigma_0 > 0$,

$$R(\check{\varphi}) - R(\widetilde{\varphi}) \ge c\sigma^{-2}\|\check{\varphi}^1 - \widetilde{\varphi}^1\|^2 - C\left(\frac{\log \sigma}{\sigma}\right)^4.$$

The left side is non-positive because $\check{\varphi}$ minimizes $R(\varphi)$, so we get $\|\check{\varphi}^1 - \widetilde{\varphi}^1\| \le \sigma^{-\tau}$ for, say, $\tau = 0.9$. Now consider the functions $f(\varphi^2) = S_2(\widetilde{\varphi}^1, \varphi^2)$ and $\check{f}(\varphi^2) = S_2(\check{\varphi}^1, \varphi^2)$. The given condition implies that f is strongly convex and has a local and global minimizer in V_0 given by $\widetilde{\varphi}^2$. Applying

the bound $\|\check{\varphi}^1 - \widetilde{\varphi}^1\| \le \sigma^{-\tau}$, we get that $\|f - \check{f}\| \le C\sigma^{-\tau}$ and $\|\nabla^2 f - \nabla^2 \check{f}\| \le C\sigma^{-\tau}$, for some constant $C > 0$ and any sufficiently small neighborhood V_0 of $\widetilde{\varphi}$. Then applying Lemma 2.7, $\check{f}$ is also strongly convex on V_0, with a local and global minimizer in V_0 given by some point $\bar{\varphi}^2$ for which $\|\bar{\varphi}^2 - \widetilde{\varphi}^2\| \le C'\sigma^{-\tau}$. This implies

$$S_2(\check{\varphi}^1, \check{\varphi}^2) - S_2(\check{\varphi}^1, \bar{\varphi}^2) \ge c\|\check{\varphi}^2 - \bar{\varphi}^2\|^2.$$

Since S_1 depends only on φ^1 and not on φ^2, we have by Lemma 4.7 that

$$R(\check{\varphi}) - R((\check{\varphi}^1, \bar{\varphi}^2, \widetilde{\varphi}^3, \dots, \widetilde{\varphi}^L)) \ge c\sigma^{-4}\|\check{\varphi}^2 - \bar{\varphi}^2\|^2 - C\left(\frac{\log\sigma}{\sigma}\right)^6.$$

Then, since this is again non-positive, we obtain $\|\check{\varphi}^2 - \bar{\varphi}^2\| \le \sigma^{-\tau}$, and hence also $\|\check{\varphi}^2 - \widetilde{\varphi}^2\| \le C\sigma^{-\tau}$. Now applying this argument to $f(\varphi^3) = S_3(\widetilde{\varphi}^1, \widetilde{\varphi}^2, \varphi^3)$ and $\check{f}(\check{\varphi}^1, \check{\varphi}^2, \varphi^3)$, we obtain similarly $\|\check{\varphi}^3 - \widetilde{\varphi}^3\| \le C\sigma^{-\tau}$. Iterating this argument yields $\|\check{\varphi} - \widetilde{\varphi}\| \le C\sigma^{-\tau}$ for a constant $C > 0$. For any neighborhood V_0, large enough $\sigma_0 > 0$ (depending on V_0), and all $\sigma > \sigma_0$, this implies that this minimizer $\check{\varphi}$ belongs to the interior of V_0, and hence must be a critical point of $R(\varphi)$. Then the strong convexity in part (b) implies that this is the unique critical point in V_0, which shows (c). □

4.4. **Local landscape and Fisher information.** We apply Lemma 4.13 to analyze the Fisher information $I(\theta_*) = \nabla^2_\theta R(\theta_*)$ and the local landscape of $R(\theta)$ near θ_*. By rotational symmetry of $R(\theta)$, the same statements hold locally around each point in the orbit $\mathcal{O}_{\theta_*}$.

Recall the transcendence basis φ in Lemma 4.3, and the decompositions $d = d_1 + \ldots + d_L$ and $\varphi = (\varphi^1, \dots, \varphi^L)$ according to the sequence of subspaces $\mathcal{R}^G_{\le \ell}$ and their transcendence degrees. Lemma 4.5 establishes that φ is a local reparametrization around generic points θ_*, and we will analyze the landscape in this reparametrization.

Theorem 4.14. *Fix a choice of transcendence basis $\varphi = (\varphi^1, \dots, \varphi^L)$ satisfying Lemma 4.3, and let $\theta_* \in \mathbb{R}^d$ be a generic point where $\mathsf{d}_\theta\varphi(\theta_*)$ is non-singular. For some constants $C, c, \sigma_0 > 0$ and some neighborhood U of θ_*, and for all $\sigma \ge \sigma_0$,*

(a) In the reparametrization by φ, $R(\varphi)$ is strongly convex on $\varphi(U)$ with $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \ge c\sigma^{-2L}$.

(b) The Fisher information matrix $I(\theta_*)$ has d_ℓ eigenvalues belonging to $[c\sigma^{-2\ell}, C\sigma^{-2\ell}]$ for each $\ell = 1, \dots, L$, where $d_\ell = \mathrm{trdeg}(\mathcal{R}^G_{\le \ell}) - \mathrm{trdeg}(\mathcal{R}^G_{\le \ell-1})$.

(c) For any polynomial $\psi \in \mathcal{R}^G_{\le \ell}$, there is a constant $C > 0$ (depending also on ψ) such that

$$\nabla_\theta\psi(\theta_*)^\top I(\theta_*)^{-1}\nabla_\theta\psi(\theta_*) \le C\sigma^{2\ell}.$$

Note that part (c) describes the limiting variance in (1.9) for estimating $\psi(\theta_*)$ by the plug-in maximum likelihood estimate $\psi(\hat{\theta})$.

The proof of Theorem 4.14 relies on the following linear-algebraic result for any σ-dependent matrix with the graded block structure of Definition 4.12, and large enough σ.

Lemma 4.15. *Suppose $H \equiv H(\sigma) \in \mathbb{R}^{d\times d}$ has a graded block structure with respect $(\varphi^1, \dots, \varphi^L)$. Let d_ℓ be the dimension of each subvector φ^ℓ. Let $H_{:\ell,:\ell}$ and $(H^{-1})_{:\ell,:\ell}$ denote the submatrices consisting of the upper-left $\ell \times \ell$ blocks in the $L \times L$ block decompositions of H and H^{-1}. Then for some constants $C, c, \sigma_0 > 0$ and all $\sigma > \sigma_0$:*

(a) H has d_ℓ eigenvalues belonging to $[c\sigma^{-2\ell}, C\sigma^{-2\ell}]$ for each $\ell = 1, \dots, L$. In particular, $\lambda_{\min}(H) \ge c\sigma^{-2L}$.

(b) For each ℓ where $d_1 + \ldots + d_\ell > 0$, $\lambda_{\min}(H_{:\ell,:\ell}) \ge c\sigma^{-2\ell}$.

(c) For each ℓ where $d_1 + \ldots + d_\ell > 0$, $\lambda_{\max}((H^{-1})_{:\ell,:\ell}) \le C\sigma^{2\ell}$.

Proof. We first show part (b). This holds for the smallest ℓ where $d_1 + \ldots + d_\ell > 0$, by the definition of the graded block structure. Assume inductively that it holds for $\ell \le L - 1$, and consider $\ell + 1$

where $d_{\ell+1} > 0$. For any unit vector $v = (v_{:\ell}, v_{\ell+1})$ where $v_{:\ell} \in \mathbb{R}^{d_1+\ldots+d_\ell}$ and $v_{\ell+1} \in \mathbb{R}^{d_{\ell+1}}$, we have by the induction hypothesis and Cauchy-Schwarz

$$\begin{aligned} v^\top H_{:(\ell+1),:(\ell+1)} v &= v_{:\ell}^\top H_{:\ell,:\ell} v_{:\ell} + v_{\ell+1}^\top H_{\ell+1,\ell+1} v_{\ell+1} + 2 v_{:\ell}^\top H_{:\ell,\ell+1} v_{\ell+1} \\ &\geq c\sigma^{-2\ell} \|v_{:\ell}\|^2 + c\sigma^{-2(\ell+1)} \|v_{\ell+1}\|^2 - 2C\sigma^{-2(\ell+1)} \|v_{:\ell}\| \|v_{\ell+1}\| \\ &\geq \Big(c\sigma^{-2\ell} - (2C/c)\sigma^{-2(\ell+1)} \Big) \|v_{:\ell}\|^2 + (c/2)\sigma^{-2(\ell+1)} \|v_{\ell+1}\|^2. \end{aligned}$$

For large σ, we get $v^\top H_{:(\ell+1),:(\ell+1)} v \geq c'\sigma^{-2(\ell+1)}$ and some $c' > 0$. Hence (b) holds by induction for each $\ell = 1, \ldots, L$.

Next, we show part (a). That $\lambda_{\min}(H) \geq c\sigma^{-2L}$ follows from (b). For the first statement, for any ℓ where $d_\ell > 0$, write $H = H^{(\ell-1)} + R^{(\ell-1)}$ where $H^{(\ell-1)}$ equals $H_{:(\ell-1),:(\ell-1)}$ on the upper-left $(\ell-1) \times (\ell-1)$ blocks and is 0 elsewhere, and $R^{(\ell-1)}$ is the remainder. Part (a) implies that $H^{(\ell-1)}$ has $d_1 + \ldots + d_{\ell-1}$ eigenvalues at least $c\sigma^{-2(\ell-1)}$, and remaining eigenvalues 0. The graded block structure condition implies $\|R^{(\ell-1)}\| \leq C\sigma^{-2\ell}$ for a constant $C > 0$. Then for a constant $c' > 0$ and all large σ, Weyl's inequality implies that H has $d_{\ell-1}$ eigenvalues at least $c'\sigma^{-2(\ell-1)}$, and remaining eigenvalues at most $C\sigma^{-2\ell}$. Since this result holds for every $\ell = 1, \ldots, L$, this implies part (a).

Finally, for part (c), denote $G^{(\ell)} = [(H^{-1})_{:\ell,:\ell}]^{-1}$. We claim that for all ℓ where $d_1 + \ldots + d_\ell > 0$, this matrix $G^{(\ell)}$ has a graded block structure with respect to $(\varphi^1, \ldots, \varphi^\ell)$. That is to say, there are constants $C, c > 0$ such that for all large σ and all $1 \leq j, k \leq \ell$,

$$C\sigma^{-2j} \geq \lambda_{\max}(G^{(\ell)}_{jj}) \geq \lambda_{\min}(G^{(\ell)}_{jj}) \geq c\sigma^{-2j} \quad \text{and} \quad \|G^{(\ell)}_{jk}\| \leq C\sigma^{-2\max(j,k)}. \tag{4.31}$$

For $\ell = L$, we have $G^{(\ell)} = H$, so this holds by assumption. Assume inductively that it holds for $\ell + 1$, and consider ℓ where $d_{\ell+1} > 0$. Applying the definition of $G^{(\ell)}$ and the Schur complement identity,

$$[G^{(\ell)}]^{-1} = ([G^{(\ell+1)}]^{-1})_{:\ell,:\ell} = \Big(G^{(\ell+1)}_{:\ell,:\ell} - G^{(\ell+1)}_{:\ell,\ell+1} [G^{(\ell+1)}_{\ell+1,\ell+1}]^{-1} G^{(\ell+1)}_{\ell+1,:\ell} \Big)^{-1}.$$

Then

$$G^{(\ell)} = G^{(\ell+1)}_{:\ell,:\ell} - G^{(\ell+1)}_{:\ell,\ell+1} [G^{(\ell+1)}_{\ell+1,\ell+1}]^{-1} G^{(\ell+1)}_{\ell+1,:\ell}.$$

We have $\|G^{(\ell+1)}_{:\ell,\ell+1}\| \leq C'\sigma^{-2(\ell+1)}$ and $\|[G^{(\ell+1)}_{\ell+1,\ell+1}]^{-1}\| \leq C'\sigma^{2(\ell+1)}$ for some $C' > 0$, by the induction hypothesis. For large enough σ, applying the induction hypothesis also to each block of $G^{(\ell+1)}_{:\ell,:\ell}$, we get that (4.31) holds for ℓ (and some constants $C, c > 0$ different from those for $\ell+1$). Hence (4.31) holds by induction for each $\ell = 1, \ldots, L$. Then, applying part (b) to this matrix $G^{(\ell)}$ in place of H, we get that $\lambda_{\min}(G^{(\ell)}) \geq c\sigma^{-2\ell}$, which implies $\lambda_{\max}((H^{-1})_{:\ell,:\ell}) \leq C\sigma^{2\ell}$. This establishes (c). □

Proof of Theorem 4.14. We first show that θ_* is a pseudo-local-minimizer with respect to this reparametrization $\varphi = (\varphi^1, \ldots, \varphi^L)$. For this, we apply the form

$$S_\ell(\varphi) = \frac{1}{2(\ell!)} \|T_\ell(\varphi) - T_\ell(\varphi_*)\|_{\mathrm{HS}}^2 + Q_\ell(\varphi)$$

provided in Lemma 4.8, where $T_\ell(\varphi)$ and $Q_\ell(\varphi)$ are shorthand for $T(\theta(\varphi))$ and $Q(\theta(\varphi))$. Differentiating in φ,

$$\begin{aligned} \nabla_\varphi S_\ell(\varphi) &= \frac{1}{\ell!} \mathsf{d}_\varphi T_\ell(\varphi)^\top (T_\ell(\varphi) - T_\ell(\varphi_*)) + \nabla_\varphi Q_\ell(\varphi), \\ \nabla^2_\varphi S_\ell(\varphi) &= \frac{1}{\ell!} \mathsf{d}_\varphi T_\ell(\varphi)^\top \mathsf{d}_\varphi T_\ell(\varphi) + \frac{1}{\ell!} \sum_i (T_\ell(\varphi)_i - T_\ell(\varphi_*)_i) \nabla^2_\varphi T_\ell(\varphi)_i + \nabla^2_\varphi Q_\ell(\varphi). \end{aligned}$$

Here, $T_\ell(\varphi)_i$ is the i^{th} entry of $T_\ell(\varphi)$ and the summation is over all multi-indices i. Note that Q_ℓ is in the algebra generated by $\mathcal{R}^G_{\leq \ell-1}$, so Lemma 4.5(c) ensures that Q_ℓ depends only on $(\varphi^1, \ldots, \varphi^{\ell-1})$.

Thus, evaluating the above at $\varphi = \varphi_*$ and restricting to the coordinates φ^ℓ yields

$$\nabla_{\varphi^\ell} S_\ell(\varphi_*) = 0, \qquad \nabla^2_{\varphi^\ell} S_\ell(\varphi_*) = \frac{1}{\ell!} \mathsf{d}_{\varphi^\ell} T_\ell(\varphi_*)^\top \mathsf{d}_{\varphi^\ell} T_\ell(\varphi_*).$$

In particular, $\nabla^2_{\varphi^\ell} S_\ell(\varphi_*) \succeq 0$. To see that $\nabla^2_{\varphi^\ell} S_\ell(\varphi_*)$ has full rank d_ℓ, observe that every degree-ℓ polynomial of $\theta \in \mathbb{R}^d$ is a linear combination of entries of the tensors $\theta^{\otimes 1}, \dots, \theta^{\otimes \ell}$. Thus, symmetrizing by G, every polynomial in $\mathcal{R}^G_{\le \ell}$ is a linear combination of entries of $T_1, \dots, T_\ell$ (monomials). This means that $\varphi^\ell = f(T_1(\varphi), \dots, T_\ell(\varphi))$ for some linear function $f : \mathbb{R}^{d+d^2+\dots+d^\ell} \to \mathbb{R}^{d_\ell}$. Differentiating both sides in φ^ℓ and observing that $T_1, \dots, T_{\ell-1}$ do not depend on φ^ℓ by Lemma 4.5(c), we obtain

$$\mathrm{Id} = (\mathsf{d}_{T_\ell} f)(\mathsf{d}_{\varphi^\ell} T_\ell),$$

where the left side is the $d_\ell \times d_\ell$ identity. Thus $\mathsf{d}_{\varphi^\ell} T_\ell$ has full rank d_ℓ, so $\nabla^2_{\varphi^\ell} S_\ell(\varphi_*) \succ 0$ and θ_* is a pseudo-local-minimizer.

Then part (a) of the theorem follows immediately from Lemma 4.13(b). For (b) and (c), note that since $\nabla_\theta R(\theta_*) = 0$, we have from (4.3) that

$$I(\theta_*) \equiv \nabla^2_\theta R(\theta_*) = \Big(\mathsf{d}\varphi(\theta_*)^\top \cdot \nabla^2_\varphi R(\varphi_*) \cdot \mathsf{d}\varphi(\theta_*) \Big)$$

where $\varphi_* = \varphi(\theta_*)$. Then setting $\tilde{V} = \mathsf{d}\varphi(\theta_*)^{-1}$, Lemma 4.13 shows that $\nabla^2_\varphi R(\varphi_*) = \tilde{V}^\top I(\theta_*) \tilde{V}$ has a graded block structure. For any polynomial $\psi \in \mathcal{R}^G_{\le \ell}$, Lemma 4.5 shows that ψ is an analytic function of $\varphi^1, \dots, \varphi^\ell$, and hence that $\nabla_\varphi \psi = \tilde{V}^\top \nabla_\theta \psi$ is non-zero only in its first ℓ blocks. Writing

$$\nabla_\theta \psi(\theta_*)^\top I(\theta_*)^{-1} \nabla_\theta \psi(\theta_*) = \Big(\tilde{V}^\top \nabla_\theta \psi(\theta_*) \Big)^\top \Big(\tilde{V}^\top I(\theta_*) \tilde{V} \Big)^{-1} \Big(\tilde{V}^\top \nabla_\theta \psi(\theta_*) \Big),$$

part (c) then follows from Lemma 4.15(c). Also, by the QR decomposition, there is a non-singular lower-triangular matrix W for which $V = \tilde{V} W$ is orthogonal. It may be verified from Definition 4.12 that the matrix $V^\top I(\theta_*) V = W^\top (\tilde{V}^\top I(\theta_*) \tilde{V}) W$ also has a graded block structure, for modified constants C, c, σ_0. As the eigenvalues of $V^\top I(\theta_*) V$ are the same as those of $I(\theta_*)$, this and Lemma 4.15(a) show part (b). □

The following then shows that with high probability for $n \gg \sigma^{4L-2} \log \sigma$, the empirical risk $R_n(\varphi)$ is also strongly convex with a local minimizer in $\varphi(U)$.

Corollary 4.16. *In the setting of Theorem 4.14, for (θ_*, d, G)-dependent constants $C, c, c', \sigma_0 > 0$, with probability at least $1 - \sigma^C e^{-c\sigma^{-4L+2} n} - C e^{-cn^{2/3}}$ for all $\sigma > \sigma_0$, we have $\lambda_{\min}(\nabla^2_\varphi R_n(\varphi)) \ge c' \sigma^{-2L}$ for all $\varphi \in \varphi(U)$, and $R_n(\theta)$ has a local minimizer in U.*

Proof. For any bounded neighborhood U and constant $c_0 > 0$, applying Lemma 2.10 with $t = c_0 \sigma^{-2L}$, we have $\sup_{\theta \in U} |R_n(\theta) - R(\theta)|, \|\nabla R_n(\theta) - \nabla R(\theta)\|, \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \le c_0 \sigma^{-2L}$ with probability at least $1 - \sigma^C e^{-c\sigma^{-4L+2} n} - C e^{-cn^{2/3}}$. Then setting $C_0 = \sup_{\theta \in U} \|\mathsf{d}_\theta \varphi\|, \|\nabla^2_\theta \varphi_i\|$ and applying (4.2) and (4.3), on this event we have $\sup_{\varphi \in \varphi(U)} |R_n(\varphi) - R(\varphi)|, \|\nabla^2_\varphi R_n(\varphi) - \nabla^2_\varphi R(\varphi)\| \le (C_0^2 + dC_0) c_0 \sigma^{-2L}$. Picking c_0 small enough and applying Theorem 4.14 and Lemma 2.7, we see that $\lambda_{\min}(\nabla^2_\varphi R_n(\varphi)) \ge c' \sigma^{-2L}$ over $\varphi \in \varphi(U)$, and $R_n(\varphi)$ has a local minimizer in $\varphi(U)$. Then $R_n(\theta)$ has a local minimizer in U. □

4.5. **Globally benign landscapes at high noise.** In the following three subsections, we apply the tools of Section 4.3 to analyze three examples in which the landscapes of $R(\theta)$ and $R_n(\theta)$ are *globally* benign in this high-noise regime $\sigma > \sigma_0(\theta_*, d, G)$, for generic $\theta_* \in \mathbb{R}^d$.

In each example, for each fixed point $\widetilde{\theta} \in \mathbb{R}^d$, we study the landscape of $R(\theta)$ near $\widetilde{\theta}$ using a local reparametrization $\varphi = (\varphi^1, \dots, \varphi^L)$ around $\widetilde{\theta}$. Note that, in general, we cannot use the same

reparametrization φ at all points $\widetilde{\theta} \in \mathbb{R}^d$, as we must handle non-generic points where $\mathsf{d}_\theta \varphi(\widetilde{\theta})$ is singular for any particular map φ, even if the true parameter θ_* is generic.

We will combine these local statements over a large enough ball $\{\theta \in \mathbb{R}^d : \|\theta\| \le M\}$ using a compactness argument. The following result strengthens Lemma 2.9 to provide a lower bound for $\|\nabla R(\theta)\|$ outside this ball.

Lemma 4.17. *For some (θ_*, d, G)-dependent constants $M, \rho, c, \sigma_0 > 0$ and all $\sigma > \sigma_0$, if $\|\theta\| > M$, then $\|\nabla R(\theta)\| > c\sigma^{-4}$. If, in addition, $\|\mathbb{E}_g[g\theta] - \mathbb{E}_g[g\theta_*]\| > \rho$, then $\|\nabla R(\theta)\| > c\sigma^{-2}$.*

Proof. First suppose that $\mathbb{E}_g[g] = 0$. Then Lemma 4.9 implies $S_1(\theta) = 0$ and

$$\nabla S_2(\theta) = \mathbb{E}_g[g\theta\theta^\top g^\top]\theta - \mathbb{E}_g[g\theta_*\theta_*^\top g^\top]\theta.$$

For every unit vector $v \in \mathbb{R}^d$, we have $v^\top \mathbb{E}_g[gvv^\top g^\top]v = \mathbb{E}_g[\|v^\top gv\|^2] \ge 1/K$ where $K = |G|$, because $g = \mathrm{Id}$ with probability $1/K$. Thus $\|\mathbb{E}_g[gvv^\top g^\top]v\| \ge 1/K$, so $\|\mathbb{E}_g[g\theta\theta^\top g^\top]\theta\| \ge \|\theta\|^3/K \ge M^3/K$ when $\|\theta\| > M$. Bounding $\|\mathbb{E}_g[g\theta_*\theta_*^\top g^\top]\| \le \|\theta_*\|^2$, we get $\|\nabla S_2(\theta)\| > c$ for any sufficiently large constant M and some constant $c > 0$. If $\|\theta\| > C\sigma^{2/3}$ for a large enough constant $C > 0$, then Lemma 2.9 shows $\|\nabla R(\theta)\| \ge c\sigma^{-2}$, whereas if $C\sigma^{2/3} \ge \|\theta\| \ge M$, then this argument and Lemma 4.7 show $\|\nabla R(\theta)\| \ge c\sigma^{-4}$. This establishes the claim when $\mathbb{E}_g[g] = 0$.

If $\mathbb{E}_g[g] \ne 0$, apply Lemmas 2.3 and 2.4 to write $R(\theta) = R^{\mathrm{Id}}(\theta_1) + R^{G_2}(\theta_2)$, where θ_1 is the component of θ orthogonal to the kernel of $\mathbb{E}_g[g]$. Then $\|\nabla R(\theta)\|^2 = \|\nabla R^{\mathrm{Id}}(\theta_1)\|^2 + \|\nabla R^{G_2}(\theta_2)\|^2$. Recall from the proof of Lemma 2.4 that $\mathbb{E}_g[g]$ is the projection orthogonal to its kernel, so we have $\|\mathbb{E}_g[g\theta] - \mathbb{E}_g[g\theta_*]\| = \|\theta_1 - \theta_{1,*}\|$. Since $R^{\mathrm{Id}}(\theta_1)$ is the risk for the single Gaussian model $\mathcal{N}(\theta_{1,*}, \sigma^2 \mathrm{Id})$, $\nabla R^{\mathrm{Id}}(\theta_1) = (\theta_1 - \theta_{1,*})/\sigma^2$. Thus, if $\|\mathbb{E}_g[g\theta] - \mathbb{E}_g[g\theta_*]\| > \rho$, then Lemmas 2.9 and Lemma 4.7 combine to yield $\|\nabla R(\theta)\| \ge \|\nabla R^{\mathrm{Id}}(\theta_1)\| \ge c\sigma^{-2}$, as above. Otherwise, $\|\theta_2 - \theta_{2,*}\| > M/2$, and applying the above argument for the mean-zero group G_2 shows $\|\nabla R(\theta)\| \ge \|\nabla R^{G_2}(\theta_2)\| \ge c\sigma^{-4}$. □

4.5.1. *Discrete rotations in $\mathbb{R}^2$.* We consider first the group of K-fold discrete rotations on $\mathbb{R}^2$: For a fixed integer K, we have

$$G = \{\mathrm{Id}, h, h^2, \dots, h^{K-1}\} \cong \mathbb{Z}/K\mathbb{Z} \tag{4.32}$$

where

$$h = \begin{pmatrix} \cos 2\pi/K & -\sin 2\pi/K \\ \sin 2\pi/K & \cos 2\pi/K \end{pmatrix} \tag{4.33}$$

is the counterclockwise rotation in the plane by the angle $2\pi/K$. For fixed $\theta_* \ne 0$ and for any $\theta \ne 0$, denote

$$t(\theta) = \arccos \frac{\langle \theta, \theta_* \rangle}{\|\theta\|\|\theta_*\|}$$

as the angle formed by θ and θ_*.

The special case of $K = 2$ and $G = \{+\mathrm{Id}, -\mathrm{Id}\}$ is subsumed by results of [XHM16, Corollary 3], which imply that the global landscape of $R(\theta)$ is benign for all $\sigma > 0$. Thus, we consider here the setting where $K \ge 3$.

Theorem 4.18. *Let G be the group of rotations (4.32) on $\mathbb{R}^2$, with $K \ge 3$. Consider $\theta_* \ne 0$. There exists a (θ_*, K)-dependent constant σ_0 such that for all $\sigma > \sigma_0$, the landscape of $R(\theta)$ is globally benign. More quantitatively, there are (θ_*, K)-dependent constants $\rho, c > 0$ such that when $\sigma > \sigma_0$,*

(a) For each $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$, reparametrizing by $\varphi = (\|\theta\|, t(\theta))$ on $B_\rho(\widetilde{\theta})$, we have the strong convexity $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \ge c\sigma^{-2K}$ for all $\varphi \in \varphi(B_\rho(\widetilde{\theta}))$.

(b) For each $\theta \in \mathbb{R}^d$ satisfying $\|\theta\| - \|\theta_*\| \in (-\rho, \rho)$ and $\theta \notin \bigcup_{\widetilde{\theta} \in \mathcal{O}_{\theta_*}} B_\rho(\widetilde{\theta})$, either $\|\nabla R(\theta)\| \ge c\sigma^{-2K}$ or $\lambda_{\min}(\nabla^2 R(\theta)) \le -c\sigma^{-2K}$.

(c) For each $\theta \in \mathbb{R}^d$ satisfying $\|\theta\| - \|\theta_*\| \notin (-\rho, \rho)$, either $\|\nabla_\theta R(\theta)\| \ge c\sigma^{-4}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta)) \le -c\sigma^{-4}$.

The proof rests on the following lemma, which characterizes the functions $S_\ell(\theta)$ in (4.9) for this discrete rotation group.

Lemma 4.19. *Let G be the group of rotations (4.32) on $\mathbb{R}^2$, with $K \geq 3$. Then*
(a) $S_1(\theta) = 0$ and $S_2(\theta) = \|\theta\|^4/8 - \|\theta\|^2\|\theta_*\|^2/4$.
(b) For each $\ell \in \{3, \dots, K-1\}$, $S_\ell(\theta) = p_\ell(\|\theta\|^2)$ for some univariate polynomial $p_\ell : \mathbb{R} \to \mathbb{R}$ (with coefficients depending on θ_*).
(c) For $\ell = K$ and some polynomial $p_K : \mathbb{R} \to \mathbb{R}$ (with coefficients depending on θ_*),

$$S_K(\theta) = -\frac{1}{2^{K-1}K!}\|\theta\|^K\|\theta_*\|^K \cos(K \cdot t(\theta)) + p_K(\|\theta\|^2).$$

Proof. Let $z = (\theta_*)_1 + \mathbf{i}(\theta_*)_2$ and $w = \theta_1 + \mathbf{i}\theta_2$ as elements of $\mathbb{C}$. Let $\zeta = e^{2\pi\mathbf{i}/K}$, and denote the set of K^{th} roots of unity by $X_K = \{1, \zeta, \dots, \zeta^{K-1}\}$. Then $\zeta^k z = (h^k\theta_*)_1 + \mathbf{i}(h^k\theta_*)_2$ where h is the generator (4.33), and similarly for w and θ. Notice that for $a = a_1 + \mathbf{i}a_2$ and $b = b_1 + \mathbf{i}b_2$ we have

$$\langle (a_1, a_2), (b_1, b_2)\rangle = a_1b_1 + a_2b_2 = \frac{1}{4}\Big((a+\bar{a})(b+\bar{b}) - (a-\bar{a})(b-\bar{b})\Big) = \frac{1}{2}\Big(a\bar{b} + \bar{a}b\Big).$$

Then

$$\mathbb{E}_g[\langle\theta_*, g\theta\rangle^2] = \frac{1}{4K}\sum_{\zeta\in X_K}\left(\zeta^{-1}z\bar{w} + \zeta\bar{z}w\right)^2 = \frac{|z|^2|w|^2}{2} = \frac{\|\theta\|^2\|\theta_*\|^2}{2},$$

where we have used $K \geq 3$ and

$$\sum_{\zeta\in X_K}\zeta^a = \begin{cases} K & \text{if } a \equiv 0 \mod K \\ 0 & \text{if } a \not\equiv 0 \mod K \end{cases} \tag{4.34}$$

for the second equality. Similarly $\mathbb{E}_g[\langle\theta, g\theta\rangle^2] = \|\theta\|^4/2$, and (a) follows from Lemma 4.9.

Applying this argument for a general term $M_{\ell,m}(\pi \mid \theta, \theta_*)$, we have

$$\begin{aligned}
&M_{\ell,m}(\pi \mid \theta, \theta_*) \\
&= \mathbb{E}_{g_1,\dots,g_{|\pi|}}\left[\prod_{j=1}^{m}\Big\langle g_{\pi(2j-1)}\theta,\ g_{\pi(2j)}\theta\Big\rangle \cdot \prod_{j=2m+1}^{\ell+m}\Big\langle\theta_*,\ g_{\pi(j)}\theta\Big\rangle\right] \\
&= \frac{1}{2^\ell K^{|\pi|}}\sum_{i_1,\dots,i_{|\pi|}=0}^{K-1}\left[\prod_{j=1}^{m}\Big((\zeta^{i_{\pi(2j-1)}-i_{\pi(2j)}} + \zeta^{i_{\pi(2j)}-i_{\pi(2j-1)}})|w|^2\Big)\prod_{j=2m+1}^{\ell+m}(\zeta^{-i_{\pi(j)}}z\bar{w} + \zeta^{i_{\pi(j)}}\bar{z}w)\right] \\
&= \frac{|w|^{2m}}{2^\ell K^{|\pi|}}\sum_{\zeta_1,\dots,\zeta_{|\pi|}\in X_K}\left[\prod_{j=1}^{m}(\zeta_{\pi(2j-1)}/\zeta_{\pi(2j)} + \zeta_{\pi(2j)}/\zeta_{\pi(2j-1)})\prod_{j=2m+1}^{\ell+m}(\zeta_{\pi(j)}^{-1}z\bar{w} + \zeta_{\pi(j)}\bar{z}w)\right].
\end{aligned}$$

Expanding into polynomials of $z, \bar{z}, w, \bar{w}$, this expression is a linear combination with constant coefficients of terms of the form

$$\sum_{\zeta_1,\dots,\zeta_{|\pi|}\in X_K}|w|^{2m+2a}|z|^{2a}w^b\bar{z}^b\zeta_1^{c_1}\cdots\zeta_{|\pi|}^{c_{|\pi|}}.$$

Here, the exponents satisfy $a \geq 0$, $2a + |b| = (\ell+m) - 2m = \ell - m$, $\sum_i c_i = b$, $\sum_i |c_i| \leq \ell + m$, and $|c_i| \leq m + (\ell + m - 2m) = \ell$ for each i. By (4.34), these terms vanish unless each c_i is a multiple of K. In particular, for $\ell < K$, the condition $|c_i| \leq \ell$ implies that the only non-zero terms must have $c_1 = \dots = c_{|\pi|} = b = 0$. Then $M_{\ell,m}(\pi \mid \theta, \theta_*)$ is a polynomial in $|w|^2 = \|\theta\|^2$. Since $S_\ell(\theta)$ is a linear combination of such terms $M_{\ell,m}(\pi \mid \theta, \theta_*)$, this shows (b).

For (c), if $\ell = K$, the only non-zero terms which are not a polynomial of $\|\theta\|^2$ must have $b \neq 0$, so that the condition $2a + |b| = K - m$ requires $m < K$. Then since $\sum_i |c_i| \leq K + m < 2K$,

there is some i^* with $c_{i^*} \in \{-K, K\}$ and $c_j = 0$ for all $j \neq i^*$. Such terms can only appear in $M_{K,m}(\pi \mid \theta, \theta_*)$ when $m = 0$ and $\pi = \{\{1, \dots, K\}\}$, for which we have

$$M_{K,0}(\{\{1, \dots, K\}\} \mid \theta, \theta_*) = \frac{1}{2^K}(z^K \bar{w}^K + \bar{z}^K w^K).$$

Writing $w = \|\theta\| e^{ir}$ and $z = \|\theta_*\| e^{ir_*}$, this is

$$M_{K,0}(\{\{1, \dots, K\}\} \mid \theta, \theta_*) = \frac{\|\theta\|^K \|\theta_*\|^K}{2^K}(e^{i(r-r_*)K} + e^{i(r_*-r)K}) = \frac{\|\theta\|^K \|\theta_*\|^K}{2^{K-1}} \cos(Kt(\theta)).$$

Substituting into (4.9) and recalling that the remaining terms are polynomial in $\|\theta\|^2$ shows (c). □

Proof of Theorem 4.18. For each point $\widetilde{\theta} \in \mathbb{R}^d$, we consider a local reparametrization by φ in a neighborhood $U_{\widetilde{\theta}}$ of $\widetilde{\theta}$. At $\widetilde{\theta} = 0$, we take the reparametrization to be $\varphi = \theta$. At each $\widetilde{\theta} \neq 0$, we take it to be $\varphi = (\|\theta\|, t(\theta))$. We then apply Lemmas 4.10 and 4.13 on $U_{\widetilde{\theta}}$.

For $\widetilde{\theta} = 0$, observe that $\nabla^2_\theta S_2(\widetilde{\theta}) = -\frac{1}{2}\|\theta_*\|^2 \operatorname{Id} \prec 0$. For $\widetilde{\theta} \neq 0$ where $\|\widetilde{\theta}\| \neq \|\theta_*\|$, set $\widetilde{\varphi} = \varphi(\widetilde{\theta})$. Observe that $S_2(\varphi) = \varphi_1^4/8 - \varphi_1^2\varphi_{1,*}^2/4$, so $\nabla_{\varphi_1} S_2(\widetilde{\varphi}) = \frac{1}{2}\widetilde{\varphi}_1(\widetilde{\varphi}_1^2 - \widetilde{\varphi}_{1,*}^2) \neq 0$. In both cases, Lemma 4.10 implies that $\|\nabla_\theta R(\theta)\| \geq c\sigma^{-4}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta)) \leq -c\sigma^{-4}$, for some $c, \sigma_0 > 0$ and all $\sigma > \sigma_0$ and $\theta \in U_{\widetilde{\theta}}$.

For $\widetilde{\theta} \notin \mathcal{O}_{\theta_*}$ where $\|\widetilde{\theta}\| = \|\theta_*\|$, observe that $S_1, \dots, S_{K-1}$ depend only on φ_1. For S_K, applying $\widetilde{\varphi}_1 = \varphi_{1,*}$, we have

$$\nabla_{\varphi_2} S_K(\widetilde{\varphi}) = \frac{1}{2^{K-1}(K-1)!}\varphi_{1,*}^{2K} \sin(K\widetilde{\varphi}_2), \qquad \nabla^2_{\varphi_2} S_K(\widetilde{\varphi}) = \frac{K}{2^{K-1}(K-1)!}\varphi_{1,*}^{2K} \cos(K\widetilde{\varphi}_2). \quad (4.35)$$

Then either $\nabla_{\varphi_2} S_K(\widetilde{\varphi}) \neq 0$ (when $\widetilde{\varphi}_2 \notin \{j\pi/K : j = 0, 1, \dots, 2K-1\}$), or $\lambda_{\min}(\nabla^2_{\varphi_2} S_K(\widetilde{\varphi})) < 0$ (when $\widetilde{\varphi}_2 \in \{j\pi/K : j = 1, 3, 5, \dots, 2K-1\}$). So Lemma 4.10 implies that $\|\nabla_\theta R(\theta)\| \geq c\sigma^{-2K}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta)) \leq -c\sigma^{-2K}$ for all $\sigma > \sigma_0$ and $\theta \in U_{\widetilde{\theta}}$.

Finally, for $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$, (4.35) verifies that $\widetilde{\varphi} = \varphi(\widetilde{\theta})$ is a pseudo-local-minimizer in the parametrization by φ. Then Lemma 4.13 implies $R(\theta)$ has a unique local minimizer in $U_{\widetilde{\theta}}$ and $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \geq c\sigma^{-2K}$ for all $\varphi \in \varphi(U_{\widetilde{\theta}})$ and $\sigma > \sigma_0$. This unique local minimizer must be $\widetilde{\theta}$ itself, since $\widetilde{\theta}$ is a global minimizer of $R(\theta)$.

The constants $c, \sigma_0 > 0$ above depend on $\widetilde{\theta}$. By compactness, for any $M > 0$, there is a finite collection of points $\widetilde{\theta}$ where the neighborhoods $U_{\widetilde{\theta}}$ cover $\{\theta \in \mathbb{R}^2 : \|\theta\| \leq M\}$, and the above statements then hold for uniform choices of $c, \sigma_0 > 0$ in their union. For a sufficiently small constant $\rho > 0$, this establishes all claims of the theorem for points $\theta \in \mathbb{R}^d$ where $\|\theta\| \leq M$, and the result for $\|\theta\| > M$ follows from Lemma 4.17. □

The following then shows that with high probability for $n \gg \sigma^{4K-2} \log \sigma$, the empirical risk $R_n(\theta)$ is also globally benign and satisfies the same properties.

Corollary 4.20. *For some (θ_*, K)-dependent constants $C, c > 0$, the statements of Theorem 4.18 hold also for $R_n(\theta)$, with probability at least $1 - \sigma^C e^{-c\sigma^{-4K+2}n} - Ce^{-cn^{2/3}}$.*

Proof. We first apply Lemma 2.10 with $t = c_0\sigma^{-2K}$ and bounded radius $r = \|\theta_*\| + \rho$. For this r, we have $\sup_{\theta \in B_r} |R_n(\theta) - R(\theta)|, \|\nabla R_n(\theta) - \nabla R(\theta)\|, \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \leq c_0\sigma^{-2K}$ with probability at least $1 - \sigma^C e^{-c\sigma^{-4K+2}n} - Ce^{-cn^{2/3}}$. We next apply Lemma 2.10 with $t = c_0\sigma^{-4}$ and radius $r = C_0\sigma^2$. For this r, we have $\sup_{\theta \in B_r} \|\nabla R_n(\theta) - \nabla R(\theta)\|, \|\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\| \leq c_0\sigma^{-4}$ with probability at least $1 - \sigma^C e^{-c\sigma^{-6}n} - Ce^{-cn^{2/3}}$. Applying these for sufficiently small c_0 and combining with Theorem 4.18, Lemma 2.8, and Lemma 2.7 yields the corollary. □

4.5.2. *All permutations in $\mathbb{R}^d$.* Consider any dimension $d \geq 1$, and let $G \cong S_d$ be the symmetric group of all permutations of coordinates in $\mathbb{R}^d$. Here, the size of the group is $K = d!$. Define the symmetric power sums in θ by

$$p_k(\theta) = \frac{1}{d}\sum_{j=1}^{d} \theta_j^k,$$

and (for fixed $\theta_* \in \mathbb{R}^d$) the Vandermonde varieties by

$$\mathcal{V}_k = \Big\{\theta \in \mathbb{R}^d : p_\ell(\theta) = p_\ell(\theta_*) \text{ for all } \ell = 1, \dots, k\Big\}. \tag{4.36}$$

Note that the map $\theta \mapsto (p_1(\theta), \dots, p_d(\theta))$ is injective on $\{\theta \in \mathbb{R}^d : \theta_1 \leq \dots \leq \theta_d\}$ (see [Kos89, Corollary 1.2]), so $\mathcal{V}_d = \mathcal{O}_{\theta_*}$.

Theorem 4.21. *Let $G \cong S_d$ be the symmetric group acting on $\mathbb{R}^d$ by permutation of coordinates. For generic $\theta_* \in \mathbb{R}^d$, there exists a (θ_*, d)-dependent constant $\sigma_0 > 0$ such that the global landscape of $R(\theta)$ is benign for all $\sigma > \sigma_0$. More quantitatively, there are (θ_*, d)-dependent constants $\rho, c > 0$ such that when $\sigma > \sigma_0$,*

(a) For each $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$, reparametrizing by the symmetric power sums $\varphi = (p_1, \dots, p_d)$ in $B_\rho(\widetilde{\theta})$, we have the strong convexity $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \geq c\sigma^{-2d}$ for all $\varphi \in \varphi(B_\rho(\widetilde{\theta}))$.

(b) Denote $\mathcal{V}_\ell^\rho = \{\theta \in \mathbb{R}^d : \operatorname{dist}(\theta, \mathcal{V}_\ell) < \rho\}$, where $\mathcal{V}_0^\rho = \mathbb{R}^d$. Then for each $\ell = 1, \dots, d$ and each $\theta \in \mathcal{V}_{\ell-1}^\rho \setminus \mathcal{V}_\ell^\rho$, either $\|\nabla_\theta R(\theta)\| \geq c\sigma^{-2\ell}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta)) \leq -c\sigma^{-2\ell}$.

The proof rests on the following lemma, which characterizes the functions $S_\ell(\theta)$ in this example.

Lemma 4.22. *Let $G \cong S_d$ be the symmetric group acting on $\mathbb{R}^d$ by permutation of coordinates. For each $\ell = 1, \dots, d$, some constant $a_\ell > 0$, and some polynomials $q_\ell, r_\ell : \mathbb{R}^{\ell-1} \to \mathbb{R}$ with coefficients depending on θ_* and such that*

$$q_\ell\Big(p_1(\theta_*), \dots, p_{\ell-1}(\theta_*)\Big) = 0,$$

we have

$$S_\ell(\theta) = a_\ell\big(p_\ell(\theta)^2 - p_\ell(\theta_*)\big)^2 + q_\ell\big(p_1(\theta), \dots, p_{\ell-1}(\theta)\big) \cdot p_\ell(\theta) + r_\ell\big(p_1(\theta), \dots, p_{\ell-1}(\theta)\big). \tag{4.37}$$

Proof. We apply Lemma 4.8 and the fact that the symmetric power sums $p_1(\theta), p_2(\theta), \dots$ generate $\mathcal{R}^G$ as an algebra over $\mathbb{R}$ (see [Mac15, Eq. (2.12)]). Thus, any polynomial $\varphi \in \mathcal{R}^G_{\leq \ell}$ may be written as

$$\varphi(\theta) = c_\varphi p_\ell(\theta) + q_\varphi(p_1(\theta), \dots, p_{\ell-1}(\theta))$$

for some $c_\varphi \in \mathbb{R}$ and some polynomial q_φ with real coefficients. In particular, applying this to each entry of the moment tensor $T_\ell(\theta)$ in Lemma 4.8, we obtain the form (4.37) where

$$a_\ell = \sum_\varphi \frac{c_\varphi^2}{2(\ell!)}$$

and

$$q_\ell\big(p_1(\theta), \dots, p_{\ell-1}(\theta)\big) = \sum_\varphi \frac{c_\varphi}{\ell!} \cdot \Big(q_\varphi\big(p_1(\theta), \dots, p_{\ell-1}(\theta)\big) - q_\varphi\big(p_1(\theta_*), \dots, p_{\ell-1}(\theta_*)\big)\Big),$$

with both summations taken over all entries of $T_\ell(\theta)$. We have $a_\ell > 0$ strictly because the diagonal entries of $T_\ell(\theta)$ are given by

$$T_\ell(\theta)_{i,\dots,i} = \frac{1}{d!}\sum_{\sigma \in S_d} \theta^\ell_{\sigma(i)} = \frac{(d-1)!}{d!}\sum_{i=1}^{d} \theta_i^\ell = p_\ell(\theta),$$

so that $c_\varphi = 1$ for these entries. ∎

The derivative of this map $\varphi = (p_1, \dots, p_d)$ is singular at points $\widetilde{\theta}$ having repeated entries. To analyze the landscape of $R(\theta)$ near such points, we use the following known (and non-trivial) facts about the symmetric power sums and Vandermonde varieties.

Lemma 4.23. *Let $\mathcal{V}_k$ be the Vandermonde variety (4.36), with $\mathcal{V}_0 = \mathbb{R}^d$. For each $k \in \{1, \dots, d\}$ and any generic $\theta_* \in \mathbb{R}^d$,*

(a) Each point $\theta \in \mathcal{V}_k$ has at least k distinct entries.
(b) $\mathcal{V}_{k-1}$ is a nonsingular algebraic variety, and $p_k(\theta)$ is a Morse function on $\mathcal{V}_{k-1}$.
(c) The critical points of the restriction $p_k|_{\mathcal{V}_{k-1}}$ are the points $\theta \in \mathcal{V}_{k-1}$ having exactly $k-1$ distinct entries.
(d) If θ is a local minimizer or local maximizer of $p_k|_{\mathcal{V}_{k-1}}$, then it is also a global minimizer or global maximizer of $p_k|_{\mathcal{V}_{k-1}}$.

Proof. For (a), fixing any integer multiplicities $d_1, \dots, d_{k-1} \ge 0$ summing to d, the image of the polynomial function $F : \mathbb{R}^{k-1} \to \mathbb{R}^k$ given by

$$F(x_1, \dots, x_{k-1}) = \left(\frac{1}{d} \sum_{j=1}^{k-1} d_j x_j^\ell : \ell = 1, \dots, k \right)$$

is a constructible set in the Zariski topology on $\mathbb{R}^k$, by Chevalley's theorem (see [Har92, Theorem 3.16]). By [Har92, Theorem 11.12], the Zariski closure of this image has dimension at most $k-1$, so its complement is generic. Taking the intersection of these complements over the finitely many choices of $d_1, \dots, d_{k-1}$, we find that the complement of the set

$$\Big\{ (p_1(\theta), \dots, p_k(\theta)) : \theta \text{ has at most } k-1 \text{ distinct coordinates} \Big\}$$

is also generic in $\mathbb{R}^k$. We conclude that for generic $\theta_* \in \mathbb{R}^d$, the point $(p_1(\theta_*), \dots, p_k(\theta_*))$ does not belong to the above set, meaning that each point $\theta \in \mathcal{V}_k$ has at least k distinct coordinates.

For (b), the gradient of p_ℓ is given by

$$\nabla p_\ell(\theta) = \frac{\ell}{d} (\theta_1^{\ell-1}, \dots, \theta_d^{\ell-1}).$$

Thus, if $\nabla p_1, \dots, \nabla p_k$ are linearly dependent, then there is a non-zero polynomial P of degree at most $k-1$ for which $P(\theta_i) = 0$ for every $i = 1, \dots, d$. Since P has at most $k-1$ real roots, this implies that θ has at most $k-1$ distinct coordinates. Applying (a), this shows that for generic θ_*, the vectors $\nabla p_1, \dots, \nabla p_k$ are linearly independent at every $\theta \in \mathcal{V}_k(\theta_*)$, so $\mathcal{V}_k(\theta_*)$ is nonsingular. The remaining two statements then follow from the results of [Arn86, Theorems 5, 6, and 7]; see also [Kos89]. □

Proof of Theorem 4.21. For each $\widetilde{\theta} \in \mathbb{R}^d$, we consider a local reparametrization by φ in a neighborhood $U_{\widetilde{\theta}}$. If k is the number of distinct entries of $\widetilde{\theta}$, then we take the first k functions in φ to be the symmetric power sums $p_1(\theta), \dots, p_k(\theta)$. As shown in the proof of Lemma 4.23 above, the gradients $\nabla p_1, \dots, \nabla p_k$ must be linearly independent at $\widetilde{\theta}$. We arbitrarily pick $d-k$ remaining functions to complete $(p_1, \dots, p_k)$ into the local reparametrization φ. Denote $\widetilde{\varphi} = \varphi(\widetilde{\theta})$ and $\varphi_* = \varphi(\theta_*)$.

We apply Lemmas 4.10 and 4.13 on each neighborhood $U_{\widetilde{\theta}}$. Fix $\ell \in \{1, \dots, d\}$ and consider $\widetilde{\theta} \in \mathcal{V}_{\ell-1} \setminus \mathcal{V}_\ell$. By Lemma 4.23(a), $\widetilde{\theta}$ has at least $\ell - 1$ distinct coordinates, so the first $\ell - 1$ coordinates of φ are $(\varphi_1, \dots, \varphi_{\ell-1}) = (p_1, \dots, p_{\ell-1})$. Denote $\varphi^\ell = (\varphi_\ell, \dots, \varphi_d)$, and note that $S_1, \dots, S_{\ell-1}$ are functions only of $\varphi_1, \dots, \varphi_{\ell-1}$. Furthermore, recalling (4.37) and applying

$$q_\ell(\widetilde{\varphi}_1, \dots, \widetilde{\varphi}_{\ell-1}) = q_\ell(\varphi_{1,*}, \dots, \varphi_{\ell-1,*}) = 0$$

and the chain rule,

$$\nabla_{\varphi^\ell} S_\ell(\widetilde{\varphi}) = 2a_\ell \Big(p_\ell(\widetilde{\varphi}) - p_\ell(\varphi_*) \Big) \nabla_{\varphi^\ell} p_\ell(\widetilde{\varphi}),$$

$$\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi}) = 2a_\ell \Big(\nabla_{\varphi^\ell} p_\ell(\widetilde{\varphi}) \nabla_{\varphi^\ell} p_\ell(\widetilde{\varphi})^\top + (p_\ell(\widetilde{\varphi}) - p_\ell(\varphi_*)) \cdot \nabla^2_{\varphi^\ell} p_\ell(\widetilde{\varphi}) \Big).$$

Since $\widetilde{\theta} \notin \mathcal{V}_\ell$, we have $p_\ell(\widetilde{\varphi}) \neq p_\ell(\varphi_*)$. Then either $\nabla_{\varphi^\ell} S_\ell(\widetilde{\varphi}) \neq 0$, or

$$\nabla_{\varphi^\ell} p_\ell(\widetilde{\varphi}) = 0 \qquad \text{and} \qquad \nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi}) = 2a_\ell (p_\ell(\widetilde{\varphi}) - p_\ell(\varphi_*)) \cdot \nabla^2_{\varphi^\ell} p_\ell(\widetilde{\varphi}).$$

In this latter case, note that φ^ℓ is a local chart for $\mathcal{V}_{\ell-1}$ around $\widetilde{\varphi}$, so $\widetilde{\varphi}$ is a critical point of $p_\ell|_{\mathcal{V}_{\ell-1}}$. The Morse condition of Lemma 4.23(b) implies that all eigenvalues of $\nabla^2_{\varphi^\ell} p_\ell(\widetilde{\varphi})$ are non-zero. If $\nabla^2_{\varphi^\ell} p_\ell(\widetilde{\varphi})$ has both positive and negative eigenvalues, then this guarantees that $\lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi})) < 0$. Otherwise, $\widetilde{\varphi}$ is a local minimizer or local maximizer of $p_\ell|_{\mathcal{V}_{\ell-1}}$. If it is a local minimizer, then all eigenvalues of $\nabla^2_{\varphi^\ell} p_\ell(\widetilde{\varphi})$ are positive. Lemma 4.23(d) also implies that $p_\ell(\widetilde{\varphi}) < p_\ell(\varphi_*)$, so all eigenvalues of $\nabla^2_{\varphi^\ell} S_\ell(\widetilde{\varphi})$ are negative. The case where $\widetilde{\varphi}$ is a local maximizer of $p_\ell|_{\mathcal{V}_{\ell-1}}$ is similar. Combining these observations and applying Lemma 4.10, we get that either $\|\nabla_\theta R(\theta)\| \geq c\sigma^{-2\ell}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta)) \leq -c\sigma^{-2\ell}$ for all $\theta \in U_{\widetilde{\theta}}$ and $\sigma > \sigma_0$.

For $\widetilde{\theta} \in \mathcal{V}_d = \mathcal{O}_{\theta_*}$, we have $\varphi = (p_1, \dots, p_d)$, so that each S_ℓ depends only on $(\varphi_1, \dots, \varphi_\ell)$ and

$$\nabla_{\varphi_\ell} S_\ell(\widetilde{\varphi}) = 2a_\ell(\widetilde{\varphi}_\ell - \varphi_{\ell,*}) = 0, \qquad \nabla^2_{\varphi_\ell} S_\ell(\widetilde{\varphi}) = 2a_\ell > 0.$$

Thus $\widetilde{\theta}$ is a pseudo-local-minimizer in the reparametrization by φ. Lemma 4.13 implies that $\widetilde{\theta}$ is the unique critical point of $R(\theta)$ in $U_{\widetilde{\theta}}$, and that $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \geq c\sigma^{-2d}$ for all $\varphi \in \varphi(U_{\widetilde{\theta}})$ and $\sigma > \sigma_0$.

Fixing any $M > 0$ and taking a finite collection of these sets $U_{\widetilde{\theta}}$ which cover the compact set $\{\theta \in \mathbb{R}^d : \|\theta\| \leq M\}$, the above results hold for uniform choices of constants $c, \sigma_0 > 0$ in their union. Then for a sufficiently small constant $\rho > 0$, the claims of the theorem hold for all $\theta \in \mathbb{R}^d$ with $\|\theta\| \leq M$, and the result for $\|\theta\| > M$ follows again from Lemma 4.17. □

The following then shows that with high probability for $n \gg \sigma^{4d-2} \log \sigma$, the empirical risk $R_n(\theta)$ is also globally benign and satisfies the same properties. The proof is the same as that of Corollary 4.20, and we omit this for brevity.

Corollary 4.24. *For some (θ_*, d)-dependent constants $C, c > 0$, the statements of Theorem 4.21 hold also for $R_n(\theta)$, with probability at least $1 - \sigma^C e^{-c\sigma^{-4d+2}n} - Ce^{-cn^{2/3}}$.*

4.5.3. *General groups.* We provide a general condition under which the landscape of $R(\theta)$ is globally benign for high noise, which captures the structure of the previous two examples.

Let $M_\ell : \mathbb{R}^d \to \mathbb{R}^{d+d^2+\dots+d^\ell}$ be the combined vectorized moment map

$$M_\ell(\theta) = \Big(T_1(\theta), \dots, T_\ell(\theta) \Big).$$

For fixed $\theta_* \in \mathbb{R}^d$, recall $P_\ell(\theta) = \|T_\ell(\theta) - T_\ell(\theta_*)\|^2_{\mathrm{HS}}$ from Lemma 4.8, and define the moment varieties

$$\mathcal{V}_\ell = \Big\{ \theta \in \mathbb{R}^d : M_\ell(\theta) = M_\ell(\theta_*) \Big\}, \qquad \mathcal{V}_0 = \mathbb{R}^d.$$

We denote by $P_\ell|_{\mathcal{V}_{\ell-1}}$ the restriction of the function P_ℓ to $\mathcal{V}_{\ell-1}$. We will assume that each $\mathcal{V}_\ell$ is nonsingular and has the same dimension $\bar{d}_\ell$ at every point. We then denote by $\nabla P_\ell|_{\mathcal{V}_{\ell-1}} \in \mathbb{R}^{\bar{d}_\ell}$ and $\nabla^2 P_\ell|_{\mathcal{V}_{\ell-1}} \in \mathbb{R}^{\bar{d}_\ell \times \bar{d}_\ell}$ the gradient and Hessian of the restriction $P_\ell|_{\mathcal{V}_{\ell-1}}$ with respect to any choice of local chart on $\mathcal{V}_{\ell-1}$. Note that the conditions below do not depend on the specific choice of chart.

Theorem 4.25. *Let $\theta_* \in \mathbb{R}^d$ be generic, and let L be the constant in Lemma 4.3. Suppose that*

$$\mathcal{V}_L = \mathcal{O}_{\theta_*}$$

and that for every $\ell \geq 1$, $\mathsf{d}_\theta M_\ell$ has constant rank on $\mathcal{V}_\ell$. Suppose also, for each $\ell = 1, \dots, L$ and each $\theta \in \mathcal{V}_{\ell-1}$, that either (1) $\nabla P_\ell|_{\mathcal{V}_{\ell-1}}(\theta) \neq 0$, (2) $\lambda_{\min}(\nabla^2 P_\ell|_{\mathcal{V}_{\ell-1}}(\theta)) < 0$, or (3) $\theta \in \mathcal{V}_\ell$. Then

there exists a (θ_*, d, G)-dependent constant $\sigma_0 > 0$ such that the landscape of $R(\theta)$ is globally benign for all $\sigma > \sigma_0$.

More quantitatively, there exist (θ_*, d, G)-dependent constants $c, \rho > 0$ such that when $\sigma > \sigma_0$,

(a) For each $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$, there is a local reparametrization $\varphi : B_\rho(\widetilde{\theta}) \to \mathbb{R}^d$ such that $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \geq c\sigma^{-2L}$ for all $\varphi \in \varphi(B_\rho(\widetilde{\theta}))$.

(b) Denote $\mathcal{V}^\rho_\ell = \{\theta \in \mathbb{R}^d : \mathrm{dist}(\theta, \mathcal{V}_\ell) < \rho\}$. Then for each $\ell = 1, \ldots, L$ and each $\theta \in \mathcal{V}^\rho_{\ell-1} \setminus \mathcal{V}^\rho_\ell$, either $\|\nabla R(\theta)\| \geq c\sigma^{-2\ell}$ or $\lambda_{\min}(\nabla^2 R(\theta)) \leq -c\sigma^{-2\ell}$.

With probability at least $1 - \sigma^C e^{-c'\sigma^{-4L+2}n} - Ce^{-c'n^{2/3}}$, the same statements hold for the empirical risk $R_n(\theta)$.

Proof. For generic $\theta_* \in \mathbb{R}^d$, Lemma 4.5 implies that $\mathsf{d}_\theta M_\ell$ has rank $d_1 + \ldots + d_\ell$ at θ_*. Then by the given assumption that $\mathsf{d}_\theta M_\ell$ has constant rank over $\mathcal{V}_\ell$, this rank must be $d_1 + \ldots + d_\ell$, and $\mathcal{V}_\ell$ is a manifold of dimension $\bar{d}_\ell = d - (d_1 + \ldots + d_\ell)$.

For any point $\widetilde{\theta} \notin \mathcal{O}_{\theta_*}$, there must exist $\ell \in \{1, \ldots, L\}$ where $\widetilde{\theta} \in \mathcal{V}_0, \ldots, \mathcal{V}_{\ell-1}$ and $\widetilde{\theta} \notin \mathcal{V}_\ell$. For each $k = 1, \ldots, \ell - 1$, since $\mathsf{d}_\theta M_k$ has rank $d_1 + \ldots + d_k$, we may pick d_k coordinates φ^k of the moment tensor T_k such that $(\varphi^1, \ldots, \varphi^{\ell-1})$ have linearly independent gradients at $\widetilde{\theta}$. Let us complete the parametrization by $d - (d_1 + \ldots + d_{\ell-1})$ additional coordinates φ^ℓ, so that $\varphi = (\varphi^1, \ldots, \varphi^\ell)$ has non-singular derivative at $\widetilde{\theta}$. Then for some neighborhood $U_{\widetilde{\theta}}$ of $\widetilde{\theta}$, φ forms a local reparametrization on $U_{\widetilde{\theta}}$, and Lemma 4.5(c) ensures that each polynomial $\psi \in \mathcal{R}^G_{\leq \ell-1}$ is a function only of $(\varphi^1, \ldots, \varphi^{\ell-1})$ in this reparametrization. In particular, the manifold $\mathcal{V}_{\ell-1}$ is defined by $\varphi^1(\theta) = \varphi(\theta_*), \ldots, \varphi^{\ell-1}(\theta) = \varphi^{\ell-1}(\theta_*)$ on $U_{\widetilde{\theta}}$, so that the remaining coordinates φ^ℓ form a local chart for $\mathcal{V}_{\ell-1}$. By Lemma 4.8, $S_1, \ldots, S_{\ell-1}$ are functions only of $(\varphi^1, \ldots, \varphi^{\ell-1})$, and

$$\nabla_{\varphi^\ell} S_\ell(\varphi) = \frac{1}{2(\ell!)} \nabla_{\varphi^\ell} P_\ell(\varphi), \qquad \nabla^2_{\varphi^\ell} S_\ell(\varphi) = \frac{1}{2(\ell!)} \nabla^2_{\varphi^\ell} P_\ell(\varphi).$$

Since $\widetilde{\theta} \notin \mathcal{V}_\ell$, the given condition in the lemma implies that either $\nabla_{\varphi^\ell} S_\ell(\varphi) \neq 0$ or $\lambda_{\min}(\nabla^2_{\varphi^\ell} S_\ell(\varphi)) < 0$. Then by Lemma 4.10, for $\sigma > \sigma_0$ and large enough σ_0, there is a neighborhood $U_{\widetilde{\theta}}$ of $\widetilde{\theta}$ on which either $\|\nabla R(\theta)\| \geq c\sigma^{-2\ell}$ or either $\lambda_{\min}(\nabla^2 R(\theta)) \leq -c\sigma^{-2\ell}$.

If $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$, then Theorem 4.14 shows that there is a neighborhood $U_{\widetilde{\theta}}$ where, parametrizing by the full transcendence basis φ of Lemma 4.3, we have $\nabla^2_\varphi R(\varphi) \geq c\sigma^{-2L}$ on $\varphi(U_{\widetilde{\theta}})$. Taking a finite collection of these neighborhoods $U_{\widetilde{\theta}}$ which cover the compact set $\{\theta \in \mathbb{R}^d : \|\theta\| \leq M\}$, these constants $c, \sigma_0 > 0$ are uniform over this set. Then for small enough $\rho > 0$, this establishes the claims of the theorem for $R(\theta)$ and $\|\theta\| \leq M$. The claims for $\|\theta\| > M$ follow from Lemma 4.17, and the statements for the empirical risk follow from the same proof as Corollary 4.20. □

4.6. **Global landscape for cyclic permutations in $\mathbb{R}^d$.** Consider the group of cyclic permutations of coordinates in dimension d. We have

$$G = \{\mathrm{Id}, h, h^2, \ldots, h^{d-1}\} \cong \mathbb{Z}/d\mathbb{Z} \tag{4.38}$$

where the generator

$$h = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{d \times d} \tag{4.39}$$

cyclically rotates coordinates by one position. Here, the size of the group is $K = d$. Since this is the same as the group of all permutations when $d \in \{1, 2\}$, we consider $d \geq 3$.

We change to the Fourier basis: Index $\mathbb{R}^d$ and $\mathbb{C}^d$ by $0, 1, \dots, d-1$ and define the d^{th} root-of-unity $\omega = e^{2\pi \mathbf{i}/d}$. For all $k \in \mathbb{Z}$, let

$$v_k(\theta) = \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} \omega^{jk} \theta_j \tag{4.40}$$

be the coordinates of the normalized Fourier transform of θ. Note that $v_0(\theta)$ is real, and $v_{d/2}(\theta)$ is also real for even d. Denote

$$r_k(\theta) = |v_k(\theta)|, \qquad v_{k,*} = v_k(\theta_*), \qquad r_{k,*} = r_k(\theta_*).$$

Let us set

$$\mathcal{I} = \{1, \dots, \lfloor \tfrac{d-1}{2} \rfloor\},$$

denote the unit circle as

$$\mathcal{S} \cong [0, 2\pi),$$

and write $\mathrm{Arg}(z) \in \mathcal{S}$ for the complex argument of $z \in \mathbb{C}$. Identifying $t_{-i} = -t_i$ for $i \in \mathcal{I}$ and $t_i \in \mathcal{S}$, we consider the two surrogate functions $F^+ : \mathcal{S}^{|\mathcal{I}|} \to \mathbb{R}$ and $F^- : \mathcal{S}^{|\mathcal{I}|} \to \mathbb{R}$ defined by

$$\begin{aligned} F^{\pm}(t_1, \dots, t_{|\mathcal{I}|}) = -\Bigg(& \frac{1}{6} \sum_{\substack{i,j,k \in \mathcal{I} \cup -\mathcal{I} \\ i+j+k \equiv 0 \bmod d}} r_{i,*}^2 r_{j,*}^2 r_{k,*}^2 \cos(t_i + t_j + t_k) \\ & \pm \mathbf{1}\{d \text{ is even}\} \cdot \frac{1}{2} \sum_{\substack{i,j \in \mathcal{I} \cup -\mathcal{I} \\ i+j \equiv d/2 \bmod d}} r_{i,*}^2 r_{j,*}^2 r_{d/2,*}^2 \cos(t_i + t_j) \Bigg). \end{aligned} \tag{4.41}$$

We have $F^+ = F^-$ when d is odd, and in this case we will only refer to F^+.

For generic $\theta_* \in \mathbb{R}^d$ and $\sigma > \sigma_0 \equiv \sigma_0(\theta_*, d)$, the following shows that local minimizers of $R(\theta)$ are in correspondence with local minimizers of these surrogate functions on the manifold $\mathcal{S}^{|\mathcal{I}|}$.

Theorem 4.26. *Let G be the cyclic group (4.38) acting on $\mathbb{R}^d$, where $d \geq 3$. Suppose $\theta_* \in \mathbb{R}^d$ has $v_{k,*} \neq 0$ for all $k \not\equiv 0 \bmod d$. Then for some (θ_*, d)-dependent constants $c, \rho, \sigma_0 > 0$ and all $\sigma > \sigma_0$:*

(a) For each local minimizer $\tilde{t}$ of $F^+(t)$ where $\lambda_{\min}(\nabla^2 F^+(\tilde{t})) > 0$, there is a unique local minimizer of $R(\theta)$ in the ball $B_\rho(\widetilde{\theta})$, and a local reparametrization φ such that $\lambda_{\min}(\nabla^2_\varphi R(\varphi)) \geq c\sigma^{-6}$ for all $\varphi \in \varphi(B_\rho(\widetilde{\theta}))$. Here, $\widetilde{\theta} \in \mathbb{R}^d$ is the point where $r_k(\widetilde{\theta}) = r_{k,*}$ for all $k \in \mathbb{Z}$, $v_0(\widetilde{\theta}) = v_{0,*}$, $v_{d/2}(\widetilde{\theta}) = v_{d/2,*}$ if d is even, and $\mathrm{Arg}(v_k(\widetilde{\theta})) = \mathrm{Arg}(v_{k,*}) + \tilde{t}_k$ for each $k \in \mathcal{I}$.

(b) If d is even, then in addition, for each local minimizer $\tilde{t} \in \mathcal{S}^{|\mathcal{I}|}$ of $F^-(t)$ where $\lambda_{\min}(\nabla^2 F^-(\tilde{t})) > 0$, the same statement of (a) holds over $B_\rho(\widetilde{\theta})$ for $\widetilde{\theta} \in \mathbb{R}^d$ defined by the same conditions as in (a), except with $v_{d/2}(\widetilde{\theta}) = -v_{d/2,*}$ in place of $v_{d/2}(\widetilde{\theta}) = v_{d/2,*}$.

(c) If $F^+(t)$ and $F^-(t)$ are Morse on $\mathcal{S}^{|\mathcal{I}|}$, then (a) and (b) characterize all of the local minimizers of $R(\theta)$. For each $\theta \in \mathbb{R}^d$ outside the union of the balls $B_\rho(\widetilde{\theta})$ in (a) and (b), either $\|\nabla_\theta R(\theta)\| \geq c\sigma^{-6}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta)) \leq -c\sigma^{-6}$.

The following corollary will then follow from an analysis of the landscape of the functions $F^\pm$ for small d.

Corollary 4.27. *Let G be the cyclic group (4.38) acting on $\mathbb{R}^d$.*

(a) For $d \leq 5$ and generic $\theta_* \in \mathbb{R}^d$, there exists a (θ_*, d)-dependent constant $\sigma_0 > 0$ such that the landscape of $R(\theta)$ is globally benign for all $\sigma > \sigma_0$.

(b) For $d = 6$, there exists an open subset $U \subset \mathbb{R}^d$ and a constant $\sigma_0 > 0$ such that for all $\theta_* \in U$ and $\sigma > \sigma_0$, $R(\theta)$ has a local minimizer not belonging to $\mathcal{O}_{\theta_*}$.

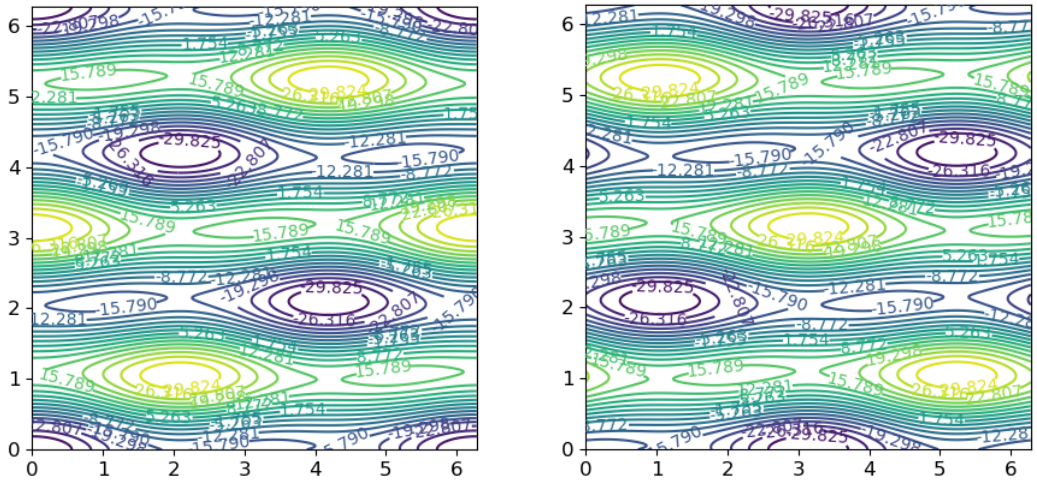


FIGURE 4.1. Contours of the functions $F^+(t_1, t_2)$ (left) and $F^-(t_1, t_2)$ (right) corresponding to θ_* in (4.42), for the group of cyclic permutations acting in dimension $d = 6$. Each function F^+ and F^- is periodic over $t_1, t_2 \in \mathcal{S} \cong [0, 2\pi)$ and has six local minimizers. Together, these twelve local minimizers of $F^{\pm}(t_1, t_2)$ correspond to six global minimizers and six spurious local minimizers of $R(\theta)$ under high noise.

For (θ_*, d)-dependent constants $C, c > 0$, the same statements hold for the empirical risk $R_n(\theta)$ with probability at least $1 - \sigma^C e^{-c\sigma^{-10}n} - Ce^{-cn^{2/3}}$.

Our proof of Corollary 4.27(b) will exhibit a particular point

$$\theta_* \approx (2.86, -0.82, -0.82, 0.41, -0.82, -0.82) \tag{4.42}$$

belonging to the open set U, for which $R(\theta)$ has spurious local minimizers. Contour maps of $F^+(t_1, t_2)$ and $F^-(t_1, t_2)$ for this point θ_* are displayed in Figure 4.1. It may be verified that F^+ and F^- each has six local minimizers given by

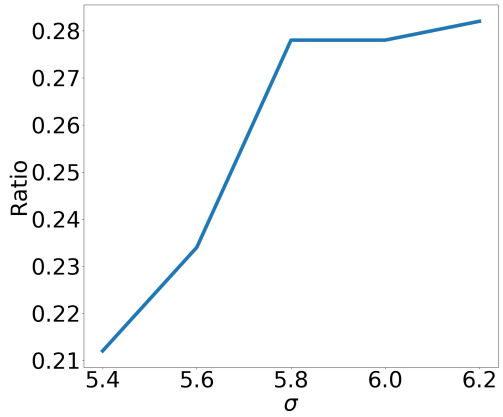
$$(t_1, t_2) = (0, 0), (\pi/3, 2\pi/3), (2\pi/3, 4\pi/3), (\pi, 0), (4\pi/3, 2\pi/3), (5\pi/3, 4\pi/3).$$

The corresponding twelve points $\widetilde{\theta}$ constitute the orbits $\mathcal{O}_{\theta_*}$ and $\mathcal{O}_{\mu_*}$ for a second point

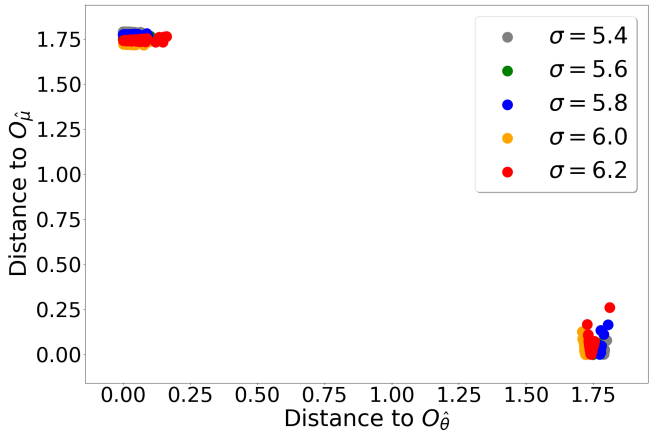
$$\mu_* \approx (2.04, 0.00, -1.63, 1.22, -1.63, 0.00). \tag{4.43}$$

Theorem 4.26 implies that for large σ and large n, the empirical risk $R_n(\theta)$ has (with high probability) twelve local minimizers, belonging to two orbits $\mathcal{O}_{\hat{\theta}}$ and $\mathcal{O}_{\hat{\mu}}$ where $(\hat{\theta}, \hat{\mu})$ are close to (θ_*, μ_*).

Simulation results in Figure 4.2 verify this behavior: We used the accelerated gradient descent (AGD) method described in Section 1.3 to minimize $R_n(\theta)$, with $n = 1{,}000{,}000$ samples at various noise levels σ. For each noise level, the underlying data $Y_1, \dots, Y_n$ was fixed, and simulations were performed with 500 random initializations $\theta^{(0)} \sim \mathcal{N}(0, \mathrm{Id})$. At noise levels $\sigma \le 5.2$, all simulations converged to the orbit of a point $\hat{\theta}$ near θ_*, suggesting a benign landscape for $R_n(\theta)$. For $\sigma \ge 5.4$, a fraction of simulations converged to the orbit of a second local minimizer $\hat{\mu}$ near μ_*, and this fraction stabilized to be roughly 28%. This value 28% may be understood as the "size" of the domain of attraction for the spurious local minimizers $\mathcal{O}_{\hat{\mu}}$ relative to that for the global minimizers $\mathcal{O}_{\hat{\theta}}$, for this particular example of θ_* in (4.42) and our simulation parameters.



(A) The fraction of AGD runs converging to the spurious local minimizers $\mathcal{O}_{\hat{\mu}}$ at different noise levels.



(B) Distances from the 250th AGD iterate to the orbits $\mathcal{O}_{\hat{\theta}}$ and $\mathcal{O}_{\hat{\mu}}$ for each run.

FIGURE 4.2. Results of applying Nesterov-accelerated gradient descent (AGD) to minimize $R_n(\theta)$ for cyclic permutations in dimension $d = 6$, with $n = 1{,}000{,}000$ samples and θ_* as in (4.42). AGD was applied from 500 random initializations for noise levels σ between 5.0 and 6.2. AGD converges to a point near $\mathcal{O}_{\hat{\theta}}$ or $\mathcal{O}_{\hat{\mu}}$ in all cases. For $\sigma = 6.2$, we find $\hat{\theta} \approx (2.84, -0.82, -0.85, 0.42, -0.79, -0.79)$ and $\hat{\mu} \approx (2.08, -0.03, -1.47, 1.17, -1.53, -0.21)$, which are close to (θ_*, μ_*) in (4.42) and (4.43).

Now proceeding with the proof of Theorem 4.26, let us fix $\theta_* \in \mathbb{R}^d$ such that $v_{k,*} \neq 0$ for all $k \not\equiv 0 \bmod 0$. For each $k \in \mathbb{Z}$, define

$$t_k(\theta) = \begin{cases} \operatorname{Arg}\big(v_k(\theta)\big) - \operatorname{Arg}\big(v_{k,*}\big) & \text{if } v_k(\theta) \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $t_k(\theta) \in \mathcal{S}$. In particular, note that

$$v_k(\theta) = \overline{v_{-k}(\theta)}, \qquad r_k(\theta) = r_{-k}(\theta), \qquad t_k(\theta) = -t_{-k}(\theta),$$

because $\theta \in \mathbb{R}^d$ is real-valued.

The proof of Theorem 4.26 rests on the following lemma, which describes the first three terms S_1, S_2, S_3 of the expansion (4.1).

Lemma 4.28. *Fix $\theta_* \in \mathbb{R}^d$ where $v_{k,*} \neq 0$ for all $k \not\equiv 0 \bmod d$. Then for some polynomial $q : \mathbb{R}^{d-1} \to \mathbb{R}$ with coefficients depending on θ_*,*

$$S_1(\theta) = -v_{0,*}v_0(\theta) + \frac{1}{2}v_0(\theta)^2 \tag{4.44}$$

$$S_2(\theta) = \sum_{i=1}^{d-1}\left(-\frac{1}{2}r_{i,*}^2 r_i(\theta)^2 + \frac{1}{4}r_i(\theta)^4\right) \tag{4.45}$$

$$S_3(\theta) = -\frac{1}{6}\sum_{\substack{i,j,k=1 \\ i+j+k\equiv 0 \bmod d}}^{d-1} r_{i,*}r_{j,*}r_{k,*}r_i(\theta)r_j(\theta)r_k(\theta)\cos\Big(t_i(\theta)+t_j(\theta)+t_k(\theta)\Big) + q\Big(r_1(\theta)^2,\dots,r_{d-1}(\theta)^2\Big). \tag{4.46}$$

Proof. Let $e = (1, \dots, 1)/\sqrt{d} \in \mathbb{R}^{d\times 1}$ and let $V \in \mathbb{R}^{d\times(d-1)}$ complete the orthonormal basis. Then the columns of V span the kernel of $\mathbb{E}_g[g]$, and Lemma 2.4 applies with $G_2 = \{V^\top h^k V : k = 0, \dots, d-1\} \subset \mathrm{O}(d-1)$ for the generator h in (4.39). Thus, noting that $e^\top\theta = v_0(\theta)$, we have

$$R(\theta) = R^{\mathrm{Id}}(v_0(\theta)) + R^{G_2}(V^\top\theta).$$

Applying the series expansion (4.1) to each of R, R^{Id}, and R^{G_2}, we have the analogous decomposition

$$S_\ell(\theta) = S_\ell^{\mathrm{Id}}(v_0(\theta)) + S_\ell^{G_2}(V^\top\theta)$$

for every $\ell \ge 1$. Note that

$$R^{\mathrm{Id}}(v_0(\theta)) = \frac{v_0(\theta)^2}{2\sigma^2} - \frac{v_{0,*}v_0(\theta)}{\sigma^2}$$

by (2.2), so that $S_1^{\mathrm{Id}}(v_0(\theta)) = -v_{0,*}v_0(\theta) + v_0(\theta)^2/2$, and $S_\ell^{\mathrm{Id}}(v_0(\theta)) = 0$ for all $\ell \ge 2$.

To compute the terms $S_\ell^{G_2}(V^\top\theta)$, we apply Lemma 4.9. For $\ell = 1$, since $\mathbb{E}_{g\sim\mathrm{Unif}(G_2)}[g] = 0$, we have $S_1^{G_2}(V^\top\theta) = 0$, so we get (4.44). For $\ell = 2$,

$$S_2^{G_2}(V^\top\theta) = -\frac{1}{2}\mathbb{E}_g[\langle V^\top\theta_*, (V^\top gV)V^\top\theta\rangle^2] + \frac{1}{4}\mathbb{E}_g[\langle V^\top\theta, (V^\top gV)V^\top\theta\rangle^2].$$

Introduce $P = VV^\top = \mathrm{Id} - ee^\top$ and the partial Fourier matrix $F \in \mathbb{C}^{(d-1)\times d}$ such that $F\theta = (v_1(\theta), \dots, v_{d-1}(\theta)) \in \mathbb{C}^{d-1}$. Denote $v = F\theta$, and let $v_* = F\theta_*$. Then note that

$$Ph^kP = F^*D^kF$$

where $D = \mathrm{diag}(\omega, \omega^2, \dots, \omega^{d-1})$, so

$$S_2^{G_2}(V^\top\theta) = -\frac{1}{2d}\sum_{k=0}^{d-1}\langle v_*, D^kv\rangle^2 + \frac{1}{4d}\sum_{k=0}^{d-1}\langle v, D^kv\rangle^2.$$

We may write

$$\frac{1}{d}\sum_{k=0}^{d-1}\langle v_*, D^kv\rangle^2 = \frac{1}{d}\sum_{k=0}^{d-1}\left(\sum_{i=1}^{d-1}\overline{v_{i,*}}\cdot\omega^{ki}v_i\right)^2 = \sum_{i,j=1}^{d-1}\overline{v_{i,*}v_{j,*}}v_iv_j\left(\frac{1}{d}\sum_{k=0}^{d-1}\omega^{ki+kj}\right).$$

Applying

$$\sum_{k=0}^{d-1} w^{jk} = \begin{cases} d & \text{if } j \equiv 0 \mod d \\ 0 & \text{if } j \not\equiv 0 \mod d, \end{cases} \tag{4.47}$$

and also $v_i = \overline{v_{-i}}$ and $v_{i,*} = \overline{v_{-i,*}}$, this yields

$$\frac{1}{d}\sum_{k=0}^{d-1}\langle v_*, D^kv\rangle^2 = \sum_{i,j=1}^{d-1}\overline{v_{i,*}v_{j,*}}v_iv_j\cdot\mathbf{1}\{i+j\equiv 0 \bmod d\} = \sum_{i=1}^{d-1}|v_{i,*}|^2\cdot|v_i|^2 = \sum_{i=1}^{d-1}r_{i,*}^2r_i(\theta)^2.$$

A similar computation shows $d^{-1}\sum_k\langle v, D^kv\rangle^2 = \sum_{i=1}^{d-1}r_i(\theta)^4$, which yields (4.45).

For $\ell = 3$, applying Lemma 4.9 and similar arguments,

$$\begin{aligned} S_3^{G_2}(V^\top\theta) &= \frac{1}{d}\sum_{p=0}^{d-1}\left(-\frac{\langle v_*, D^pv\rangle^3}{6} + \frac{\langle v, D^pv\rangle^3}{12}\right) \\ &\qquad + \frac{1}{d^2}\sum_{p,q=0}^{d-1}\left(\frac{\langle D^pv, D^qv\rangle\langle v_*, D^pv\rangle\langle v_*, D^qv\rangle}{2} - \frac{\langle D^pv, D^qv\rangle\langle v, D^pv\rangle\langle v, D^qv\rangle}{3}\right) \\ &= \sum_{i,j,k=1}^{d-1}\left[\left(-\frac{\overline{v_{i,*}v_{j,*}v_{k,*}}v_iv_jv_k}{6} + \frac{|v_i|^2|v_j|^2|v_k|^2}{12}\right)\mathbf{1}\{i+j+k\equiv 0 \bmod d\}\right. \end{aligned}$$

$$+\left(\frac{|v_i|^2\overline{v_{j,*}v_{k,*}}v_jv_k}{2}-\frac{|v_i|^2|v_j|^2|v_k|^2}{3}\right)\mathbf{1}\{i+k\equiv 0 \bmod d,\ -i+j\equiv 0 \bmod d\}\Bigg].$$

Observe that for the second term, we must have $k\equiv -i$ and $j\equiv i$, in which case $\overline{v_{j,*}v_{k,*}}v_jv_k = |v_i|^2|v_{i,*}|^2$. Then applying also $\overline{v_{i,*}}v_i = r_{i,*}r_ie^{\mathbf{i}t_i}$ (where we write r_i, t_i for $r_i(\theta), t_i(\theta)$), for some polynomial $q:\mathbb{R}^{d-1}\to\mathbb{R}$ we get

$$S_3^{G_2}(V^\top\theta) = -\frac{1}{6}\sum_{i,j,k=1}^{d-1} r_{i,*}r_{j,*}r_{k,*}r_ir_jr_ke^{\mathbf{i}(t_i+t_j+t_k)}\mathbf{1}\{i+j+k\equiv 0 \bmod d\} + q(r_1^2,\dots,r_{d-1}^2).$$

Taking the real part on both sides yields (4.46). □

Proof of Theorem 4.26. For each $\widetilde{\theta}\in\mathbb{R}^d$, we construct a local reparametrization $\varphi=(\varphi^1,\varphi^2,\varphi^3)$ as follows: Let $\varphi^1(\theta)=v_0(\theta)$. For each $k\in\mathcal{I}$, if $v_k(\widetilde{\theta})\neq 0$, then include $r_k(\theta)$ as a coordinate of φ^2. If $v_k(\widetilde{\theta})=0$, then include $\operatorname{Re}v_k(\theta)$ and $\operatorname{Im}v_k(\theta)$ as two coordinates of φ^2. If d is even, include also $v_{d/2}(\theta)$ as a coordinate of φ^2. Then for each $k\in\mathcal{I}$ where $v_k(\widetilde{\theta})\neq 0$, include $t_k(\theta)$ as a coordinate of φ^3. If there are m coordinates $k\in\mathcal{I}$ where $v_k(\widetilde{\theta})\neq 0$, then $\varphi^3\in\mathbb{R}^m$ and $\varphi^2\in\mathbb{R}^{d-m-1}$. It is easily verified that this defines a local reparametrization in some neighborhood $U_{\widetilde{\theta}}$ around every $\widetilde{\theta}\in\mathbb{R}^d$. Note that S_1 depends only on φ^1, and S_2 on φ^1 and φ^2.

We now apply Lemmas 4.10 and 4.13. Let $\widetilde{\varphi}=\varphi(\widetilde{\theta})$. For $\widetilde{\theta}\in\mathbb{R}^d$ where $v_0(\widetilde{\theta})\neq v_{0,*}$, we have $\nabla_{\varphi^1}S_1(\widetilde{\varphi})\neq 0$. For $\widetilde{\theta}\in\mathbb{R}^d$ where $v_k(\widetilde{\theta})\neq 0$ and $r_k(\widetilde{\theta})\neq r_{k,*}$ for some $k\in\mathcal{I}$, we similarly have $\nabla_{\varphi^2}S_2(\widetilde{\varphi})\neq 0$, because the derivative of S_2 in the coordinate r_k is non-zero. For $\widetilde{\theta}\in\mathbb{R}^d$ where $v_k(\widetilde{\theta})=0$ for some $k\in\mathcal{I}$, let us write $r_k(\theta)^2=(\operatorname{Re}v_k(\theta))^2+(\operatorname{Im}v_k(\theta))^2$ in (4.45). Differentiating S_2 twice in these variables $\operatorname{Re}v_k(\theta)$ and $\operatorname{Im}v_k(\theta)$ and evaluating at $\operatorname{Re}v_k(\widetilde{\theta})=\operatorname{Im}v_k(\widetilde{\theta})=0$, we get that the Hessian of S_2 in these variables is $-r_{k,*}^2\operatorname{Id}$. Thus, $\lambda_{\min}(\nabla^2_{\varphi^2}S_2(\widetilde{\varphi}))<0$. Finally, for even d and $\widetilde{\theta}\in\mathbb{R}^d$ where $v_{d/2}(\widetilde{\theta})\notin\{+v_{d/2,*},-v_{d/2,*}\}$, let us write $r_{d/2}(\theta)^2=v_{d/2}(\widetilde{\theta})^2$ in (4.45). Then either $v_{d/2}(\widetilde{\theta})\neq 0$ and $\nabla_{\varphi^2}S_2(\widetilde{\varphi})\neq 0$, or $v_{d/2}(\widetilde{\theta})=0$ and $\lambda_{\min}(\nabla_{\varphi^2}S_2(\widetilde{\varphi}))<0$. In all of these cases, Lemma 4.10 implies either $\|\nabla_\theta R(\theta)\|\geq c\sigma^{-4}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta))\leq -c\sigma^{-4}$, for all $\theta\in U_{\widetilde{\theta}}$ and $\sigma>\sigma_0$.

It remains to consider those points $\widetilde{\theta}\in\mathbb{R}^d$ where $v_0(\widetilde{\theta})=v_{0,*}$ and $r_k(\widetilde{\theta})=r_{k,*}\neq 0$ for all $k\in\mathbb{Z}$. For such $\widetilde{\theta}$, we have $\varphi^3\equiv(t_1,\dots,t_{|\mathcal{I}|})\in\mathbb{R}^{|\mathcal{I}|}$. When d is odd, the summation defining (4.46) may be written as that over $i,j,k\in\mathcal{I}\cup-\mathcal{I}$ with $i+j+k\equiv 0 \bmod d$, and the restriction of $S_3(\varphi)$ to points $\varphi\in\mathbb{R}^d$ where $r_k=r_{k,*}$ for all $k\in\mathbb{Z}$ coincides with $F^+(t)$. When d is even, we may isolate the terms of the summation in (4.46) where some coordinate, say k, equals $d/2$. Then we must have $i+j\equiv d/2 \bmod d$, and the constraint $i,j\not\equiv 0 \bmod d$ is equivalent to $i,j\in\mathcal{I}\cup-\mathcal{I}$. When $v_{d/2}=v_{d/2,*}$, we have $t_k=0$ so $\cos(t_i+t_j+t_k)=\cos(t_i+t_j)$. In this case, $S_3(\varphi)$ restricted to $r_k=r_{k,*}$ is the function $F^+(t)$, where the factor $1/2$ is produced from $1/6$ by considering the three symmetric settings where i, j, or k is $d/2$. When $v_{d/2}=-v_{d/2,*}$, we have $t_k=\pi$, so $\cos(t_i+t_j+t_k)=-\cos(t_i+t_j)$. In this case, $S_3(\varphi)$ restricted to $r_k=r_{k,*}$ is the function $F^-(t)$.

Thus, if $\tilde{t}=\widetilde{\varphi}^3$ is not a critical point of $F^\pm(t)$, then $\nabla_{\varphi^3}S_3(\widetilde{\varphi})\neq 0$. If $\tilde{t}$ is a critical point where $\lambda_{\min}(\nabla^2F^\pm(t))<0$, then also $\lambda_{\min}(\nabla_{\varphi^3}S_3(\widetilde{\varphi}))<0$. In these cases, Lemma 4.10 implies that either $\|\nabla_\theta R(\theta)\|\geq c\sigma^{-6}$ or $\lambda_{\min}(\nabla^2_\theta R(\theta))\leq -c\sigma^{-6}$, for all $\theta\in U_{\widetilde{\theta}}$ and $\sigma>\sigma_0$. If $\tilde{t}$ is a critical point of $F^\pm(t)$ where $\lambda_{\min}(\nabla^2F^\pm(t))>0$, then $\widetilde{\varphi}$ is a pseudo-local-minimizer of $R(\theta)$, and Lemma 4.10 implies both that there is a unique local minimizer of $R(\varphi)$ in $\varphi(U_{\widetilde{\theta}})$ and that $\nabla^2_\varphi R(\varphi)\geq c\sigma^{-6}$ on $\varphi(U_{\widetilde{\theta}})$. Finally, if $F^\pm(t)$ is Morse, then this accounts for all possible points $\widetilde{\theta}$.

Taking a finite collection of these sets $U_{\widetilde{\theta}}$ which cover $\{\theta:\|\theta\|\leq M\}$, the above constants $c,\sigma_0>0$ may be chosen to be uniform over this finite cover. Then for small enough $\rho>0$, the

above arguments establish the claims of the theorem for $\|\theta\| \le M$. The result for $\|\theta\| > M$ follows from Lemma 4.17. □

The same statements then hold for the empirical risk $R_n(\theta)$, with high probability when $n \gg \sigma^{10} \log \sigma$. The proof is the same as Corollary 4.20, and we omit this for brevity.

Corollary 4.29. *For some (θ_*, d)-dependent constants $C, c > 0$, the statements of Theorem 4.26 hold also for $R_n(\theta)$, with probability at least $1 - \sigma^C e^{-c\sigma^{-10}n} - Ce^{-cn^{2/3}}$.*

Finally, let us analyze the functions $F^{\pm}$ in dimensions up to $d = 6$.

Proof of Corollary 4.27. For part (a), the result for $d = 1$ or 2 follows from the analysis of all permutations in Theorem 4.21. For $d \in \{3, 4, 5\}$, by Theorem 4.26 and Corollary 4.29, it suffices to show that $F^{\pm}(t)$ is a Morse function of $t \in \mathcal{S}^{|\mathcal{I}|}$, and its only local minimizers correspond to the points $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$.

For $d = 3$ or 4, $\mathcal{I} = \{1\}$, so $F^{\pm}(t)$ is a function of a single scalar argument in $t_1 \in \mathcal{S}$. We may directly compute its first and second derivatives. For $d = 3$, we get

$$\nabla F^+(t_1) = r_{1,*}^6 \sin(3t_1), \quad \nabla^2 F^+(t_1) = 3r_{1,*}^6 \cos(3t_1).$$

Then F^+ is Morse and there are six critical points, three of which are the local minimizers $\{0, 2\pi/3, 4\pi/3\}$. These correspond to the three points $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$. For $d = 4$, we have

$$\nabla F^{\pm}(t_1) = \pm 2r_{1,*}^6 \sin(2t_1), \quad \nabla^2 F^{\pm}(t_1) = \pm 4r_{1,*}^6 \cos(2t_1).$$

Each function F^+ and F^- is Morse with four critical points. For F^+, there are two local minimizers $\{0, \pi\}$, and for F^-, there are two local minimizers $\{\pi/2, 3\pi/2\}$. These correspond to the four points $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$.

For $d = 5$, we have $\mathcal{I} = \{1, 2\}$. Let us abbreviate

$$s_i = r_{i,*}^2, \qquad u_1 = 2t_1 - t_2, \qquad u_2 = t_1 + 2t_2.$$

Then

$$\nabla F^+(t) = \Big(2s_1^2 s_2 \sin u_1 + s_1 s_2^2 \sin u_2, \;\; -s_1^2 s_2 \sin u_1 + 2s_2 s_1^2 \sin u_2\Big),$$
$$\nabla^2 F^+(t) = \begin{pmatrix} 4s_1^2 s_2 \cos u_1 + s_1 s_2^2 \cos u_2 & -2s_1^2 s_2 \cos u_1 + 2s_1 s_2^2 \cos u_2 \\ -2s_1^2 s_2 \cos u_1 + 2s_1 s_2^2 \cos u_2 & s_1^2 s_2 \cos u_1 + 4s_1 s_2^2 \cos u_2 \end{pmatrix}.$$

From this, we may also compute

$$\det \nabla^2 F^+(t) = 25 s_1^3 s_2^3 \cos u_1 \cos u_2,$$
$$\operatorname{Tr} \nabla^2 F^+(t) = 5s_1^2 s_2 \cos u_1 + 5s_1 s_2^2 \cos u_2.$$

For generic θ_* and hence generic (s_1, s_2), the condition $\nabla F^+(t) = 0$ for a critical point requires $\sin u_1 = \sin u_2 = 0$. We have $\det \nabla^2 F^+(t) \neq 0$ at such points, so F^+ is Morse. The condition $\nabla^2 F^+(t) \succ 0$ for a local minimizer then requires $\det H^+(t) > 0$ and $\operatorname{Tr} H^+(t) > 0$, so we must have $\cos u_1 = \cos u_2 = 1$, and hence and $t_1 + 2t_2 \equiv 2t_1 - t_2 \equiv 0 \bmod 2\pi$. This implies that $5t_1 \equiv 0 \bmod 2\pi$, and there are five local minimizers $(t_1, t_2) = (0, 0)$, $(2\pi/5, 4\pi/5)$, $(4\pi/5, 8\pi/5)$, $(6\pi/5, 2\pi/5)$, or $(8\pi/5, 6\pi/5)$. These correspond to the five points $\widetilde{\theta} \in \mathcal{O}_{\theta_*}$. Together with Theorem 4.26, this shows part (a).

For part (b), when $d = 6$, we also have $\mathcal{I} = \{1, 2\}$. Writing

$$s_i = r_i^2, \qquad u_1 = 2t_1 - t_2, \qquad u_2 = t_1 + t_2$$

we may compute

$$F^{\pm}(t) = -\Big(s_1^2 s_2 \cos u_1 + \tfrac{1}{3} s_2^3 \cos(3t_2) \pm 2s_1 s_2 s_3 \cos u_2\Big),$$

$$\nabla F^{\pm}(t) = \Big(2s_1^2 s_2 \sin u_1 \pm 2s_1 s_2 s_3 \sin u_2,\ -s_1^2 s_2 \sin u_1 + s_2^3 \sin(3t_2) \pm 2s_1 s_2 s_3 \sin u_2\Big),$$
$$\nabla^2 F^{\pm}(t) = \begin{pmatrix} 4s_1^2 s_2 \cos u_1 \pm 2s_1 s_2 s_3 \cos u_2 & -2s_1^2 s_2 \cos u_1 \pm 2s_1 s_2 s_3 \cos u_2 \\ -2s_1^2 s_2 \cos u_1 \pm 2s_1 s_2 s_3 \cos u_2 & s_1^2 s_2 \cos u_1 + 3s_2^3 \cos(3t_2) \pm 2s_1 s_2 s_3 \cos u_2 \end{pmatrix}.$$

Let us exhibit a point θ_* where F^+ and F^- have spurious local minimizers: Consider $(s_1, s_2, s_3) = (1, 4, 1)$ so that $(r_{1,*}, r_{2,*}, r_{3,*}) = (1, 2, 1)$. Suppose for simplicity that $v_{0,*} = 0$ and $\operatorname{Arg} v_{1,*} = \operatorname{Arg} v_{2,*} = \operatorname{Arg} v_{3,*} = 0$. Moving back to the standard basis, this corresponds to the point θ_* previously defined in (4.42). It may be verified that for this point θ_*, $F^+(t)$ and $F^-(t)$ are both Morse, each with six local minimizers at

$$(t_1, t_2) = (0, 0), (\pi/3, 2\pi/3), (2\pi/3, 4\pi/3), (\pi, 0), (4\pi/3, 2\pi/3), (5\pi/3, 4\pi/3).$$

The corresponding twelve points $\widetilde{\theta}$ constitute the orbits $\mathcal{O}_{\theta_*}$ and $\mathcal{O}_{\mu_*}$ for μ_* defined in (4.43). Thus, for $\theta_* \in \mathbb{R}^6$ given by (4.42), some $\sigma_0 \equiv \sigma_0(\theta_*) > 0$, and all $\sigma > \sigma_0$, Theorem 4.26 guarantees that there are six additional local minimizers of $R(\theta)$ near the points of $\mathcal{O}_{\mu_*}$.

We complete the proof by showing that this holds not just at the single point (4.42), but for all θ_* in an open neighborhood U of this point. For this, consider $\nabla F^{\pm}$ and $\nabla^2 F^{\pm}$ as functions jointly of $(s_1, s_2, s_3, t_1, t_2)$. At $(s_1, s_2, s_3) = (1, 4, 1)$, since $F^{\pm}(t)$ are Morse, for each of the above six pairs (t_1, t_2) and each choice of sign $\pm$, the Hessian $\nabla_t^2 F^{\pm}(s_1, s_2, s_3, t_1, t_2)$ is non-singular. Then the implicit function theorem guarantees that, for each of these twelve choices, there is an open neighborhood V of $(s_1, s_2, s_3) = (1, 4, 1)$ and two functions $a_1(s_1, s_2, s_3), a_2(s_1, s_2, s_3)$ on V such that $(a_1(1, 4, 1), a_2(1, 4, 1)) = (t_1, t_2)$ and

$$\nabla_t F^{\pm}(s_1, s_2, s_3, a_1(s_1, s_2, s_3), a_2(s_1, s_2, s_3)) = 0$$

for all $(s_1, s_2, s_3) \in V$. Taking V sufficiently small, by continuity we may also ensure

$$\nabla_t^2 F^{\pm}(s_1, s_2, s_3, a_1(s_1, s_2, s_3), a_2(s_1, s_2, s_3)) \succ 0$$

for all $(s_1, s_2, s_3) \in V$. Then, denoting by $V_\cap$ the intersection of these twelve open neighborhoods V, and defining the open set

$$U = \Big\{\theta_* \in \mathbb{R}^d : (r_1(\theta_*)^2, r_2(\theta_*)^2, r_3(\theta_*)^2) \in V_\cap\Big\},$$

for each $\theta_* \in U$ there are twelve non-degenerate local minimizers of $F^{\pm}(t)$. Applying Theorem 4.26 again, for each point $\theta_* \in U$ and all $\sigma > \sigma_0$, this implies that $R(\theta)$ has twelve local minimizers, six of which do not belong to the true orbit $\mathcal{O}_{\theta_*}$. □

Appendix A. Auxiliary lemmas and proofs

A.1. **Cumulants and cumulant bounds.** The order-ℓ cumulant $\kappa_\ell(X)$ of a random variable X is defined recursively by the moment-cumulant relations

$$\mathbb{E}[X^\ell] = \sum_{\text{partitions } \pi \text{ of } [\ell]} \prod_{S \in \pi} \kappa_{|S|}(X).$$

More generally, for random variables $X_1, \dots, X_\ell$, the mixed cumulants $\kappa_{|S|}(X_k : k \in S)$ for $S \subseteq [\ell]$ are defined recursively by the moment-cumulant relations

$$\mathbb{E}\left[\prod_{i \in T} X_i\right] = \sum_{\text{partitions } \pi \text{ of } T} \prod_{S \in \pi} \kappa_{|S|}(X_k : k \in S).$$

These relations may be Möbius-inverted to obtain the explicit definition

$$\kappa_\ell(X_1, \dots, X_\ell) = \sum_{\text{partitions } \pi \text{ of } [\ell]} (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{S \in \pi} \mathbb{E}_g\left[\prod_{i \in S} X_i\right] \tag{A.1}$$

where $|\pi|$ is the number of sets in π (see [McC18, Sec. 2.3.4]). If $X_1 = \ldots = X_\ell = X$, then $\kappa_\ell(X_1, \ldots, X_\ell) = \kappa_\ell(X)$. The mixed cumulant $\kappa_\ell(X_1, \ldots, X_\ell)$ is multi-linear and permutation-invariant in its ℓ arguments. We have $\kappa_1(X) = \mathbb{E}[X]$, $\kappa_2(X) = \mathrm{Var}[X]$, and $\kappa_2(X_1, X_2) = \mathrm{Cov}[X_1, X_2]$.

The cumulant generating function of a random variable X is $K_X(s) = \log \mathbb{E}[e^{sX}]$. This has the formal series expansion

$$K_X(s) = \sum_{\ell=1}^{\infty} \kappa_\ell(X) \frac{s^\ell}{\ell!}. \tag{A.2}$$

If $K_X(s)$ exists in a neighborhood of 0, then its ℓ^{th} derivative at 0 is $K_X^{(\ell)}(0) = \kappa_\ell(X)$. Similarly, the cumulant generating function of a random vector $u \in \mathbb{R}^d$ is $K_u(\theta) = \log \mathbb{E}[e^{\langle \theta, u \rangle}]$. This has the formal series expansion

$$K_u(\theta) = \sum_{\ell_1, \ldots, \ell_d = 1}^{\infty} \frac{\theta_1^{\ell_1} \ldots \theta_d^{\ell_d}}{\ell_1! \ldots \ell_d!} \kappa_{\ell_1 + \ldots + \ell_d}(u_1, \ldots, u_1, \ldots, u_d, \ldots, u_d)$$

where in $\kappa_{\ell_1 + \ldots + \ell_d}(u_1, \ldots, u_1, \ldots, u_d, \ldots, u_d)$, each u_j appears ℓ_j times. The ℓ^{th} derivative at 0 is

$$\nabla^\ell K_u(0) = \kappa_\ell(u) \in (\mathbb{R}^d)^{\otimes \ell},$$

where $\kappa_\ell(u)$ denotes the order-ℓ cumulant tensor of u. This has entries, for $i_1, \ldots, i_\ell \in [d]$,

$$\kappa_\ell(u)_{i_1, \ldots, i_\ell} = \kappa_\ell(u_1, \ldots, u_1, \ldots, u_d, \ldots, u_d)$$

where each coordinate u_j appears ℓ_j times if ℓ_j of the indices $i_1, \ldots, i_\ell$ equal j. The first two cumulant tensors are $\kappa_1(u) = \mathbb{E}[u]$ and $\kappa_2(u) = \mathrm{Cov}[u]$.

More generally, if $K_u(\theta)$ exists in a neighborhood of θ, a reweighted exponential family law $p(u|\theta)$ may be defined by the expectation

$$\mathbb{E}[f(u) \mid \theta] = \mathbb{E}[f(u) e^{\langle \theta, u \rangle - K_u(\theta)}] = \frac{\mathbb{E}[f(u) e^{\langle \theta, u \rangle}]}{\mathbb{E}[e^{\langle \theta, u \rangle}]}.$$

Then $\nabla^\ell K_u(\theta) = \kappa_\ell(u \mid \theta)$, the order-$\ell$ cumulant tensor of this reweighted law (see [LC06, Theorem 1.5.10]).

The following result provides an upper bound bound for these cumulants when $X, X_1, \ldots, X_\ell$ are bounded random variables. This bound is tight up to an exponential factor in ℓ, as may be seen for $X \sim \mathrm{Unif}([0,1])$ where $\kappa_\ell(X) = B_\ell/\ell$ and B_ℓ is the ℓ^{th} Bernoulli number (see [BKS20, Example 2.7]), satisfying $|B_{2\ell}| \sim 4\sqrt{\pi \ell}(\ell/(\pi e))^{2\ell}$.

Lemma A.1. *(a) If $|X| \le m$ almost surely, then $|\kappa_\ell(X)| \le (m\ell)^\ell$.*
(b) If $|X_i| \le m_i$ almost surely for each $i = 1, \ldots, \ell$, then $|\kappa_\ell(X_1, \ldots, X_\ell)| \le \ell^\ell m_1 \ldots m_\ell$.
(c) If $|X| \le m$ almost surely, then the series (A.2) is absolutely convergent for $|s| < 1/(me)$.

Proof. We apply (A.1). Enumerating over $v = |\pi|$, we have

$$\sum_{\text{partitions } \pi \text{ of } [\ell]} (|\pi| - 1)! = \sum_{v=1}^{\ell} \frac{(v-1)!}{v!} \sum_{\ell_1 + \ldots + \ell_v = \ell} \binom{\ell}{\ell_1, \ldots, \ell_v} = \sum_{v=1}^{\ell} \frac{1}{v} \cdot v^\ell = \sum_{v=1}^{\ell} v^{\ell-1} \le \ell^\ell,$$

so (b) follows from (A.1). Specializing to $X_1 = \ldots = X_\ell$ yields (a), and (c) follows from (a) and the bound $\ell! \ge \ell^\ell / e^\ell$. □

A.2. **Reparametrization by invariant polynomials.** We prove Lemmas 4.3 and 4.5. Parts of these are well-known, but we provide a brief proof here for convenience.

We recall the more usual definition of transcendence degree for two fields $E \subset F$, where $\mathrm{trdeg}(F/E)$ is the maximum number of elements in F that are algebraically independent over E. We verify also in the proof of Lemma 4.3 that our definition of $\mathrm{trdeg}(A)$ for any subset $A \subseteq \mathcal{R}^G$ coincides with $\mathrm{trdeg}(\mathbb{R}(A)/\mathbb{R})$, where $\mathbb{R}(A)$ is the field of rational functions generated by A.

Proof of Lemma 4.3. Consider any subsets $A' \subseteq A \subseteq \mathcal{R}^G$, where A' is algebraically independent. Call A' *maximal* in A if $A' \cup \{a\}$ is algebraically dependent for every $a \in A \setminus A'$. Let A' be maximal in A, and suppose $|A'| = k$. Let $\mathbb{R}(A)$ and $\mathbb{R}(A')$ be the fields of G-invariant rational functions generated by A and A'. Algebraic independence of A' implies that $\mathrm{trdeg}(\mathbb{R}(A')/\mathbb{R}) = k$. Maximality of A' implies that each $a \in A$ is algebraic over $\mathbb{R}(A')$. Then $\mathbb{R}(A)$ is an algebraic extension of $\mathbb{R}(A')$, so $\mathrm{trdeg}(\mathbb{R}(A)/\mathbb{R}(A')) = 0$, hence $\mathrm{trdeg}(\mathbb{R}(A)/\mathbb{R}) = k$. This verifies that every such maximal algebraically independent set A' of A has the same cardinality, which coincides with $\mathrm{trdeg}(\mathbb{R}(A)/\mathbb{R})$.

Letting $\mathbb{R}(\theta_1, \dots, \theta_d)$ and $\mathbb{R}(\mathcal{R}^G)$ be the fields of all rational functions and all G-invariant rational functions in θ, respectively, $\mathbb{R}(\theta_1, \dots, \theta_d)$ is an algebraic extension of $\mathbb{R}(\mathcal{R}^G)$ (see [CLO92, Lemma 11]), so $\mathrm{trdeg}(\mathbb{R}(\theta_1, \dots, \theta_d)/\mathbb{R}(\mathcal{R}^G)) = 0$. Since $\mathrm{trdeg}(\mathbb{R}(\theta_1, \dots, \theta_d)/\mathbb{R}) = d$, this shows $\mathrm{trdeg}(\mathcal{R}^G) = \mathrm{trdeg}(\mathbb{R}(\mathcal{R}^G)/\mathbb{R}) = d$. Thus $\mathrm{trdeg}(\mathcal{R}^G_{\le L}) = d$ for some $L \ge 1$, and there exists a smallest such L. To construct φ, let φ^1 be any maximal algebraically independent subset of $\mathcal{R}^G_1$. The above implies that the cardinality of φ^1 is $d_1 = \mathrm{trdeg}(\mathcal{R}^G_1)$. These polynomials have degree exactly 1. Now extend this to any maximal algebraically independent subset (φ^1, φ^2) of $\mathcal{R}^G_2$. The above implies that the cardinality of φ^2 is $d_2 = \mathrm{trdeg}(\mathcal{R}^G_2) - \mathrm{trdeg}(\mathcal{R}^G_1)$. If $d_2 > 0$, then the polynomials of φ^2 must have degree exactly 2, by maximality of φ^1. We may iterate this procedure to obtain $(\varphi^1, \dots, \varphi^L)$. □

Proof of Lemma 4.5. For parts (a) and (b), recall by [ER93, Theorem 2.3] that $\varphi_1, \dots, \varphi_k$ are algebraically independent if and only if $\nabla\varphi_1, \dots, \nabla\varphi_k$ are linearly independent over the field of rational functions $\mathbb{C}(\theta_1, \dots, \theta_d)$. For part (a), this linear independence means that some maximal $k \times k$ minor of the $k \times d$ derivative $\mathsf{d}_\theta\varphi$ does not vanish in $\mathbb{C}(\theta_1, \dots, \theta_d)$. Then that same maximal minor does not vanish in $\mathbb{C}$ for generic $\theta \in \mathbb{R}^d$, showing linear independence for generic θ. For part (b), linear independence at any point θ implies that some maximal minor of $\mathsf{d}_\theta\varphi$ does not vanish and hence $\nabla\varphi_1, \dots, \nabla\varphi_k$ are linearly independent over $\mathbb{C}(\theta_1, \dots, \theta_d)$, implying algebraic independence.

For part (c), let us arbitrarily extend $(\varphi_1, \dots, \varphi_k)$ to a system of coordinates $\varphi = (\varphi_1, \dots, \varphi_d)$, where $\mathsf{d}\varphi$ is non-singular in a neighborhood of $\widetilde{\theta}$. (Here, $\varphi_{k+1}, \dots, \varphi_d$ are general analytic functions and need not belong to $\mathcal{R}^G$.) By the inverse function theorem, there is a neighborhood U of $\widetilde{\theta}$ and corresponding neighborhood $\varphi(U)$ of $\varphi(\widetilde{\theta})$ for which θ is an analytic function of $\varphi \in \varphi(U)$. Then any polynomial $\psi \in \mathcal{R}^G_{\le \ell}$ is such that $\psi(\theta)$ is also an analytic function of $\varphi \in \varphi(U)$. Let us write this function as $\psi = f(\varphi)$. Then $\psi(\theta) = f(\varphi(\theta))$ for all $\theta \in U$, so by the chain rule,

$$\mathsf{d}\psi(\theta) = \mathsf{d}_\varphi f(\varphi) \cdot \mathsf{d}\varphi(\theta). \tag{A.3}$$

By part (b), since $(\varphi_1, \dots, \varphi_k, \psi)$ are algebraically dependent, the gradients $\nabla\varphi_1, \dots, \nabla\varphi_k, \nabla\psi$ must be linearly dependent at every $\theta \in U$. So $\nabla\psi = \mathsf{d}\psi^\top$ belongs to the span of $\nabla\varphi_1, \dots, \nabla\varphi_k$ at every $\theta \in U$. Since $\mathsf{d}\varphi(\theta)$ is a non-singular matrix, this and (A.3) imply that $\nabla_\varphi f = \mathsf{d}_\varphi f^\top$ has coordinates $k+1, \dots, d$ equal to 0 for every $\varphi \in \varphi(U)$. So f is in fact an analytic function of only the first k variables $\varphi_1, \dots, \varphi_k$ over $\varphi(U)$, which is the statement of part (c). □

A.3. **Concentration inequality for $\sum_i \|\varepsilon_i\|^3$.** We prove the inequality (2.22). We use the following concentration result, which specializes [AW15, Theorem 1.2] to Gaussian random variables.

Theorem A.2 ([AW15])**.** *Suppose $f : \mathbb{R}^m \to \mathbb{R}$ is D times continuously-differentiable, and $\nabla^D f(x)$ is uniformly bounded over $x \in \mathbb{R}^m$. Let $\varepsilon \in \mathbb{R}^m$ have i.i.d. $\mathcal{N}(0,1)$ coordinates. Then for a constant*

$c \equiv c(D) > 0$,

$$\mathbb{P}[|f(\varepsilon) - \mathbb{E}f(\varepsilon)| \geq t] \leq 2e^{-c\eta_f(t)}.$$

Here,

$$\eta_f(t) = \min\left(\min_{\text{partitions } \mathcal{J} \text{ of } [D]} \left(\frac{t}{\sup_{x\in\mathbb{R}^m} \|\nabla^D f(x)\|_{\mathcal{J}}} \right)^{2/|\mathcal{J}|}, \right.$$
$$\left. \min_{1\leq d\leq D-1} \min_{\text{partitions } \mathcal{J} \text{ of } [d]} \left(\frac{t}{\|\mathbb{E}[\nabla^d f(\varepsilon)]\|_{\mathcal{J}}} \right)^{2/|\mathcal{J}|} \right)$$

where $|\mathcal{J}| \equiv K$ is the number of sets in the partition $\mathcal{J} = \{J_1, \dots, J_K\}$ of $[d]$, and

$$\|A\|_{\mathcal{J}} = \sup\left(\sum_{i_1,\dots,i_d=1}^{m} a_{i_1,\dots,i_d} \prod_{k=1}^{K} x^{(k)}_{(i_\ell:\ell\in J_k)} : \|x^{(k)}\|_{\mathrm{HS}} \leq 1 \text{ for all } k = 1, \dots, K \right).$$

In this expression, $x^{(k)}$ denotes an order-$|J_k|$ tensor in $(\mathbb{R}^m)^{\otimes|J_k|}$, and $x^{(k)}_{(i_\ell:\ell\in J_k)}$ is its entry at the indices $(i_\ell : \ell \in J_k)$.

To show (2.22), let us write the coordinates of ε_i as ε_{ij}. We consider

$$f(\varepsilon_1, \dots, \varepsilon_n) = \sum_{i=1}^{n} \|\varepsilon_i\|^3$$

as a function of the $m = nd$ standard Gaussian variables ε_{ij}, and apply the above result with $D = 3$ and this function $f : \mathbb{R}^{nd} \to \mathbb{R}$. We analyze $\eta_f(t)$: Applying $\partial_{\varepsilon_{ij}}\|\varepsilon_i\| = \varepsilon_{ij}/\|\varepsilon_i\|$, a direct computation yields

$$\begin{aligned}
\partial_{\varepsilon_{ij}} f &= 3\|\varepsilon_i\|\varepsilon_{ij}, \\
\partial_{\varepsilon_{ij}}\partial_{\varepsilon_{ik}} f &= 3\|\varepsilon_i\|\mathbf{1}\{j=k\} + 3\varepsilon_{ij}\varepsilon_{ik}/\|\varepsilon_i\|, \\
\partial_{\varepsilon_{ij}}\partial_{\varepsilon_{ik}}\partial_{\varepsilon_{i\ell}} f &= 3(\varepsilon_{i\ell}\mathbf{1}\{j=k\} + \varepsilon_{ik}\mathbf{1}\{j=\ell\} + \varepsilon_{ij}\mathbf{1}\{k=\ell\})/\|\varepsilon_i\| - 3\varepsilon_{ij}\varepsilon_{ik}\varepsilon_{i\ell}/\|\varepsilon_i\|^3,
\end{aligned}$$

and all other partial derivatives up to order three are 0. Taking expectations above and applying sign invariance of ε_{ij}, we have $\mathbb{E}[\nabla f] = 0$ and $\mathbb{E}[\nabla^2 f] = c\,\mathrm{Id}$ (in dimension $nd \times nd$) for a constant $c > 0$. Then $\|\mathbb{E}[\nabla f]\|_{\{1\}} = 0$, $\|\mathbb{E}[\nabla^2 f]\|_{\{1,2\}} = \|\mathbb{E}[\nabla^2 f]\|_{\mathrm{HS}} = c\sqrt{n}$, and $\|\mathbb{E}[\nabla^2 f]\|_{\{1\},\{2\}} = \|\mathbb{E}[\nabla^2 f]\| = c$. Thus

$$\min_{1\leq d\leq D-1} \min_{\text{partitions } \mathcal{J} \text{ of } [d]} \left(\frac{t}{\|\mathbb{E}[\nabla^d f(\varepsilon)]\|_{\mathcal{J}}} \right)^{2/|\mathcal{J}|} \geq c' \min(t^2/n, t). \tag{A.4}$$

The third derivative $A = \nabla^3 f$ has n non-zero blocks of size $d \times d \times d$, with entries uniformly bounded in the range $[-12, 12]$. We observe that for $\mathcal{J} = \{\{1, 2, 3\}\}$,

$$\|A\|_{\{1,2,3\}} = \|A\|_{\mathrm{HS}} \leq C\sqrt{n}.$$

For $\mathcal{J} = \{\{1, 2\}, \{3\}\}$, denote by $B_1, \dots, B_n$ the n blocks of d consecutive coordinates in $[nd]$, and by $\|z_B\|_2^2 = \sum_{i\in B} z_i^2$. Then, since $a_{ijk} = 0$ unless i, j, k belong to the same such block,

$$\begin{aligned}
\|A\|_{\{1,2\},\{3\}} &= \sup\left(\sum_{i,j,k=1}^{nd} a_{ijk} y_{ij} z_k : \|Y\|_{\mathrm{HS}} \leq 1, \|z\|_2 \leq 1 \right) \\
&= \sup\left(\sum_{i,j=1}^{nd} \left(\sum_{k=1}^{nd} a_{ijk} z_k \right)^2 : \|z\|_2 \leq 1 \right)^{1/2}
\end{aligned}$$

$$= \sup\left(\sum_{\ell=1}^{n}\sum_{i,j\in B_\ell}\left(\sum_{k\in B_\ell} a_{ijk}z_k\right)^2 : \|z\|_2 \le 1\right)^{1/2}$$
$$\le C\sup\left(\sum_{\ell=1}^{n}\|z_{B_\ell}\|^2 : \|z\|_2 \le 1\right)^{1/2} = C.$$

Similarly $\|A\|_{\{1,3\},\{2\}}, \|A_{\{2,3\},\{1\}}\| \le C$, and we also have $\|A\|_{\{1\},\{2\},\{3\}} \le \|A\|_{\{1,2\},\{3\}} \le C$. Combining with (A.4), $\eta_f(t) \ge c'\min(t^{2/3}, t, t^2/n)$ for a constant $c' > 0$. Then applying Theorem A.2 with $t = n$,

$$\mathbb{P}\left[n^{-1}\Big(f(\varepsilon_1,\dots,\varepsilon_n) - \mathbb{E}[f(\varepsilon_1,\dots,\varepsilon_n)]\Big) \ge 1\right] \le 2e^{-cn^{2/3}}.$$

As $n^{-1}\mathbb{E}[f(\varepsilon_1,\dots,\varepsilon_n)] = C_1$ for a constant $C_1 > 0$, this shows (2.22) for $C_0 = 1 + C_1$.

References

[ABL+18] Emmanuel Abbe, Tamir Bendory, William Leeb, João M Pereira, Nir Sharon, and Amit Singer. Multireference alignment is easier with an aperiodic translation distribution. *IEEE Transactions on Information Theory*, 65(6):3565–3584, 2018.

[APS18] Emmanuel Abbe, João M Pereira, and Amit Singer. Estimation in the group action channel. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 561–565. IEEE, 2018.

[Arn86] V. I. Arnold. Hyperbolic polynomials and Vandermonde's mapping. *Functional Analysis and Applications*, 20:52–53, 1986.

[AS48] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Applied Mathematics Series. U.S. Government Printing Office, 1948.

[AW15] Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3-4):531–586, 2015.

[BBLS18] Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer. Heterogeneous multireference alignment: A single pass approach. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018.

[BBM+17] Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on signal processing*, 66(4):1037–1050, 2017.

[BBS19] Tamir Bendory, Alberto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *arXiv preprint arXiv:1908.00574*, 2019.

[BBS20] Tamir Bendory, Alberto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE Signal Processing Magazine*, 37(2):58–76, 2020.

[BBSK+17] Afonso S Bandeira, Ben Blum-Smith, Joe Kileel, Amelia Perry, Jonathan Weed, and Alexander S Wein. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

[BGPS17] Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak. Rapid solution of the cryo-EM reconstruction problem by frequency marching. *SIAM Journal on Imaging Sciences*, 10(3):1170–1195, 2017.

[BKS20] Sara C. Billey, Matjaž Konvalinka, and Joshua P. Swanson. Asymptotic normality of the major index on standard tableaux. *Adv. in Appl. Math.*, 113:101972, 36, 2020.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[Bru19] Victor-Emmanuel Brunel. Learning rates for Gaussian mixtures under group action. In *Conference on Learning Theory*, pages 471–491, 2019.

[BRW17] Afonso S Bandeira, Philippe Rigollet, and Jonathan Weed. Optimal rates of estimation for multireference alignment. *arXiv preprint arXiv:1702.08546*, 2017.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[CLO92] David Cox, John Little, and Donal O'Shea. Invariant theory of finite groups. In *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, pages 306–344. Springer New York, New York, NY, 1992.

[CM00] Christopher S Coffey and Keith E Muller. Properties of doubly-truncated gamma variables. *Communications in Statistics-Theory and Methods*, 29(4):851–857, 2000.

[DAC+88] Jacques Dubochet, Marc Adrian, Jiin-Ju Chang, Jean-Claude Homo, Jean Lepault, Alasdair W McDowall, and Patrick Schultz. Cryo-electron microscopy of vitrified specimens. *Quarterly reviews of biophysics*, 21(2):129–228, 1988.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[EEB13] Hans Elmlund, Dominika Elmlund, and Samy Bengio. PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure*, 21(8):1299–1306, 2013.

[ER93] Richard Ehrenborg and Gian-Carlo Rota. Apolarity and canonical forms for homogeneous polynomials. *European Journal of Combinatorics*, 14(3):157 – 181, 1993.

[Fra06] Joachim Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, 2006.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[Har92] J. Harris. *Algebraic Geometry: A First Course*. Graduate Texts in Mathematics. Springer, 1992.

[HBC+90] Richard Henderson, J M Baldwin, T A Ceska, F Zemlin, E A Beckmann, and Kenneth H Downing. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of molecular biology*, 213(4):899–929, 1990.

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[JZB+16] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in neural information processing systems*, pages 4116–4124, 2016.

[Kos89] V. P. Kostov. On the geometric properties of Vandermonde's mapping and on the problem of moments. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 112(3-4):203–211, 1989.

[LC06] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[Mac15] I. G. Macdonald. *Symmetric functions and Hall polynomials*. Oxford Classic Texts in the Physical Sciences. The Clarendon Press, Oxford University Press, New York, second edition, 2015. With contribution by A. V. Zelevinsky and a foreword by Richard Stanley, Reprint of the 2008 paperback edition [MR1354144].

[MBM18] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[McC18] Peter McCullagh. *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC, 2018.

[Mit20] Boris Samuilovich Mityagin. The zero set of a real analytic function. *Matematicheskie Zametki*, 107(3):473–475, 2020.

[Nes13] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[PRFB17] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature methods*, 14(3):290, 2017.

[PWB+19] Amelia Perry, Jonathan Weed, Afonso S Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multireference alignment. *SIAM Journal on Mathematics of Data Science*, 1(3):497–517, 2019.

[SDCS10] Fred J. Sigworth, Peter C. Doerschuk, Jose-Maria Carazo, and Sjors H.W. Scheres. Chapter Ten - An introduction to maximum-likelihood methods in Cryo-EM. In Grant J. Jensen, editor, *Cryo-EM, Part B: 3-D Reconstruction*, volume 482 of *Methods in Enzymology*, pages 263 – 294. Academic Press, 2010.

[SGV+07] Sjors HW Scheres, Haixiao Gao, Mikel Valle, Gabor T Herman, Paul PB Eggermont, Joachim Frank, and Jose-Maria Carazo. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature methods*, 4(1):27–29, 2007.

[Sig98] Fred J Sigworth. A maximum-likelihood approach to single-particle image refinement. *Journal of structural biology*, 122(3):328–339, 1998.

[SNRGL+07] Sjors HW Scheres, Rafael Núñez-Ramírez, Yacob Gómez-Llorente, Carmen San Martín, Paul PB Eggermont, and José María Carazo. Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure*, 15(10):1167–1177, 2007.

[SQW15] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[SQW16] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

[SQW18] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[SVN+05] Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology*, 348(1):139–149, 2005.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

[Ver18] Roman Vershynin. *High Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[XHM16] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

E-mail address: zhou.fan@yale.edu

E-mail address: yisun@math.columbia.edu

E-mail address: tianhao.wang@yale.edu

E-mail address: yihong.wu@yale.edu