Proceedings of Machine Learning Research vol 125:1–67, 2020 33rd Annual Conference on Learning Theory

# Learning Polynomials in Few Relevant Dimensions

**Sitan Chen** SITANC@MIT.EDU
*MIT*

**Raghu Meka** RAGHUM@CS.UCLA.EDU *UCLA*

**Editors:** Jacob Abernethy and Shivani Agarwal

## Abstract

Polynomial regression is a basic primitive in learning and statistics. In its most basic form the goal is to fit a degree $d$ polynomial to a response variable $y$ in terms of an $n$-dimensional input vector $x$. This is extremely well-studied with many applications and has sample and runtime complexity $\Theta(n^d)$. Can one achieve better runtime if the intrinsic dimension of the data is much smaller than the ambient dimension $n$?

Concretely, we are given samples $(x, y)$ where $y$ is a degree at most $d$ polynomial in an unknown $r$-dimensional projection (the relevant dimensions) of $x$. This can be seen both as a generalization of phase retrieval and as a special case of learning multi-index models where the link function is an unknown low-degree polynomial. Note that without distributional assumptions, this is at least as hard as junta learning.

In this work we consider the important case where the covariates are Gaussian. We give an algorithm that learns the polynomial within accuracy $\epsilon$ with sample complexity that is roughly $N = O_{r,d}(n \log^2(1/\epsilon)(\log n)^d)$ and runtime $O_{r,d}(Nn^2)$. Prior to our work, no such results were known even for the case of $r = 1$.

We introduce a new *filtered PCA* approach to get a warm start for the true subspace and use *geodesic SGD* to boost to arbitrary accuracy; our techniques may be of independent interest, especially for problems dealing with subspace recovery or analyzing SGD on manifolds.

**Keywords:** polynomial regression, index models, Riemannian optimization, phase retrieval, PCA

## 1. Introduction

Consider the classical *polynomial regression* problem in learning and statistics. In its most basic form, we receive samples of the form $(x, y)$ with $x \in \mathbb{R}^n$ coming from some distribution and $y$ is $P(x)$ for a degree at most $d$ polynomial in $x$. Our goal is to *learn* the polynomial $P$. Here learning could either mean learning the coefficients of $P$ or even finding some other function that gets small prediction error (as in find $Q$ with $E[(Q(x) - P(x))^2] \ll Var(y)$).

Polynomial regression of course is one of the most basic primitives in statistics and machine learning especially in the more general *non-realizable* case. For example, it is crucial in many kernalization applications, and it gives the best known PAC learning algorithms for various central complexity classes such as constant-depth circuits Linial et al. (1993), intersection of halfspaces Klivans et al. (2004), DNFs Klivans and Servedio (2004), convex sets Klivans et al. (2008); Vempala (2010a), the last of which even exploits intrinsic dimension as we do but for a different problem.

The basic bound for polynomial regression is that one can achieve good error with sample complexity and run-time that are $O(n^d)$. This dependence is also necessary (the space of degree $d$

polynomials is of dimension $\approx n^d$) even when $y = P(x)$. But often, such high complexity either in run-time or sample requirements is not feasible for many applications.

This begs the question: can we formulate natural and useful scenarios where one can beat $n^d$ complexity? One such example is the work of Andoni et al. (2014) who study *sparse polynomials* and achieve complexity that is $f(d)poly(n, s)$ where $s$ is sparsity (in a suitable basis).

Motivated by the rich body of work on *phase retrieval* (see, e.g., Candes et al. (2013, 2015); Conca et al. (2015); Netrapalli et al. (2013) and references therein), work on *multi-index models* in learning (see Section 1.2 below) and the above broad question, we study the question of learning polynomials that depend on few relevant dimensions. We call such polynomials *low-rank polynomials*:

**Definition 1** *A degree $d$ polynomial $P : \mathbb{R}^n \to \mathbb{R}$ is of rank $r$ if there exists a degree $d$ polynomial $p : \mathbb{R}^r \to \mathbb{R}$ and vectors $u_1^*, \ldots, u_r^* \in \mathbb{R}^n$ such that $P(x) = p(\langle u_1^*, x\rangle, \langle u_2^*, x\rangle, \ldots, \langle u_r^*, x\rangle)$. We will refer to $p$ as the* link polynomial *and $U^* \triangleq span(u_1^*, \ldots, u_r^*)$ as the* hidden subspace.

In other words, even though the ambient dimension of the polynomial $P$ is $n$, its *intrinsic dimension* is only $r$. If we knew the subspace spanned by $u_1^*, \ldots, u_r^*$, then we could learn $P$ with sample-complexity that does not depend on $n$ at all and run-time that is linear in $n$ (and not $n^d$). Here, there are many natural notions of learning $P$ one could consider. Arguably the two most important goals are 1) to recover the hidden subspace $U^*$ spanned by $u_1^*, \ldots, u_r^*$, and 2) to find a polynomial $q$ that is close to $P$?

Concretely, we are given samples $(x, y)$ where $y = P(x)$ and $P$ is a low-rank polynomial. For most natural distributions $y$, one can show it is information-theoretically possible to learn $P$ with sample-complexity that is only $O_{d,r}(n)$. That is, the dependence on the ambient dimension is only linear. Can we achieve this goal *efficiently*? Henceforth, by efficient we mean that the sample-complexity and run-time are at most some fixed polynomial in $n$ that is of the form $O(f(r, d)n^c)$ for universal constant $c$.

As desirable as the above goal is, it might be too good to be true for general distributions. For example, if $x$ is uniform on the hypercube $\{1, -1\}^n$, then the above question can encode the problem of learning $k$-*juntas*. There, we are given samples $(x, f(x))$ where $x \in_u \{\pm 1\}^n$ and $f$ is a function of at most $k$ variables, and the goal is to recover the indices of the relevant variables. Despite much attention, the best algorithms run in time $n^{\Omega(k)}$, and achieving $f(k)poly(n)$ sample complexity is an outstanding challenge conjectured to be computationally hard Mossel et al. (2003). The connection to rank is that any $k$-junta is a polynomial of rank and degree at most $k$.

Nevertheless, it makes sense to ask the question for other natural distributions. The most basic question in this vein (as we will further motivate later) is the case when $x$ is Gaussian:

**Question 1** *Given samples $(x, y = P(x))$ where $x \sim \mathcal{N}(0, \mathbf{I}_n)$, and $P$ is an unknown degree-$d$, rank-$r$ polynomial, can one approximately recover the subspace defining $P$ efficiently? Can we efficiently approximate $P$? Further, what is the dependence on the error $\epsilon$?*

Note that while we ask the question for isotropic Gaussian covariates, our guarantees immediately carry over to general Gaussians, because the space of low-rank polynomials is affine invariant. Before stating our results, we first briefly discuss different ways of looking at the above question.

**Learning Multi-Index Models** While we motivated the above problem from the context of polynomial regression, an equally valid way to introduce it is from the perspective of learning *multi-index models* in Gaussian space.

Here, we are given samples from a distribution $(x, y)$ where $x \sim \mathcal{N}(0, \mathbf{I}_n)$ and

$$y = g(\langle u_1^*, x\rangle, \langle u_2^*, x\rangle, \ldots, \langle u_r^*, x\rangle),$$

where $g : \mathbb{R}^r \to \mathbb{R}$ is some unknown *link function* and $u_1^*, u_2^*, \ldots, u_r^*$ are unknown orthonormal vectors, and the goal is to learn the subspace $U^*$ spanned by $u_1^*, \ldots, u_r^*$.

The main question we study is the case where the unknown link function $g$ is a low-degree polynomial. Most relevant to the present work is the recent work of Dudeja and Hsu (2018) which we discuss next. There is a tremendous amount of work on learning multi-index models, and we refer to Dudeja and Hsu (2018) for a detailed overview of previous work. Dudeja and Hsu (2018) address the case where $g$ is *smooth* in a Lipschitz sense quantified by a parameter $R$. They show:

1. For *single-index models* (i.e. when $r = 1$): an algorithm that takes $\tilde{O}(n^{O(R^2)}) + n/\epsilon^2)$ samples and computes a direction $u$ that is $\epsilon$-close to the hidden direction.

2. For *multi-index models*: an algorithm that takes $\tilde{O}(n^{O(rR^2)}) + n/\epsilon^2)$ samples and computes a direction $u$ that has at least $1 - \epsilon$ of its $\ell_2$-mass in the span of the unknown $u_1^*, u_2^*, \ldots, u_r^*$.

Firstly, note that while most works on learning multi-index models assume some sort of Lipschitz-smoothness of the link function, polynomials are a natural class of link functions that do not satisfy such smoothness. More importanty, unlike existing works on multi-index models, our main goal is to achieve near-linear sample complexity, run-time scaling with $n^c$ for $c$ independent of $r, d$, and polylogarithmic dependence on the error $\epsilon$.

**Generalizing Phase Retrieval** Further impetus for the above problem comes from the vast literature on phase retrieval. Here, one is given samples of the form $(x, \langle w, x\rangle^2)$ where $x$ is typically Gaussian for most provable guarantees Candes et al. (2013, 2015); Conca et al. (2015); Netrapalli et al. (2013), and the goal is to learn $w$. Besides being natural by itself, the problem is extremely important in practice: as is explained in the references above, in certain physical devices one only observes the *amplitudes* of linear measurements (corresponding to $\langle w, x\rangle^2$) and not the phase. In this setting, the signal and the inputs are taken to be complex but the question is often studied over the reals as well.

Note that the low-rank polynomial in question here is rank 1 and degree 2; moreover the link polynomial $p(z) = z^2$ is even known *a priori*. In this sense, the problem we consider in this work is a substantial generalization, the study of which could potentially lead to new insights for phase retrieval, especially over more general covariate distributions.

**Connections to Tensor Decompositions** Our work also broadly fits in the category of *tensor decompositions*. A $k$-ary tensor in $n$-dimensions is a multi-dimensional array $T \in \mathbb{R}^{[n]^k}$. More relevant to the present work, one can also view a tensor $T$ as a multi-linear map from $T : (\mathbb{R}^n)^k \to \mathbb{R}$ as $T(x^1, x^2, \ldots, x^k) = \sum_{1 \le i_1 \le i_2 \le \cdots \le i_k \le n} T[i_1, i_2, \ldots, i_k] x^1_{i_1} x^2_{i_2} \ldots x^k_{i_k}$. For tensors, the term "rank" has a different meaning: a rank 1 tensor is a tensor of the form $v^1 \otimes \cdots \otimes v^k$, and in general, the rank of a tensor $T$ is the least number of rank one tensors whose sum is $T$.

The basic problem in tensor decomposition is to find a low-rank decomposition of a given tensor. Tensor decomposition algorithms have received a lot of attention recently Anandkumar et al. (2014,

2015); Ge and Ma (2015); Hopkins et al. (2015, 2016); Schramm and Steurer (2017); Ma et al. (2016) with various works studying many different aspects. The connection to our polynomial learning problem comes from the fact that a degree $d$ polynomial can be viewed as a $d$-ary tensor. Moreover, if a polynomial has rank $r$, then the corresponding $d$-ary tensor has rank roughly $O(rd)$.

However, our goals and setting are quite different from those studied in the literature. For one, we are not given access to the tensor directly but only implicitly in the form of evaluations of the symmetric multi-linear form of the tensor on random inputs. Secondly, the central goal for us is to exploit the implicit representation to run in time that is much less than the time to even store the corresponding $d$-ary tensor. As far as we can tell, existing methods for tensor decompositions do not have these properties, at least provably. It is an intriguing question to find further scenarios where one could find tensor decompositions with much better run-time, for instance for constant-rank tensors, when the tensor has a *succinct implicit representation*.

### 1.1. Main Result

Our main result is that we can indeed efficiently learn low-rank polynomials in Gaussian space. To the best of our knowledge, no such results were known even for the rank-1 case. Before stating our result formally, we have to introduce a definition to deal with *degeneracy* in the notion of low-rank.

To understand the issue, consider the example where the link polynomial $p(z_1, z_2) = z_1 + z_2$. Then, if we look at $P(x) = p(\langle w_1^*, x\rangle, \langle w_2^*, x\rangle)$, even though the polynomial is represented as a rank two polynomial, it is really only of rank one and we cannot hope to recover the span of $w_1^*, w_2^*$ but only the span of $w_1^* + w_2^*$. The following is necessary to overcome such non-identifiability issues:

**Definition 2** *(Informal; see Definition 9) A polynomial $P$ is $\alpha$-non-degenerate rank $r$ if $P$ is of rank $r$ and for any $(r-1)$-dimensional subspace $H$, the conditional variance of $P(x)$ given the projection of $x$ onto $H$ is at least $\alpha \cdot Var(p)$.*

Intuitively, there should not be a $(r-1)$-dimensional space that captures all of the variance of $P$. We give an equivalent analytic definition in Section A. Note that any rank-1 polynomial satisfies the condition with $\alpha = 1$.

**Theorem 3** *There exists a universal constant $c_0$ and for all $r, d, \alpha$, there exists $C_0(r, d, \alpha)$ such that the following holds. For all $\delta > 0$ and $\epsilon \in (0, 1)$, there is an efficient algorithm that takes $N = C_0(r, d, \alpha)(\ln(n/\delta))^{c_0 d} \cdot n \log^2(1/\epsilon)$ samples $(x, P(x))$, where $x \sim \mathcal{N}(0, \mathbf{I}_n)$ and $P$ is an unknown $\alpha$-non-degenerate rank $r$, degree-$d$ polynomial defined by hidden subspace $U^*$, and outputs*

1. *Orthonormal $u_1, \ldots, u_r \in \mathbb{S}^{n-1}$ such that $d_P(\text{span}(u_1, \ldots, u_r), U^*) \le \epsilon$*
2. *Degree $d$, $r$-variate polynomial $g$ such that $\mathbb{E}[(y - g(\langle u_1, x\rangle, \ldots, \langle u_r, x\rangle))^2] \le \epsilon \cdot Var(y)$.*

*The run-time of the algorithm is at most $O(r^{c_0 d} N n^2)$.*

This will follow from Theorem 5 and Theorem 34 later in the paper. Here, $d_P(U, U^*)$ denotes the *Procrustes distance* which is one of the standard measures for quantifying distances between subspaces. See Definition 22 for the exact definition.

Note that the run-time of the algorithm is essentially $O_{r,d}(n^3(\log n)^{O(d)})$ — a fixed polynomial in $n$ as desired[1]. The sample complexity is also essentially linear in the ambient dimension $n$ and poly-logarithmic in $1/\epsilon$. No such result was known even for the rank 1 case.

**Remark 4** *A word about the constant $C_0(r, d, \alpha)$ in the theorem. Our proof involves a compactness argument and as a result does not give an explicit upper bound on this quantity. Bounding this comes down to an extremal problem for low-degree polynomials in $r$ variables. For instance for $r = 1$, $C_0(1, d, 1)$ is essentially the inverse of*

$$\sup_{\tau} \inf_{h} (\mathbb{E}[1(|p(g)| > \tau)(g^2 - 1)]),$$

*where $g \sim \mathcal{N}(0, 1)$ and the infimum is over degree $d$ polynomials of variance 1. We believe that this quantity is at least $2^{-Cd^2}$ (as achieved by a suitably scaled degree $d$ Chebyshev polynomial). In general, our arguments can potentially yield a bound of $C(r, d, \alpha) \approx 2^{O(rd)^2}/\alpha^{\Theta(1)}$.*

*Also, we study the noiseless case where $Y = P(X)$. It is possible to modify the first part of our argument (Theorem 5) to get a version tolerant to some noise in $Y$, but we do not focus on this here. In any case, one of our main technical emphases is on getting run-time and sample complexity scaling with poly$(\log(1/\epsilon))$, which would not be possible in the presence of noise.*

### 1.2. Related Work

**Filtering Data by Thresholding** Our algorithm for obtaining a warm start (see Theorem 5) relies on filtering the data via some form of thresholding. This general paradigm has been used in other, unrelated contexts like robustness, see Shen and Sanghavi (2019, 2018); Diakonikolas et al. (2019a); Li (2018b); Diakonikolas et al. (2019b, 2017) and the references therein, though typically the points which are *smaller* than some threshold are removed, whereas our algorithm, TRIMMEDPCA, is an intriguing case where the opposite kind of filter is applied.

**Riemannian Optimization** It is beyond the scope of this paper to reliably survey the vast literature on Riemannian optimization methods, and we refer the reader to the standard references on the subject Udriste (1994); Absil et al. (2009) which mostly provide asymptotic convergence guarantees, as well as the thesis of Boumal Boumal (2014) and the references therein. Some notable lines of work include optimization with respect to orthogonality constraints Edelman et al. (1998), applications to low-rank matrix and tensor completion Mishra et al. (2013); Vandereycken (2013); Ishteva et al. (2011); Kressner et al. (2014), dictionary learning Sun et al. (2016), independent component analysis Shen et al. (2009), canonical correlation analysis Liu et al. (2015), matrix equation solving Vandereycken and Vandewalle (2010), complexity theory and operator scaling Allen-Zhu et al. (2018), subspace tracking Balzano et al. (2010); Zhang and Balzano, and building a theory of geodesically convex optimization Zhang and Sra (2016); Hosseini and Sra (2015); Zhang et al. (2016).

We remark that the update rule we use in our boosting algorithm is very similar to that of Balzano et al. (2010); Zhang and Balzano, as their and our work are based on geodesics on the Grassmannian manifold. That said, they solve a very different problem from ours, and the analysis is quite different.

1. One can save a further factor of $n$ as the $n^3$ comes from computing the top $r$ eigenvectors of a matrix which can be done better, see e.g. Allen-Zhu and Li (2016). We do not belabor this issue here.

**Single/Multi-Index Models and Other Link Functions** As mentioned above, the problem of learning low-rank polynomial is a special case of that of learning a multi-index model, for which there is also a large literature which we cannot hope to cover here. In addition to Dudeja and Hsu (2018) other works include those based on a connection to Stein's lemma Plan et al. (2017); Neykov et al. (2016); Brillinger (2012); Li (1992); Plan and Vershynin (2016); Yang et al. (2017), sliced inverse regression Babichev et al. (2018) as introduced in Li (1991), and gradient-based estimators Hristache et al. (2001b,a); Dalalyan et al. (2008). Other works consider specific link functions or families of link functions:

- $z \mapsto \text{sgn}(z)$, i.e. one-bit compressed sensing Plan and Vershynin (2013); Ai et al. (2012); Gopi et al. (2013).
- $z \mapsto |z|^2$, i.e. phase retrieval Candes et al. (2013, 2015); Conca et al. (2015); Netrapalli et al. (2013).
- $z \mapsto F(z)$ where $F : \mathbb{R}^r \to \mathbb{R}$ is computable by a constant-layer neural network Ge et al. (2017); Bakshi et al. (2018); Janzamin et al. (2015); Ge et al. (2018); Goel et al. (2016); Goel and Klivans (2017).
- $z \mapsto \mathbb{1}\left[\epsilon_i \cdot \text{sgn}(z_i)\ \forall\, i \in [r]\right]$ for signs $\epsilon \in \{\pm 1\}^r$, i.e. intersections of halfspaces Vempala (2010b); Klivans et al. (2009, 2004); Khot and Saket (2008); Vempala (2010a); Diakonikolas et al. (2018).
- $z \mapsto F(z)$ for some function $F : \mathbb{R}^r \to \{0, 1\}$, i.e. subspace juntas Vempala and Xiao (2011); De et al. (2019).

That said, none of the above seem to imply the guarantees for learning low-rank polynomials that we want, namely a run-time that is a fixed polynomial in $n$ and poly-logarithmic in $1/\epsilon$.

## 2. Outline of Algorithm and Analysis

A natural first step is to try to adapt the various techniques from the phase retrieval literature or existing works on multi-index models to the problem. But this seems challenging even for rank 1. For example, the phase retrieval problem corresponds to the polynomial $p(z) = z^2$, which is rather special (see below), and if we don't even know the polynomial, then there are further difficulties. The works on multi-index models such as Dudeja and Hsu (2018) also seem to be difficult to apply off the shelf. For one, they require smoothness of the link function. While it may be possible to circumvent the strict smoothness condition, it seems hard to find useful notions where the smoothness would not grow with the degree, leading to inefficient algorithms.

We present a different line of attack, inspired by ideas of Dudeja and Hsu (2018), Candes et al. (2015), Balzano et al. (2010). Let $P(x) = p(\langle u_1^*, x\rangle, \langle u_2^*, x\rangle, \ldots, \langle u_r^*, x\rangle)$ be the unknown $\alpha$-non-degenerate rank $r$ polynomial. For the remainder of the paper, let $\mathcal{D}$ denote the distribution $(x, y)$ where $x \sim \mathcal{N}(0, \mathbf{I}_n)$ and $y = P(x)$. Let $U^* = \text{span}(u_1^*, \ldots, u_r^*)$ be the hidden subspace. Without loss of generality assume $\text{Var}(y) = 1$.[2]

Our approach has two modular steps:

2. We can do so as our algorithms only need a good lower and upper bound on the variance $y$ which can be obtained easily.

1. **Warm start**: Obtain a "good" approximation to the true subspace $U^*$ by a modified PCA.
2. **Boost accuracy**: Use the subspace computed above as a starting point to boost the accuracy by *Riemannian stochastic gradient descent*.

We next explain the steps at a high-level. The methods to carry out each of the steps could potentially be useful elsewhere especially for problems dealing with subspace recovery.

### 2.1. Getting a Warm Start

The first step is to find a good subspace $V$ of dimension $r$ that $\epsilon$-close to $U^*$ (i.e., $d_P(V, U^*) \leq \epsilon$) in $O_{r,d}(n/\epsilon^2)$ samples. Note that identifying the subspace $U^*$ is the best we can do as the individual directions are not uniquely identifiable.

**Rank-One Case:** To motivate the algorithm, let us first focus on the rank 1 or single-index case. Here $P(x) = p(\langle u^*, x\rangle)$ where $u^* \in \mathbb{S}^{n-1}$ and our goal is to find some $u \in \mathbb{S}^{n-1}$ close to $u^*$.

To do so, we propose a modified PCA by estimating a matrix of the form $M^\phi \equiv E[\phi(y)xx^T] - E[\phi(y)]E[xx^T]$ where $\phi : \mathbb{R} \to \mathbb{R}$ is a suitable "filtering" function. The intuition behind looking at $M^\phi$ is that the matrix has kernel of dimension $n-1$ corresponding to directions orthogonal to $u^*$. Thus, the non-zero eigenvalue of $M^\phi$, if any, could help us approximate or even identify $u^*$.

But what should the function $\phi$ be? For example, for phase retrieval where $P(x) = \langle u^*, x\rangle^2$, taking $\phi(z) = z^2$ suffices. The key issue is that this choice of $\phi$ does not work for general link polynomials. For example, if the link polynomial $p$ is $p(z) = z^2 - 3$, the matrix $M^\phi$ for this particular choice of $\phi$ is identically zero.

We propose overcoming this by instead applying a simple *thresholding* filter for $\phi$. Specifically, for a parameter $\tau > 0$ to be chosen later, let

$$M^\tau \triangleq \mathbb{E}[1(|y| > \tau)(xx^T - I)].$$

We show that for all $d$ there exists $\tau \equiv \tau(d)$ that only depends on $d$ such that $M^\tau$ is a non-zero matrix. Note that this by itself is not enough for our purposes: if the least non-zero eigenvalue of $M^\tau$ were extremely small, then this would affect our sample complexity in estimating $M^\tau$. We show there exists $\tau$ such that $M^\tau$ has an eigenvalue with magnitude at least $\lambda_d > 0$ for some constant depending on $d$ only. As argued before, the corresponding eigenvector is $u^*$. The intuition behind the proof is that conditioning on $|y| > \tau$ makes $x$ more likely to be large in the relevant direction.

The above structural statement is enough to get a warm start for $u^*$ by looking at the empirical approximation of $M^\tau$: for $N$ samples, let

$$\widehat{M^\tau} \triangleq \frac{1}{N}\sum_{i=1}^{N} 1(|y_i| > \tau)(x_i x_i^T - I).$$

We can now use standard matrix concentration inequalities to argue that for $N = O_d(n/\epsilon^2)$ samples, the top eigenvector $\hat{u}$ of $\widehat{M^\tau}$ satisfies $\|u^* - \hat{u}\| \leq \epsilon$.

**Higher-Rank Case:** Extending the above to higher ranks seems much more challenging.

A natural attempt would be to look at a matrix $M^\tau$ as above for a suitable $\tau$. It is once again easy to argue that $M^\tau$ has $n - r$ vectors in its kernel corresponding to the vectors orthogonal to $U^*$. We would now like to say that for some suitable $\tau \equiv \tau(r, d)$, the top $r$ eigenvalues of $M^\tau$ are at

least $\lambda_{r,d}$. If so, we can proceed as before to get an approximation to $U^*$ (the non-zero eigenvectors are in $U^*$). While we can currently show that there is at least one such eigenvalue, we do not know if the matrix $M^\tau$ has rank at least $r$ and it seems considerably more challenging to prove. The difficulty is that unlike the rank 1 case, while conditioning on $|y|> \tau$ should intuitively bias $x$ to have large norm in the relevant directions, it is not clear if it does so in every relevant direction.

Instead, we follow an iterative strategy where we identify one direction at a time in $U^*$. This is similar in spirit to the standard technique of computing the eigenvalues of a matrix by first computing the top eigenvector, projecting it out, and then iterating.

Concretely, suppose we have identified orthonormal vectors $V = \{v_1, v_2, \ldots, v_\ell\}$ for $\ell < r$ that individually have most of their mass in $U^*$. Let $\Pi_{V^\perp}$ be the projection operator onto the space orthogonal to $v_1, \ldots, v_\ell$. Then, to compute the next direction we look at the top eigenvector of

$$M^{\ell,\tau} \triangleq \Pi_{V^\perp} \mathbb{E}[1(|y|> \tau) 1(|\langle v_i, x\rangle| \leq 1,\ \forall i \leq \ell)(xx^T - I)] \Pi_{V^\perp}.$$

As before, we argue that the top eigenvector of the above matrix will have most of its mass in $U^*$ and this gives us our next vector $v_{\ell+1}$. While the sequence of matrices we look at are a bit more complicated, standard random matrix concentration inequalities still allow us to identify the new directions with sample complexity $O_{r,d}(n)$.

In summary, we get the following:

**Theorem 5** *For all $r, d, \alpha$, there exists $C(r, d, \alpha)$ such that the following holds. For all $\delta > 0$ and $\epsilon \in (0, 1)$, there is an efficient algorithm that takes $N = C(r, d, \alpha) n \log(1/\delta)/\epsilon^2$ samples $(x, P(x))$ for $x \sim \mathcal{N}(0, \mathbf{I}_n)$ and unknown $P$ which is $\alpha$-non-degenerate of rank $r$, and outputs a subspace $U$ such that with probability at least $1 - \delta$, $d_P(U, U^*) < \epsilon$. The algorithm runs in time $O(r(Nn^2 + n^3))$.*

### 2.2. Boosting via Geodesic-Based Riemannian Gradient Descent

The results from the previous section give us a way to find a subspace $U$ that is $\epsilon$-close to the true subspace $U^*$ with sample complexity $O_{r,d}(n/\epsilon^2)$.

However, the dependence on $\epsilon$ above is problematic and quite far from what is achievable, e.g., for the special case of phase retrieval. There, results starting with work of Candes et al. (2015) show that one can get *exact* recovery of the unknown direction with sample complexity $\tilde{O}(n)$; in this case, while the sample complexity is $\tilde{O}(n)$, the *run-time* to get within error $\epsilon$ scales with $\log(1/\epsilon)$). In a similar vein, the result of Netrapalli et al. (2013) shows that one can find a vector $w$ that is $\epsilon$-close to the unknown vector with sample-complexity $\tilde{O}(n \log(1/\epsilon))$ and a similar run-time. We address this issue next and give an algorithm that achieves error $\epsilon$ with sample-complexity $\tilde{O}_{r,d}(n \log^2(1/\epsilon))$ and run-time $\tilde{O}_{r,d}(n^2 \log^2(1/\epsilon))$. In the proceeding discussion, we will use some basic terminology from differential geometry in motivating our algorithm, though we emphasize that the algorithm itself is stated solely in terms of matrices, and its proof only involves, e.g., linear algebra and concentration of measure.

First, it is important to understand what fundamentally changes when going from phase retrieval to the more general problem of learning an unknown, low-rank polynomial. At a high level, there are two closely related challenges:

1. **Unknown $r$-variate polynomial**: Unlike in phase retrieval where we know that the link polynomial is $h(z) = z^2$ *a priori*, in our setting we are not given the coefficients of the true

polynomial. The natural workaround is to simply run gradient descent jointly on the space of coefficients and the space of $n \times r$ matrices $V$. As we will see in Section 2.2.1 next, this poses novel difficulties even in the rank-1 case.

2. **Identifiability only up to rotation**: A more fundamental issue is the number of inherent symmetries in the problem, which explodes as $r$ increases. Indeed, there is an infinitely large orbit of parameters $\Theta^* = (\mathbf{c}^*, V^*)$ which give rise to the same underlying low-rank polynomial $P$, parametrized by the group of all rotations of the underlying subspace. Whereas for $r = 1$ it is easy to quotient out most of the symmetries by simply running projected gradient descent on the unit sphere, as we will see in Section 2.2.2, to define the right quotient geometry we will need to run gradient descent on a manifold for which the corresponding optimization landscape is far less straightforward. In addition, as we will see in Section 2.2.3, these symmetries also pose problems for defining and analyzing a suitable progress measure.

In light of 2), it will be good to give a name to the set of parameters $\Theta^* = (\mathbf{c}^*, V^*)$ which correspond to the underlying low-rank polynomial.

**Definition 6** *For a collection of coefficients $\mathbf{c}^*$ of a degree-$d$ $r$-variate polynomial, and a column-orthonormal matrix $V^* \in \mathbb{R}^{n\times r}$, we say that the parameters $\Theta^* = (\mathbf{c}^*, V^*)$ are a* realization *of $\mathcal{D}$ if the polynomial $p_*(z) \triangleq \sum_I c_I^* \phi_I(z)$ satisfies $P(x) = p_*(V^{*\top} x)$ for all $x \in \mathbb{R}^n$, where $\{\phi_I\}$ are the (normalized) tensor-product Hermite polynomials of degree at most $d$ over $r$ variables (see Section A.3).*

### 2.2.1. NOT KNOWING THE POLYNOMIAL: A TOY CALCULATION

The issue of not knowing $p$ manifests even in the $r = 1$ case. Below, we examine at a high level where the calculations for analyzing gradient descent for phase retrieval break down for us.

Let us try to imitate the approach of Candes et al. (2015). Let $\Theta^* = (\mathbf{c}^*, v^*)$ be one of the two possible realizations of $\mathcal{D}$ for which $v^* \in \mathbb{S}^{n-1}$, and suppose we already have a warm start of $\Theta = (\mathbf{c}, v)$, where the coefficients $\mathbf{c}$ and $\mathbf{c}^*$ define the univariate degree-$d$ polynomials $p(z) \triangleq \sum_{i=1}^d c_i \phi_i(z)$ and $p_*(z) \triangleq \sum_{i=1}^d c_i^* \phi_i(z)$ respectively. Given samples $(x^1, y^1), ..., (x^N, y^N) \sim \mathcal{D}$, a natural approach would be to analyze vanilla gradient descent over $\mathbb{R}^{d+1} \times \mathbb{R}^n$ for the empirical risk

$$L(\Theta) \triangleq \frac{1}{N} \sum_{i=1}^N (F_{x^i}(\Theta) - y_i)^2 \qquad \text{for} \qquad F_x(\Theta) \triangleq p(V^\top x).$$

To show that this converges linearly from a warm start, the first thing to show would be that the negative gradient at $\Theta$ is correlated with the direction in which we would like to move, a property that sometimes goes under the name *local curvature*. Noting that $\frac{1}{2}\nabla L(\Theta) = \frac{1}{N}\sum_{i=1}^N (F_{x^i}(\Theta) - F_{x^i}(\Theta^*)) \cdot \nabla F(x^i)(\Theta)$, using the fact that we initialize at a warm start in order to linearly approximate $F_x(\Theta) - F_x(\Theta^*)$ by $\nabla F_x(\Theta^*) \cdot \langle \Theta - \Theta^* \rangle$, and explicitly computing the gradient of $F_x$ (see Proposition 36), one can check that

$$\begin{aligned}\left\langle \frac{1}{2}\nabla L(\Theta), \Theta - \Theta^* \right\rangle &\approx \frac{1}{N}\sum_{i=1}^N \langle \nabla F_{x^i}(\Theta^*), \Theta - \Theta^* \rangle^2 \\ &= \frac{1}{N}\sum_{i=1}^N \left[ \langle v - v^*, x^i \rangle \cdot p_*'(\langle v^*, x^i \rangle) + (p - p_*)(\langle v^*, x^i \rangle) \right]^2.\end{aligned}$$

The expectation of this quantity is

$$\mu \triangleq \mathbb{E}_g \left[ \left( \langle v - v^*, g \rangle \cdot p_*'(\langle v^*, g \rangle) + (p - p_*)(\langle v^*, g \rangle) \right)^2 \right]$$

Write $v - v^* = \alpha \cdot v^* + \beta \cdot v^\perp$ for $v^\perp \in \mathbb{S}^{n-1}$ orthogonal to $v^*$, where $\alpha = \langle v, v^* \rangle - 1 \approx -\|v - v^*\|_2^2$. By some elementary calculations which we omit here, one can show that

$$\mu = \beta^2 \cdot \mathbb{E}[p_*'(x)^2] + \sum_{\ell=0}^{d} \left( (\alpha\ell + 1) \cdot c_\ell + a\sqrt{(\ell+1)(\ell+2)} \cdot c_{\ell+2} - c_\ell^* \right)^2. \tag{1}$$

In the case of phase retrieval, $p(z) = p_*(z) = z^2 = \sqrt{2} + \sqrt{2} \cdot \phi_2(z)$, so $\mathbf{c} = \mathbf{c}^* = (\sqrt{2}, 0, \sqrt{2})$ and we simply get that

$$\mu = 12\alpha^2 + 4\beta^2 \geq 4\|v - v^*\|_2^2.$$

In other words, the correlation between the negative gradient and the residual direction $v^* - v$ in which we would like to go is positive and scales with the squared norm of the residual. This simple calculation lies at the heart of the proof that vanilla gradient descent converges linearly to $v^*$ from a warm start for phrase retrieval.

More generally, if $\mathbf{c}^* = \mathbf{c}$, then the quantity in (1) will enjoy this positive scaling with $\|v^* - v\|_2^2$, and one can also show linear convergence of vanilla gradient descent. But it is apparent that when $\mathbf{c}^* \neq \mathbf{c}$, $\mu$ can be arbitrarily close to zero, e.g. by taking $\beta$ to be much smaller than $\alpha$. So when $\mathbf{c}^* \neq \mathbf{c}$, we may get stuck at spurious infinitesimal-curvature points of the optimization landscape and fail to make sufficient progress in a single step.

The basic underlying issue is simply that vanilla gradient steps can move us in unhelpful directions, e.g. we might end up moving mostly in the direction of $v$ when we should be moving in directions orthogonal to $v$. And whereas this evidently does not pose an issue when $\mathbf{c} = \mathbf{c}^*$, which corresponds to the case where we know the underlying polynomial and only need to run gradient descent to learn the hidden direction, in the case where $\mathbf{c} \neq \mathbf{c}^*$ and we must run gradient descent jointly on $v$ and $\mathbf{c}$, the usual analysis of vanilla gradient descent fails.

#### 2.2.2. NON-IDENTIFIABILITY: WHICH SPACE TO RUN SGD IN?

The workaround for the issue posed in Section 2.2.1 is clear at least in the rank-1 case: to avoid moving in the wasteful directions which are orthogonal to the current iterate $v$, simply compute the vanilla gradient and project to the orthogonal complement of $v$. We would also like to ensure that our iterates themselves are unit vectors like $v^*$, so the following two-step update rule would suffice: 1) walk against the projected gradient and then 2) project back to $\mathbb{S}^{n-1}$. In fact, one can show that this algorithm actually achieves linear convergence for learning arbitrary unknown rank-1 polynomials.

It turns out there is a principled way to extend this approach to higher rank. Indeed, the above mentioned projected gradient scheme is nothing more than (retraction-based) gradient descent on the Riemannian manifold $\mathbb{S}^{n-1}$: the orthogonal complement of $v$ is precisely the tangent space of $\mathbb{S}^{n-1}$ at $v$, and the projection back to $\mathbb{S}^{n-1}$ is a special instance of a *retraction*, roughly speaking a continuous mapping from the tangent spaces of a manifold back onto the manifold itself. We do not attempt to define these notions formally, referring the reader to, e.g. Absil et al. (2009).

The rank-$r$ analogue of $\mathbb{S}^{n-1}$ is the Grassmannian $\mathrm{G}(n, r)$ of $r$-dimensional subspaces of $\mathbb{R}^n$. However, while various retraction operations, e.g. via QR decomposition, can be constructed,

retraction-based Riemannian optimization is somewhat more difficult to analyze in our setting. Instead, we appeal to an alternative formulation of Riemannian gradient descent via geodesics.

Roughly, geodesics are acceleration-free curves on a manifold determined solely by their initial position on the manifold, initial velocity, and length. Gradient descent on a Riemannian manifold $\mathcal{M}$ via geodesics is then very simple to formulate: at an iterate $p \in \mathcal{M}$, 1) compute the gradient $\nabla$ after projecting to the tangent space at $p$, 2) walk along the geodesic that starts at $p$ and has initial velocity $\nabla$ and length $\eta$, where $\eta$ is the learning rate.

We now see what this would yield in our setting. Let $\Theta = (\mathbf{c}, V)$ be an iterate. For now, we will keep $\mathbf{c}$ fixed and describe how to update $V$, regarded as a column-orthonormal $n \times r$ matrix of basis vectors for the subspace $V$, by following the appropriate geodesic on $\mathrm{G}(n, r)$. Given a single sample $(x, y)$, define the single-sample empirical risk $L_x^{\mathbf{c}}(V) = (F_x(\Theta) - y)^2$. Let $\nabla L_x^{\mathbf{c}}(V) \in \mathbb{R}^{n \times r}$ be the vanilla gradient, where $L_x^{\mathbf{c}}(V) \triangleq L_x(\Theta)$. It turns out its projection to the tangent space at $V$ is simply $\nabla \triangleq \Pi_V^{\perp} \cdot \nabla L_x^{\mathbf{c}}(V) \in \mathbb{R}^{n \times r}$, where $\Pi_V^{\perp}$ denotes projection to the orthogonal complement of $V$ (note that this is a natural generalization of the tangent spaces for $\mathbb{S}^{n-1}$).

The geodesic $\Gamma$ with initial point $V$ and velocity $\nabla$, and length $\eta$ has a simple closed form in terms of the SVD of $\nabla$, which is made even simpler by the fact that in our setting, $\nabla$ turns out to be rank-1. We defer the details of the exact update, which can be computed in time $O(n)$, to Section C.

### 2.2.3. TRACKING PROGRESS IN BOTH $\mathbf{c}$ AND $V$

In the previous section we sketched our approach for updating our estimate $V$ for the subspace given an estimate $\mathbf{c}$ for the coefficients of the polynomial, but did not explain how to update $\mathbf{c}$. As $\mathbf{c}$ just lives in Euclidean space, we can simply update $\mathbf{c}$ to some $\mathbf{c}'$ via vanilla gradient descent on $L^V$, where $L^V(\mathbf{c}) \triangleq L(\Theta)$, and this is the approach we take.

To analyze such an approach, one would want to show that each step $(\mathbf{c}, V) \mapsto (\mathbf{c}', V')$ contracts some suitably defined progress measure. Indeed, the natural progress measure one could try analyzing is

$$\inf_{(\mathbf{c}^*, V^*) \text{ realizing } \mathcal{D}} \|\mathbf{c} - \mathbf{c}^*\|_2^2 + \|V - V^*\|_F^2. \tag{2}$$

The key difficulty here is that the minimizing realization $(\mathbf{c}^*, V^*)$ could change with each new iterate, and tracking how this changes is tricky as there is no clean non-variational proxy for (2).

Our workaround is to have our boosting algorithm alternate between two phases. For an iterate $V \in \mathbb{R}^{n \times r}$, we run the following algorithm, GEOSGD, which alternates between two phases: 1) recomputing a good $\mathbf{c}$, and 2) updating $V$ using that $\mathbf{c}$. An informal specification of this algorithm is given in Algorithm 1 below.

We will defer an exact specification of GEOSGD and the subroutines REALIGNPOLYNOMIAL and SUBSPACEDESCENT until Section C.

To analyze this scheme, rather than track progress in (2) we can simply track progress in $d_P(V, V^*) = \inf_{V^*} \|V - V^*\|_F^2$, where $V^*$ ranges over $n \times r$ matrices whose columns form an orthonormal basis for the true subspace. This progress measure is, up to constants, simply the Procrustes distance between our current subspace $V$ and the true subspace $V^*$, and can be approximated by the *chordal distance* which has a simple closed-form expression amenable to analysis.

Roughly, we will show the following:

**Algorithm 1:** GEOSGD (informal)

**Input:** Sample access to $\mathcal{D}$, warm start $V^{(0)} \in \mathbb{R}^{n \times d}$, target error $\epsilon$, failure probability $\delta$
**Output:** Estimate $(\mathbf{c}^{(T)}, V^{(T)})$ which is $\epsilon$-close to a realization of $\mathcal{D}$

**for** $0 \leq t < T$ **do**
  Run REALIGNPOLYNOMIAL using $V^{(t)}$. That is, draw samples and run vanilla gradient descent with respect to empirical risk $L^{V^{(t)}}$ over those samples to produce $\mathbf{c}^{(t)}$ which approximates the "best" choice of $\mathbf{c}$ given fixed $V^{(t)}$.
  Run SUBSPACEDESCENT initialized to $V^{(t)}$ and using $\mathbf{c}^{(t)}$. That is, draw samples and, starting from $V^{(t)}$, run a small step of geodesic gradient descent with respect to empirical risk $L_x^{\mathbf{c}}$ for each of those samples $x$. Call the result $V^{(t+1)}$
**end**
Output $V^{(T)}$.

**Theorem 7 (Informal, see Theorem 38)** *If $V$ is sufficiently close to the true subspace in Procustes distance, then running* REALIGNPOLYNOMIAL *using $V$ will yield $\mathbf{c}$ such that for the realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$ where $d_P(V, V^*) = \|V - V^*\|_F$, $\|\mathbf{c} - \mathbf{c}^*\|_2 \approx d_P(V - V^*)$.*

**Theorem 8 (Informal, see Theorem 52)** *If $V$ is sufficiently close to the true subspace in Procustes distance, If $V$ and $\mathbf{c}$ are such that $\|\mathbf{c} - \mathbf{c}^*\|_2 \approx d_P(V - V^*)$for the realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$ where where $d_P(V, V^*) = \|V - V^*\|_F$, then running* SUBSPACEDESCENT *initialized to $V$ and using $\mathbf{c}$ will yield $V'$ so that the progress measure $d_P(V, V^*)$ contracts by a factor of $1 - \tilde{O}_{r,d}(1/n)$.*

Having defined the "right" gradient descent subroutines, the proofs of Theorems 7 and 8 will be based on showing the same kind of estimates alluded to in Section 2.2.1. That is, for instance we must show that the steps in both subroutines have good correlation with the direction in which we want to go. Showing this holds with high probability will then entail exhibiting the appropriate second moment bounds. In the case of Theorem 7, we can then invoke standard hypercontractivity-based tail bounds to show concentration. In the case of Theorem 8, concentration will be more delicate as each small step of SUBSPACEDESCENT will be a geodesic gradient step with respect to a *single-sample* empirical risk $L_x^{\mathbf{c}}$. For the analysis to be doable, it is crucial that these risks be single-sample so that the geodesic steps are *rank-one* updates. But then, to show concentration over a sequence of small geodesic steps, we must invoke non-standard martingale concentration inequalities, see Section A.2.2. Intuitively, if we take the sizes of these small steps to scale with $O(1/T)$, the corresponding martingale does not move away from its starting point by too much, and the sum of the martingale differences ends up behaving more or less like a sum of iid random variables (see the beginning of Section E.2.2). We refer the reader to Sections D and E for the complete proofs of Theorems 7 and 8 respectively.

## References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *arXiv preprint arXiv:1208.6279*, 2012.

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pages 974–982, 2016.

Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 172–181, 2018.

Anima Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *arXiv preprint arXiv:1411.1488*, 2014.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, pages 36–112, 2015.

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning sparse polynomial functions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 500–510. SIAM, 2014.

Dmitry Babichev, Francis Bach, et al. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1):1507–1543, 2018.

Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. *arXiv preprint arXiv:1811.01885*, 2018.

Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *2010 48th Annual allerton conference on communication, control, and computing (Allerton)*, pages 704–711. IEEE, 2010.

V Bentkus. An inequality for tail probabilities of martingales with differences bounded from one side. *Journal of Theoretical Probability*, 16(1):161–173, 2003.

Nicolas Boumal. *Optimization and estimation on manifolds.* PhD thesis, 2014.

David R Brillinger. A generalized linear model with "gaussian" regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Aldo Conca, Dan Edidin, Milena Hering, and Cynthia Vinzant. An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346–356, 2015.

Arnak S Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(Aug):1647–1678, 2008.

Anindya De, Elchanan Mossel, and Joe Neeman. Is your function low dimensional? In *Conference on Learning Theory*, pages 979–993, 2019.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019a.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606, 2019b.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1887–1930. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/dudeja18a.html.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. *arXiv preprint arXiv:1810.06793*, 2018.

Surbhi Goel and Adam Klivans. Learning neural networks with two nonlinear layers in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *International Conference on Machine Learning*, pages 154–162, 2013.

Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.

Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 910–918, 2015.

Marian Hristache, Anatoli Juditsky, Jörg Polzehl, Vladimir Spokoiny, et al. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001a.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001b.

Mariya Ishteva, P-A Absil, Sabine Van Huffel, and Lieven De Lathauwer. Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

Subhash Khot and Rishi Saket. On hardness of learning intersection of two halfspaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 345–354, 2008.

Adam R Klivans and Rocco A Servedio. Learning dnf in time 2o (n1/3). *Journal of Computer and System Sciences*, 68(2):303–318, 2004.

Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004.

Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning geometric concepts via gaussian surface area. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550. IEEE, 2008.

Adam R Klivans, Philip M Long, and Alex K Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 588–600. Springer, 2009.

Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

Chris Junchi Li. A note on concentration inequality for vector-valued martingales with weak exponential-type tails. *arXiv preprint arXiv:1809.02495*, 2018a.

Jerry Zheng Li. *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology, 2018b.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420): 1025–1039, 1992.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

Xin-Guo Liu, Xue-Feng Wang, and Wei-Guo Wang. Maximization of matrix trace function of product stiefel manifolds. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1489–1506, 2015.

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016.

Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 206–212, 2003.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

Matey Neykov, Zhaoran Wang, and Han Liu. Agnostic estimation for misspecified phase retrieval models. In *Advances in Neural Information Processing Systems*, pages 4089–4097, 2016.

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In *Conference on Learning Theory*, pages 1760–1793, 2017.

Hao Shen, Stefanie Jegelka, and Arthur Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57(9):3498–3511, 2009.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. *arXiv preprint arXiv:1810.11874*, 2018.

Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. In *Advances in Neural Information Processing Systems*, pages 6076–6086, 2019.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016.

Constantin Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 1994.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Bart Vandereycken and Stefan Vandewalle. A riemannian optimization approach for computing low-rank solutions of lyapunov equations. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2553–2579, 2010.

Santosh S Vempala. Learning convex concepts from gaussian distributions with pca. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 124–130. IEEE, 2010a.

Santosh S Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM (JACM)*, 57(6):1–14, 2010b.

Santosh S Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *arXiv preprint arXiv:1108.3329*, 2011.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Van H Vu. Concentration of non-lipschitz functions and applications. *Random Structures & Algorithms*, 20(3):262–316, 2002.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. On stein's identity and near-optimal estimation in high-dimensional index models. *arXiv preprint arXiv:1709.08795*, 2017.

Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600, 2016.

**Organization** In Section A we introduce notation and miscellaneous technical facts that we will use in our proofs. In Section B, we give our algorithm TRIMMEDPCA for obtaining a warm start. In Section C, we give the formal specification for our boosting algorithm GEOSGD, and in Sections D and E we prove guarantees for its key subroutines. We complete the proof of correctness for GEOSGD in Section F. In Section G we give the martingale concentration inequalities we will need, and in Section H and Section I we complete proofs deferred from the body of the paper.

## Appendix A. Notations and Preliminaries

Throughout, $n$ will denote the ambient dimension, $r$ the rank of the polynomial, and $d$ the degree.

Given a vector space $U$, let $\Pi_U$ denote the orthogonal projection operator onto $U$, and let $U^\perp$ denote the orthogonal complement of $U$. We will often abuse notation and also use $U$ to refer to a set of column vectors $\{u_1, \ldots, u_\ell\}$, in which case $\text{span}(U)$ denotes the span of these vectors, $\Pi_U$ denotes $\Pi_{\text{span}(U)}$, and $\Pi_{U^\perp}$ denotes $\Pi_{\text{span}(U)^\perp}$. Given vector spaces $U \subset V$, $V \backslash U = V \cap U^\perp$ denotes the orthogonal complement of $U$ in $V$. Given $v \in \mathbb{S}^{d-1}$, we will use $\Pi_v \triangleq vv^\top$ and $\Pi_v^\perp \triangleq \mathbf{I} - vv^{top}$ to denote projection to the span of $v$ and its orthogonal complement, respectively. More generally, given $V \in \mathbb{R}^{n \times r}$ whose columns are orthonormal, we will use $\Pi_V \triangleq VV^\top$ and $\Pi_V^\perp \triangleq \mathbf{I} - VV^\top$ to denote projection to the span of the columns of $V$ and its orthogonal complement, respectively.

Given matrix $M \in \mathbb{R}^{m \times n}$, let $\|M\|_F$ denote its Frobenius norm, and $\|M\|_2$ its operator norm.

Let $\text{St}_r^n$ denote the *Stiefel manifold* of $n \times r$ matrices with orthonormal columns, and let $\text{G}(n, r)$ denote the *Grassmannian* of $r$-dimension subspaces of $n$. $\text{G}(n, r)$ can be regarded as the quotient of $\text{St}_r^n$ under the natural action of $O(r)$, that is, given any subspace $U \in \text{G}(n, r)$ and any $V \in \text{St}_r^n$ whose columns form a basis for $U$, we can associate $U$ to the equivalence class $[V] \triangleq \{V \cdot O : O \in O(r)\}$.

For $r > 0$, let $\mathcal{B}_r^n \subset \mathbb{R}^n$ denote the Euclidean ball of radius $r$ centered at the origin. When the context is clear, we will suppress the superscript $n$.

For polynomial $p : \mathbb{R}^r \to \mathbb{R}$, define $\text{Var}[p] = \mathbb{E}[(p - \mathbb{E}[p])^2]$. Given indices $\mathbf{j} \triangleq (j_1, ..., j_\ell) \in [r]^\ell$, and $z \in \mathbb{R}^r$ we will use the shorthand

$$\text{D}_{\mathbf{j}}\, p(z) \triangleq \frac{\partial}{\partial z_{j_1} \cdots \partial z_{j_\ell}} p(z). \tag{3}$$

Similarly, for $F : \mathbb{R}^{n \times r} \to \mathbb{R}$, indices $\mathbf{i} \in [n]^\ell$ and $\mathbf{j} \in [r]^\ell$, and $V \in \mathbb{R}^{n \times r}$, we will use the shorthand

$$\text{D}_{\mathbf{i},\mathbf{j}}\, F(V) \triangleq \frac{\partial}{\partial V_{i_1,j_1} \cdots \partial V_{i_\ell,j_\ell}} F(V). \tag{4}$$

### A.1. Non-degeneracy

Recall the notion of $\alpha$-non-degenrate rank $r$ polynomials introduced in Definition 2. While that notion is intuitive, it is less amenable to analysis. It turns out that the notion is essentially equivalent (up to scaling $\alpha$ by $d$) to the following and we will use this going forward.

**Definition 9** *A polynomial $h : \mathbb{R}^r \to \mathbb{R}$ is $\alpha$ non-degenerate if $M = \mathbb{E}_{g \sim \mathcal{N}(0, \mathbf{I}_r)} \left[\nabla h(g) \nabla h(g)^\top\right]$ satisifes $M \succeq \alpha \cdot \|M\|_2 \mathbf{I}_r$.*

*We say a rank $r$ polynomial $P : \mathbb{R}^n \to \mathbb{R}$ is $\alpha$ non-degenerate if $P$ is non-degenerate in the $r$-dimensional space corresponding to the relevant directions. That is, there exist orthonormal vectors $u_1, \ldots, u_r$ such that $P(x) = h(\langle u_1, x\rangle, \ldots, \langle u_r, x\rangle)$ and $h$ is $\alpha$ non-degenerate.*

While it is not clear immediately from the definition, the notion above does not depend on the specific basis chosen. Henceforth, fix constant $\alpha_{\mathsf{ndg}} > 0$. we will let $\mathcal{P}_{n,r,d}^{\alpha_{\mathsf{ndg}}}$ denote the set of all $\alpha_{\mathsf{ndg}}$ non-degenerate rank $r$ polynomials $P$ of degree at most $d$ in $n$ variables that satisfy the normalization conditions $\mathbb{E}_{X \sim \mathcal{N}(0, \mathbf{I}_n)}[P(X)] = 0$ and $\mathbb{E}_{g \sim \mathcal{N}(0, \mathbf{I}_r)} \left[\nabla h(g) \nabla h(g)^\top\right] \preceq \mathbf{I}_n$. We write $\mathcal{P}_{r,d}^{\alpha_{\mathsf{ndg}}}$ for $\mathcal{P}_{r,r,d}^{\alpha_{\mathsf{ndg}}}$.

Finally, we will use the following elementary property of non-degeneracy.

**Fact 1** *If $P \in \mathcal{P}^{\alpha_{\mathsf{ndg}}}_{n,r,d}$, then $\alpha_{\mathsf{ndg}}/d \leq Var[P(X)] \leq r$.*

**Proof** It suffices to consider $n = r$. For the upper bound, we have $\mathrm{Var}[P] \leq \mathbb{E}_g \left[\|\nabla p_*(g)\|_2^2\right] \leq r$ by taking traces in the definition of non-degeneracy and invoking Lemma 19 below.

For the lower bound, we have $\mathrm{Var}[P] \geq \mathbb{E}_g \left[\|\nabla p_*(g)\|_2^2\right]/rd \geq \alpha_{\mathsf{ndg}}/d$ by taking traces and invoking Lemma 18 below. ■

### A.2. Concentration Inequalities

In this section we record some concentration inequalities. Let $\zeta_1, ..., \zeta_T$ be independent atom variables which each take values in Euclidean space.

#### A.2.1. STANDARD CONCENTRATION

We will need the following matrix concentration inequality in our analysis of TRIMMEDPCA.

**Lemma 10 (Vershynin (2010))** *Let $\phi : \mathbb{R} \to [0, 1]$ be any function. Let $M = \mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{I}_n)}[\phi(x) \cdot (xx^\top - \mathbf{I})]$. If $x_1, ..., x_N \sim \mathcal{N}(0, \mathbf{I}_n)$ for $N = \Omega(\{n \vee \log(1/\delta)\}/\epsilon^2)$, then*

$$\Pr\left[\left\|\mathbf{M} - \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\cdot(x_ix_i^\top - \mathbf{I})\right\|_2 \geq \epsilon\right] \leq \delta.$$

**Proof** This follows from standard sub-Gaussian concentration; see e.g. Remark 5.40 in Vershynin (2010). ■

In our analysis of GEOSGD, we will also need the following standard consequence of Fact 6.

**Lemma 11** *Let $Z_1, ..., Z_T$ be iid scalar random variables which are each given by polynomials of degree $d$ in $\zeta_1, ..., \zeta_T$ respectively. If $Var[Z] \leq \sigma^2$ for each $i \in [T]$, then then for any $t > 0$,*

$$\Pr\left[\left|\frac{1}{T}\sum_{i=1}^{T}(Z_i - \mathbb{E}[Z_i])\right| \geq \frac{1}{\sqrt{T}}\cdot O(\log(1/\delta))^{d/2}\cdot\sigma\right] \leq \delta.$$

Additionally, we will need the following concentration inequality for sums of random variables which only satisfy one-sided bounds. This is a specialization of the martingale concentration result of Bentkus (2003) to the iid case, though we also need that result in its full generality for Lemma 14 below.

**Lemma 12 (Special case of Bentkus (2003))** *Let $Z_1, ..., Z_T$ be iid, mean-zero random variables. Let $c, s > 0$ be deterministic constants for which $Z_i \leq c$ with probability one and $Var[Z_i] \leq s^2$ for all $i \in [T]$. Let $\sigma = c \vee s$. Then for any $\delta > 0$,*

$$\Pr\left[\frac{1}{T}\sum_{i=1}^{T}Z_i \geq \frac{1}{\sqrt{T}}\cdot\sqrt{2}\log(1/\delta)\cdot\sigma\right] \leq \delta.$$

### A.2.2. MARTINGALE CONCENTRATION

We now generalize the two scalar concentration inequalities of Section A.2.1 to the martingale setting. In this section, let $Y(\zeta_1, ..., \zeta_T)$ be a real-valued random variable depending on the atom variables $\zeta_1, ..., \zeta_T$ which each take values in Euclidean space. Define the martingale differences $Z_i(\zeta) \triangleq \mathbb{E}[Y|\zeta_1, ..., \zeta_i] - \mathbb{E}[Y|\zeta_1, ..., \zeta_{i-1}]$. When the context is clear, we will suppress the parenthetical $\zeta$. For brevity, we will use the acronym MDS throughout to refer to martingale difference sequences.

The first lemma is the martingale analogue of Lemma 11, with the slight twist that the moment bounds only hold with high probability. The bounds are slightly weaker than those of Lemma 11 but will suffice for our applications.

**Lemma 13** *There is a constant $c_1 > 0$ for which the following holds. Let $\sigma > 0$, and suppose the atom variables $\zeta_1, ..., \zeta_T$ each take values in $\mathbb{R}^n$, and suppose the martingale differences $\{Z_i\}$ are such that for any realization of $\zeta_1, ..., \zeta_{i-1}$, $Z_i(\zeta)$ is a polynomial of degree at most $d$ in $\zeta_i$, and moreover $\Pr\left[\mathbb{E}[Z_i^2|\zeta_1, ..., \zeta_{i-1}] \le \sigma^2\right] \ge 1 - \beta$ for each $i \in [T]$. Then for any $t > 0$,*

$$\Pr\left[\max_{\ell \in [T]} \left|\sum_{i=1}^{\ell} Z_i\right| \ge (2\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot \sigma\right] \le \delta + T \cdot \beta.$$

The second lemma is the martingale analogue of Lemma 12, again with the twist that the bounds on the differences only hold with high probability.

**Lemma 14** *Let $\{c_i\}_{i\in[T]}$ and $\{s_i\}_{i\in[T]}$ be collections of positive constants, and let $\mathcal{E}_i$ be the event that $Z_i \le c_i$ and $\mathbb{E}[Z_i^2|\zeta_1, ..., \zeta_{i-1}] \le s_i^2$. Let $\sigma_i = c_i \vee s_i$, and define $\sigma^2 = \sum_i \sigma_i^2$. Then if $\Pr[\mathcal{E}_i|\zeta_1, ..., \zeta_{i-1}] \ge 1 - \beta$ for each $i \in [T]$, then for any $\delta > 0$,*

$$\Pr\left[\sum_{i=1}^{T} Z_i \ge \sqrt{2}\log(1/\delta) \cdot \sigma\right] \le \delta + T \cdot \beta.$$

## A.3. Hermite Polynomials and Gradients

For every $\ell \in \mathbb{Z}_{\ge 0}$, define the oscillator $\phi_\ell(z) = \frac{1}{(\ell!)^{1/2}}\mathrm{He}_\ell(z)$, where $\mathrm{He}_\ell$ is the degree-$\ell$ (probabilist's) Hermite polynomial. $\{\phi_\ell(z)\}$ forms an orthonormal basis for $L_2(\mathbb{R})$ with respect to the Gaussian inner product.

The following elementary identities will be useful.

**Fact 2 (Derivatives)** *For any $0 \le i \le k$, $\phi_k^{[i]} = \sqrt{\frac{k!}{(k-i)!}} \cdot \phi_{k-i}$.*

**Fact 3 (Recurrence Relation)** *For any $k \ge 0$, $x \cdot \phi_k(x) = \sqrt{k+1}\phi_{k+1}(x) + \sqrt{k} \cdot \phi_{k-1}(x)$.*

**Fact 4 (Linearization Coefficients)** *For any $a, b, c \in \mathbb{Z}_{\ge 0}$ such that $a+b \ge c$, $a+c \ge b$, $b+c \ge a$, and $a + b + c$ is even.*

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\phi_a(g)\phi_b(g)\phi_c(g)\right] = \frac{\sqrt{a! \cdot b! \cdot c!}}{\left(\frac{a+b-c}{2}\right)! \cdot \left(\frac{a-b+c}{2}\right)! \cdot \left(\frac{-a+b+c}{2}\right)!}$$

*For all other $a, b, c$, this quantity is zero.*

**Corollary 15** *For any $0 \le a \le b$,*

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g \cdot \phi_a(g)\phi_b(g)] = \mathbb{1}\left[b = a+1\right] \cdot \sqrt{a+1}$$

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}[\phi_2(g)\phi_a(g)\phi_b(g)] = \mathbb{1}\left[b = a+2\right] \cdot \sqrt{\frac{(a+1)(a+2)}{2}} + \mathbb{1}\left[b = a\right] \cdot a\sqrt{2}$$

**Fact 5** *For any $v, v' \in \mathbb{S}^{n-1}$ and $\ell, \ell' \in \mathbb{Z}_{\ge 0}$,*

$$\mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_n)}\left[\phi_\ell(\langle v, g\rangle \cdot \phi_{\ell'}(\langle v', g\rangle)\right] = \mathbb{1}\left[\ell = \ell'\right] \cdot \langle v, v'\rangle^\ell.$$

We also record some basic facts about gradients and moments of polynomials in Gaussians.

**Fact 6 (Hypercontractivity)** *For a polynomial $f : \mathbb{R}^r \to \mathbb{R}$ of degree $d$, and integer $q \ge 1$,*

$$\mathbb{E}[f(g)^q]^{1/q} \le (q-1)^{d/2}\mathbb{E}[f(g)^2]^{1/2},$$

*where the expectation is over $g \sim \mathcal{N}(0,1)$.*

**Corollary 16** *For any integer $q \ge 1$, $\mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_m)}[\|g\|_2^{2q}]^{1/q} \le (q-1)\cdot(m+1)$.*

**Proof** By Fact 6 applied to $f(g) \triangleq \|g\|_2^2$ and $d = 2$, we have that $\mathbb{E}[\|g\|_2^{2q}]^{1/q} \le \mathbb{E}[\|g\|_2^4]^{1/2}$. But it is straightforward to compute $\mathbb{E}[\|g\|_2^4] = m^2 + 2m$, from which the claim follows. ■

**Corollary 17** *For any polynomial $p \in \mathbb{R}_d[x_1, ..., x_r]$, $\mathbf{j} = (j_1, ..., j_\ell) \in [r]^\ell$, and integer $q \ge 1$,*

$$\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^q]^{1/q} \le (q-1)^{d/2} \cdot d^{\ell/2} \cdot \mathit{Var}[p]^{1/2}$$

**Proof** By Fact 6,

$$\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^q]^{1/q} \le (q-1)^{d/2}\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^2]^{1/2}$$

Write $\mathrm{D}_{\mathbf{j}}\, p$ as $\frac{\partial^\ell}{\partial x_1^{a_1}\cdots\partial x_r^{a_r}}p$, where $a_i$ is the number of entries of $\mathbf{j}$ equal to $i$, and write $p$ in the tensored Hermite basis $p = \sum_I c_I\phi_I$. By Fact 2,

$$\mathrm{D}_{\mathbf{j}}\, p(x) = \frac{\partial^\ell}{\partial x_1^{a_1}\cdots\partial x_r^{a_r}}p(x) = \sum_I c_I\left(\prod_{i\in[r]}\phi_{I_i}^{[a_i]}(x_i)\right) = \sum_I c_I\left(\prod_{i\in[r]}\sqrt{\frac{I_i!}{(I_i-a_i)!}}\phi_{I_i-a_i}(x_i)\right),$$

so by orthogonality and the fact that $a_1 + \cdots + a_r = \ell$, we see that

$$\mathbb{E}_g[(\mathrm{D}_{\mathbf{j}}\, p(g))^2] = \sum_I c_I^2 \cdot \prod_{i\in[r]}\frac{I_i!}{(I_i-a_i)!} \le \sum_{I\neq\emptyset} c_I^2 \cdot \prod_{i\in[r]} d^{a_i} = d^\ell \cdot \mathrm{Var}[p],$$

from which the claim follows. ■

We can use Corollary 17 to bound the moments of $\|\nabla p(g)\|_2^2$.

**Lemma 18** *For any polynomial $p \in \mathbb{R}_d[x_1, ..., x_r]$ and any integer $q \geq 1$,*

$$\mathbb{E}[\|\nabla p(g)\|_2^{2q}]^{1/q} \leq rd \cdot (2q-1)^d \cdot \mathit{Var}[p]$$

**Proof** We have

$$\mathbb{E}[\|\nabla p(g)\|_2^{2q}] \leq r^{q-1} \cdot \mathbb{E}[\|\nabla p(g)\|_{2q}^{2q}] = r^{q-1} \cdot \sum_{i=1}^{r} \mathbb{E}\left[\left(\frac{\partial}{\partial x_i} p(g)\right)^{2q}\right] \leq r^q \cdot (2q-1)^{dq} \cdot d^q \cdot \mathrm{Var}[p]^q,$$

where the first inequality follows by Holder's, and the last step follows by Corollary 17. ■

It will be useful to give a corresponding lower bound for $\mathbb{E}[\|\nabla p(g)\|_2^2]$:

**Lemma 19** *For any polynomial $p \in \mathbb{R}_d[x_1, ..., x_r]$, $\mathbb{E}_g[\|\nabla p(g)\|_2^2] \geq \mathit{Var}[p]$.*

**Proof** Again, write $p$ in the tensored Hermite basis $p = \sum_I c_I \phi_I$. We know that

$$\sum_i \mathbb{E}\left[\left(\frac{\partial}{\partial x_i} p(g)\right)^2\right] = \sum_I c_I^2 \cdot \sum_i I_i \geq \sum_{I \neq \emptyset} c_I^2 = \mathrm{Var}[p],$$

from which the claim follows. ■

The following more careful estimate gives something better than what Cauchy-Schwarz, Corollary 16, and Lemma 18 imply.

**Lemma 20** *For any $p \in \mathbb{R}_d[x_1, ..., x_r]$, $\mathbb{E}_g\left[\|g\|^2 \cdot \|\nabla p(g)\|_2^2\right]^{1/2} \leq O(rd) \cdot \mathit{Var}[p]^{1/2}$.*

**Proof** Take any $i, j \in [r]$. Let $q_I^{i,j}$ denote the polynomial $\prod_{\ell \in [|I|]: \ell \neq i,j} \phi_{I_\ell}(x_\ell)$. If $i = j$, then

$$\begin{aligned}
\mathbb{E}\left[g_i^2 \cdot \left(\frac{\partial}{\partial x_j} p(g)\right)^2\right] &= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,i}(g) \cdot \sqrt{I_i} \cdot g_i \cdot \phi_{I_i - 1}(x_i)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,i}(g) \cdot \sqrt{I_i} \cdot \left(\sqrt{I_i} \cdot \phi_{I_i}(g_i) + \sqrt{I_i - 1} \cdot \phi_{I_i - 2}(g_i)\right)\right)^2\right] \\
&\leq 2\left(\sum_I c_I^2 \cdot I_i^2 + \sum_I c_I^2 \cdot I_i(I_i - 1)\right) \leq 4d^2 \,\mathrm{Var}[p],
\end{aligned}$$

where the second step follows by Corollary 4, and the third step follows by the elementary inequality $(a+b)^2 \le 2a^2 + 2b^2$. Likewise, if $i \neq j$, then we have that

$$\mathbb{E}\left[g_i^2 \cdot \left(\frac{\partial}{\partial x_j} p(g)\right)^2\right] = \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,j}(g) \cdot g_i \cdot \phi_{I_i}(g_i) \cdot \sqrt{I_j} \cdot \phi_{I_j-1}(g_j)\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_I c_I \cdot q_I^{i,j}(g) \cdot \left(\sqrt{I_i+1} \cdot \phi_{I_i+1}(g_i) + \sqrt{I_i} \cdot \phi_{I_i-1}(g_i)\right) \cdot \sqrt{I_j} \cdot \phi_{I_j-1}(g_j)\right)^2\right]$$

$$\le 2\left(\sum_I c_I^2 \cdot (I_i+1) I_j + \sum_I c_I^2 \cdot I_i I_j\right) \le 4d(d+1)\,\mathrm{Var}[p] \le 5d^2\,\mathrm{Var}[p].$$

The lemma follows upon summing over $i, j \in [r]$. ■

The following basic inequality will also be useful.

**Lemma 21** *Let $\mathcal{S}$ denote the collection of all multisets $I$ of size at most $d$ consisting of elements of $[r]$. Then $\mathbb{E}\left[\left(\sum_I \phi_I(g)^2\right)^2\right] \le O(r)^{2d}$.*

**Proof** We have that

$$\mathbb{E}\left[\left(\sum_I \phi_I(g)^2\right)^2\right] \le |\mathcal{S}| \cdot \mathbb{E}\left[\sum_I \phi_I(g)^4\right] = |\mathcal{S}| \cdot 9^d \sum_I \mathbb{E}\left[\phi_I(g)^2\right] = |\mathcal{S}|^2 \cdot 9^d = O(r)^{2d},$$

where the first step follows by Cauchy-Schwarz, the second by Fact 6, the third by orthonormality of $\{\phi_I\}$, and the last by the fact that $|\mathcal{S}| = O(r)^d$. ■

### A.4. Tail Bounds

We will need the following elementary estimates for Gaussian tails and correlated Gaussians. Define $\mathrm{erf}(\beta) \triangleq \Pr_{h\sim\mathcal{N}(0,1)}[|h| \le \beta]$ and $\mathrm{erfc}(\beta) \triangleq 1 - \mathrm{erf}(\beta)$ (note we eschew the usual normalization). It is an elementary fact that under this normalization, for all $z > 0$ we have that $\mathrm{erfc}(z) \le e^{-z^2/2}$.

**Fact 7 (e.g. Proposition 2.1.2 in Vershynin (2018))**

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \le \mathit{erfc}(t) \le \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

**Fact 8** *Let $\rho \in (1/2, 1)$. For vectors $v, v' \in \mathbb{S}^{n-1}$ for which $\langle v, v'\rangle = \rho$, we have that*

$$\Pr_{x\sim\mathcal{N}(0,\mathbf{I}_d)}\left[|\langle v, x\rangle| > 1 \ \wedge\ |\langle v', x\rangle| \le 1\right] \le O(\sqrt{1-\rho^2}) \tag{5}$$

**Proof** We will bound the probability that $\langle v, x\rangle > 1$ and $\langle v', x\rangle \leq 1$, from which the desired probability bound in the claim follows up to a constant factor.

First, we may write $v = \rho \cdot v' + \sqrt{1-\rho^2} \cdot v^\perp$ for $v^\perp \in \mathbb{S}^{n-1}$ orthogonal to $v'$. Then $\langle v', x\rangle$ and $\langle v^\perp, x\rangle$ are independent standard Gaussians $g'$ and $g^\perp$. Also define $g \triangleq \langle v, x\rangle$ so that $g, g'$ are $\rho$-correlated Gaussians. Provided $g > 1$ and $g' \leq 1$, the conditional density of $g'$ relative to $g$ is given by $\int_{-\infty}^{\frac{1-\rho g}{\sqrt{1-\rho^2}}} \mathcal{N}(0,1,x)$, where $\mathcal{N}(0,1,x)$ denotes the density of the standard Gaussian at $x$. When $g > 1/\rho$, this integral is simply $\frac{1}{2} \cdot \mathrm{erfc}\left(\frac{1-\rho g}{\sqrt{1-\rho^2}}\right) \leq \frac{1}{2}\exp\left(-\frac{(1-\rho g)^2}{2(1-\rho^2)}\right)$.

We will also crudely upper bound the probability that $1 < g \leq 1/\rho$ and $g' \leq 1$ by the probability that $1 < g \leq 1/\rho$, which can be upper bounded by $\frac{1}{4}\left(\frac{1}{\rho} - 1\right) = O(1-\rho)$.

We conclude that the quantity on the left-hand side of (5) is at most

$$\begin{aligned} O(1-\rho) + \frac{1}{2}\int_{1/\rho}^{\infty} \exp\left(-\frac{(1-\rho g)^2}{2(1-\rho^2)}\right) dg &= O(1-\rho) + \frac{1}{2}\int_{0}^{\infty} \exp\left(-\frac{g^2}{2\cdot(\rho^{-2}-1)}\right) dg \\ &\leq O(1-\rho) + \frac{1}{4}\cdot\sqrt{2\pi}\cdot\sqrt{\rho^{-2}-1} \\ &= O(1-\rho) + O(\sqrt{1-\rho^2}) = O(\sqrt{1-\rho^2}), \end{aligned}$$

where the first step is by shifting the integrand, the second by standard Gaussian integration. ■

A similar argument to the above shows the following:

**Fact 9** *Let $\rho \in (1/2, 1)$. For vectors $v, v' \in \mathbb{S}^{n-1}$ for which $\langle v, v'\rangle = \rho$, and an arbitrary unit vector $v$, we have that*

$$\mathbb{E}_{x\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\langle v, x\rangle^2 \,|\, |\langle v, x\rangle| > 1 \ \wedge\ |\langle v', x\rangle| \leq 1\right] = O(1).$$

### A.5. Subspaces and Subspace Distances

**Definition 22** *Given $V, V' \in St_r^n$, the* Procrustes distance $d_P(V, V')$ *is given by*

$$d_P(V, V') \triangleq \min_{O\in O(r)} \|V - V'O\|_F.$$

*Let $0 \leq \theta_1 \leq \cdots \leq \theta_r \leq \pi/2$ be the principal angles between $V$ and $V'$. Then we also have that*

$$d_P(V, V') = 2\left(\sum_{i=1}^{r} \sin^2(\theta_i/2)\right)^{1/2}.$$

**Definition 23** *Given $V, V' \in St_r^n$, the* chordal distance $d_C(V, V')$ *is given by*

$$d_C(V, V') \triangleq (d - \|V^\top V'\|_F^2)^{1/2}$$

*Let $0 \leq \theta_1 \leq \cdots \leq \theta_r \leq \pi/2$ be the principal angles between $V$ and $V'$. Then we also have that*

$$d_C(V, V') = \left(\sum_{i=1}^{r} \sin^2\theta_i\right)^{1/2}.$$

**Fact 10 (Triangle inequality for Procrustes)** *Given any $V_1, V_2, V_3 \in St_r^n$,*

$$d_P(V_1, V_2) + d_P(V_2, V_3) \geq d_P(V_1, V_3).$$

**Lemma 24** $d_P(V, V')^2/2 \leq d_C(V, V')^2 \leq d_P(V, V')^2$.

**Proof** This follows immediately from the elementary inequality $2\sin^2(\theta/2) \leq \sin^2(\theta) \leq 4\sin^2(\theta/2)$ for $\theta \in [0, \pi/2]$. ■

Next, we give a more refined estimate for $d_P(V, V')^2 - d_C(V, V')^2$.

**Lemma 25** $d_P(V, V')^2 - d_C(V, V')^2 \leq d_P(V, V')^4$.

**Proof** From the elementary inequality $4\sin^2(\theta/2) - \sin^2(\theta) \leq \sin^4(\theta)$ for $\theta \in [0, \pi/2]$, we see that

$$d_P(V, V')^2 - d_C(V, V')^2 = \left(\sum_{i=1}^r \sin^4\theta_i\right)^2 \leq \left(\sum_{i=1}^r \sin^2\theta_i\right)^2 = d_C(V, V')^4 \leq d_P(V, V')^4$$

as claimed. ■

The following consequence of Lemma 25 will be useful in our analysis of GEOSGD.

**Lemma 26** *For $V, V^* \in St_r^n$, we have that $\|\mathbf{I} - V^\top V^*\|_2 \leq \|V - V_F^*\|$. If $V, V^*$ additionally satisfy that $\|V - V^*\|_F = d_P(V, V^*)$, then we have that $\|\mathbf{I} - V^\top V^*\|_2 \leq d_P(V, V^*)^2$.*

**Proof** It suffices to upper bound $\|\mathbf{I} - V^\top V^*\|_F$. Note that

$$\begin{aligned}\|\mathbf{I} - V^\top V^*\|_F^2 &= d - 2\mathrm{Tr}(V^\top V^*) + \|V^\top V^*\|_F^2 \\ &= \|V - V^*\|_F^2 - d_C(V, V^*)^2 \leq \|V, V^*\|_F^2,\end{aligned}$$

from which the first part of the lemma follows.

For the second bound, note that

$$\|V - V^*\|_F^2 - d_C(V, V^*)^2 = d_P(V, V^*)^2 - d_C(V, V^*)^2 \leq d_P(V, V^*)^4,$$

where the final step follows by Lemma 25. ■

The following says that if a set of $r$ orthogonal unit vectors all have large component in $U^*$, then their span is close to the true subspace in the sense of either of the distances above.

**Lemma 27** *Let $\Pi$ denote orthogonal projection to a subspace $U_1 \in G(n, \ell)$. Let $v_1, ..., v_\ell \in \mathbb{S}^{n-1}$ be orthogonal and satisfy $\|\Pi v_i\|_2 \geq 1 - \varepsilon$ for all $i \in [r]$. Let $U_2 \triangleq$ span$(\{v_i\})$. Then $d_C(U_1, U_2) \leq \varepsilon \cdot \ell$ and $d_P(U_1, U_2) \leq \sqrt{2}\varepsilon \cdot \ell$.*

**Proof** Let $V_1 \in \mathrm{St}_\ell^n$ be any frame with columns forming a basis for $U_1$, and let $V_2 \in \mathrm{St}_\ell^n$ be the frame with columns given by $\{v_i\}_{i\in[\ell]}$. Observe that

$$d_C(U_1,U_2)^2 = \ell - \|V_1^\top V_2\|_F^2 = \ell - \mathrm{Tr}\left(V_2^\top \Pi V_2\right) \geq \ell - \sum_{i=1}^{\ell} \|\Pi v_i\|_2 = \varepsilon\cdot\ell.$$

as claimed. ■

We will also need the gap-free Wedin theorem of Allen-Zhu and Li (2016):

**Lemma 28 (Allen-Zhu and Li (2016), Lemma B.3)** *Let $\epsilon,\gamma,\mu > 0$. For psd matrices $\mathbf{A},\hat{\mathbf{A}} \in \mathbb{R}^{d\times d}$ for which $\|\mathbf{A}-\hat{\mathbf{A}}\|_2 \leq \epsilon$, if $U$ is the matrix whose columns consist of the eigenvectors of $\mathbf{A}$ with eigenvalue at least $\mu$, and $\hat{U}$ is the matrix whose columns consist of the eigenvectors of $\hat{\mathbf{A}}$ with eigenvalue at most $\mu-\gamma$, then $\|\mathbf{U}^\top\hat{\mathbf{U}}\|_2 \leq \epsilon/\gamma$.*

**Claim 1** *For any $M \in \mathbb{R}^{n\times n}$ and projectors $\Pi_1,\Pi_2 \in \mathbb{R}^{n\times n}$ to subspaces $U_1,U_2 \in G(n,\ell)$, $\|\Pi_1^\top M\Pi_1 - \Pi_2^\top M\Pi_2\|_2 \leq O(\|M\|_2\cdot d_C(U_1,U_2))$.*

**Proof** We bound $\|(\Pi_1-\Pi_2)^\top M\Pi_1\|_2$ and $\|\Pi_2^\top M(\Pi_1-\Pi_2)\|_2$ and apply triangle inequality.

By sub-multiplicativity of the operator norm and the fact that projections have spectral norm 1, $\|(\Pi_1-\Pi_2)^\top M\Pi_1\|_2 \leq \|\Pi_1-\Pi_2\|_2\cdot\|M\|_2$. Finally, note that

$$\|\Pi_1-\Pi_2\|_2 \leq \|\Pi_1-\Pi_2\|_F = \sqrt{2}\cdot d_C(U_1,U_2),$$

from which the claim follows. ■

**Lemma 29** *Let $U^* \in G(n,r)$, $V \in St_r^n$, and $\epsilon > 0$. Suppose the columns $v_i$ of $V$ satisfy $\|\Pi_U v_i\|_2 \geq 1-\epsilon$ for every $i \in [\ell]$. Then there exist orthogonal vectors $v_1^*,\ldots,v_\ell^* \in U$ for which $\langle v_i, v_i^*\rangle \geq 1-\epsilon^2\ell^2/2$ for every $i\in[\ell]$.*

**Proof** Let $U \triangleq \mathrm{span}(\{v_i\})$. By Lemma 27, $d_P(U,U^*) \leq \epsilon\cdot\ell$, so there exists a frame $V^* \in \mathrm{St}_r^n$ for $U^*$ such that $\|V-V^*\|_F \leq \epsilon\cdot\ell$. Note that $\|V-V^*\|_F^2 = 2\ell - 2\mathrm{Tr}(V^\top V^*) = 2\sum_{i=1}^{\ell}(1-\langle v_i,v_i^*\rangle)$. As $v_i, v_i^*$ are unit vectors $1-\langle v_i,v_i^*\rangle \geq 0$ for every $i\in[\ell]$, so we conclude that $\langle v_i,v_i^*\rangle \geq 1-\epsilon^2\ell^2/2$ for each $i\in[\ell]$. ■

## Appendix B. Warm Start via Trimmed PCA

The main result of this section is the proof of Theorem 5. Let $\mathcal{D}$ denote the distribution $(X,Y)$ where $Y = P(X)$ is a $\alpha$ non-degerate polynomial of rank $r$ and degree at most $d$ as in the hypothesis of the theorem. Let $U^*$ be the true hidden subspace defining $P$. The proof follows the outline described in the introduction closely. To this end, for a *threshold parameter* $\tau > 0$ and a collection of unit vectors $V = \{v_1,\ldots,v_\ell\}$, define the matrix

$$\mathbf{M}_V^\tau \triangleq \Pi_{V^\perp}\cdot\left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\left[\{|y|>\tau\} \wedge \{|\langle v_i,x\rangle| \leq 1,\ \forall\, i\in[\ell]\}\right]\cdot(xx^\top - \mathbf{I})\right]\right)\cdot\Pi_{V^\perp}.$$

We will show that the above algorithm satisfies the guarantees of Theorem 5. The core of its analysis will be the following main inductive lemma.

**Algorithm 2:** TRIMMEDPCA($\mathcal{D}, \epsilon, \delta$)

**Input:** Sample access to $\mathcal{D}$, target error $\epsilon$, failure probability $\delta$

**Output:** Frame for a subspace $U$ with $d_P(U, U^*) \leq \epsilon$, with probability at least $1 - \delta$

$V_0 \leftarrow \emptyset$.

$\tau \leftarrow \tau(r, d, \alpha)$ // Lemma 30)

**for** $0 \leq \ell \leq r - 1$ **do**

  Draw $N = O_{r,d,\epsilon}(n)$ samples $(x_1, y_1), ..., (x_N, y_N)$ // Theorem 5)

  Compute an empirical approximation $\widehat{M}^\ell$ to $M^\tau_{V_\ell}$ by drawing $N = O_{r,d,\epsilon}(n)$ samples from the distribution $\mathcal{D}$.

  Let $v^{\ell+1}$ be the eigenvector with the largest eigenvalue of $\widehat{M}^\ell$.

  $V_{\ell+1} \leftarrow V_\ell \cup \{v^{\ell+1}\}$.

**end**

Output $V_r$

**Lemma 30** *There exists $\tau = \tau(r, d, \alpha)$, a constant $C = C(r, d, \alpha)$ such that the following holds. Let $V = \{v_1, \ldots, v_\ell\}$ for $\ell < r$ be orthonormal vectors such that $\|\Pi_{U^*} v_i\| \geq 1 - \rho$, and $M$ a matrix such that $\|M - M^\tau_V\| \leq \rho$. Then, the largest eigenvector $v$ of $M$ satisfies $\|\Pi_{U^*} v\| \geq 1 - C\ell^2\rho$.*

Before proving the lemma, we first show how the main theorem follows from the above.

**Proof** [Proof of Theorem 5] Let $C, \tau$ be as in the above lemma. For a $\rho_0$ to be chosen later, let $\rho_{\ell+1} = C\ell^2\rho_\ell$ for $\ell \geq 0$. Let $N = O(n \log(r/\delta)/\rho_0^2)$.

We will show by induction that $\|\Pi_{U^*} v_\ell\| \geq 1 - \rho_\ell$. Suppose we have the statement for $v_1, \ldots, v_\ell$ computed by the algorithm. Then, in the next iteration, by Lemma 10, with probability at least $1 - \delta/r$, we will have $\|\widehat{M}^\ell - M^\tau_{V_\ell}\| \leq \rho_\ell$. In this case, the top eigenvector $v_{\ell+1}$ of $\widehat{M}^\ell$ satisfies $\|Pi_{U^*} v_{\ell+1}\| \geq 1 - C\ell\rho_\ell^{1/4} = 1 - C\ell\rho_{\ell+1}$.

By a union bound over the $r$ events, we get that with probability at least $1 - \delta$, we would have computed orthonormal vectors $v_1, \ldots, v_r$ such that $\|\Pi_{U^*} v_i\| \geq 1 - \rho_r$. Now, by Lemma 27, $d_P(sp(v_1, \ldots, v_r), U^*) \leq O(\rho_r r)$.

As $\rho_r \leq C^r r^{2r} \rho_0$, the lemma follows by setting $\rho_0 = \epsilon/(Cr)^{2r}$. The overall sample complexity will be $N = O(r \cdot n \log(r/\delta)/\rho_0^2 = C(r, d, \alpha) n \log(r/\delta)/\epsilon^2$ as stated in the theorem.

Each iteration of the for loop takes time $O(n^2 N)$ to form the matrix $\widehat{M}^\ell$ and further $O(n^3)$ time to compute the top eigenvector. So the total run-time is $O(r(n^2 N + n^3))$. ■

### B.1. Proof of Lemma 30

We next prove the Lemma 30 which allows us to identify one direction at a time. The proof proceeds as follows:

1. We first show a lower bound on the largest eigenvalue of the matrix $M^\tau_V$ when the vectors $v_1, \ldots, v_\ell$ lie in the subspace $U^*$. This is the heart of the proof and follows from a compactness argument. This essentially gives a proof of the lemma when $V \subseteq U^*$ (and $M$ approximates $M^\tau_V$). See Lemmas 31, 32.

2. The second step is to reduce to the above case. Given $V$ as in the lemma, we find orthonormal vectors $V^* = \{v_1^*, \ldots, v_\ell^*\} \in U^*$ such that $\|v_i - v_i^*\| \leq O(\ell\rho)$. We then do a perturbation analysis (using elementary linear algebra) to argue that perturbing the vectors $V$ slightly will only incur a small error in the matrix $M_V^\tau$. Specifically, we will show that $\|M_V^\tau - M_{V^*}^\tau\| \leq O(\ell\rho^{1/4})$. See Lemma 33.

For brevity, in the remainder of this section let $\Pi^*$ denote orthogonal projection to the true subspace $U^* \subset \mathbb{R}^n$.

First, we show that if the vectors in $V^* = \{v_1^*, ..., v_\ell^*\}$ were vectors in the true subspace, then the top eigenvector of $\mathbf{M}_{V^*}^\tau$ will be a new vector in the subspace orthogonal to the preceding ones.

**Lemma 31** *There are absolute constants $\tau = \tau_{r,d,\alpha_{\mathsf{ndg}}} > 0$ and $\lambda = \lambda_{r,d,\alpha_{\mathsf{ndg}}} > 0$ for which the following holds. Suppose $V^* = \{v_1^*, ..., v_\ell^*\} \subset \mathbb{S}^{n-1}$ are orthogonal and is in $U^*$. Then*

1. *The kernel of $\mathbf{M}_{V^*}^\tau$ contains span$(v_1^*, ..., v_\ell^*)$ as well as the orthogonal complement of $U^*$.*
2. *The top eigenvalue of $\mathbf{M}_{V^*}^\tau$ is at least $\lambda$ and corresponds to a vector in $U^* \backslash$span$(V^*)$.*

Note that Lemma 31 already gives a nontrivial algorithmic guarantee for $\ell = 0$: given exact access to $\mathbf{M}_\emptyset^\tau$, we can recover a vector inside the true subspace by taking its top eigenvector.

**Proof** Let $\{v_i^*\}_{i\in[\ell]}$ to an orthonormal basis $\{v_i^*\}_{i\in[r]}$ of $U^*$, and let $p^*((V^*)^\top x)$ be a realization of the true low-rank polynomial, where the frame $V^* \in \mathrm{St}_r^n$ consists of these basis elements.

(Proof of 1) Certainly span$(\{v_i^*\}_{i\in[\ell]})$ lies in the kernel of $\mathbf{M}_{V^*}^\tau$ by definition. Moreover for any $v \in \mathbb{S}^{n-1}$ orthogonal to $U^*$, because $\langle v_i^*, x\rangle, ..., \langle v_r^*, x\rangle, \langle v, x\rangle$ are independent Gaussians, call them $g_1, ..., g_r, g_\perp \sim \mathcal{N}(0,1)$, we have that

$$\begin{aligned} v^{*\top}\mathbf{M}_{V^*}^\tau v &= \mathbb{E}\left[\mathbb{1}\left[\{|p^*(g_1,...,g_r)| > \tau\} \wedge \{|g_i| \leq 1 \,\forall\, i \in [\ell]\}\right] \cdot (g_\perp^2 - 1)\right] \\ &= \mathbb{E}\left[\mathbb{1}\left[\{|p^*(g_1,...,g_r)| > \tau\} \wedge \{|g_i| \leq 1 \,\forall\, i \in [\ell]\}\right]\right] \cdot \mathbb{E}\left[(g_\perp^2 - 1)\right] \\ &= 0. \end{aligned}$$

(Proof of 2) The fact that the top eigenvector lies in $U^* \backslash$span$(\{v_i^*\}_{i\in[\ell]})$ follows immediately from the fact that it must be orthogonal to both span$(\{v_i^*\}_{i\in[\ell]}$ and the orthogonal complement of $U^*$.

To get a bound on the top eigenvalue, define the quantities $Z_i \triangleq v_i^{*\top}\mathbf{M}_{V^*}^\tau v_i^*$ for $\ell < i \leq r$. Again using the fact that $\langle v_i^*, x\rangle, ..., \langle v_r^*, x\rangle$ are independent Gaussians $g_1, ..., g_r$, we have

$$\sum_{i=\ell+1}^r Z_i = \mathbb{E}\left[\mathbb{1}\left[\{|p^*(g_1,...,g_r)| > \tau\} \wedge \{|g_i| \leq 1 \,\forall\, i \in [\ell]\}\right] \cdot \left(\sum_{i>\ell} g_i^2 - (r-\ell)\right)\right].$$

We would like to lower bound this quantity, at which point by averaging over $i$ we conclude the proof of the lemma.

Let $K \subset \mathbb{R}^r$ denote the set of all points $x$ for which $|x_i| \leq 1$ for all $1 \leq i \leq \ell$ and for which $\sum_{i=\ell+1}^r x_i^2 \leq 2(r-\ell)$. For any $p \in \mathcal{P}_{r,d}^{\alpha_{\mathsf{ndg}}}$, define $\|p\|_K \triangleq \sup_{x\in K}|p(x)|$. By compactness of $K$, $\|p\|_K < \infty$ for all $p$, and furthermore $\|p\|_K$ is a continuous function of $p$. If we take $\tau = \tau(\alpha_{\mathsf{ndg}}, r, d, \ell) \triangleq \sup_{p\in\mathcal{P}_{r,d}^{\alpha_{\mathsf{ndg}}}} \|p\|_K$, then by compactness of $\mathcal{P}_{r,d}^{\alpha_{\mathsf{ndg}}}$, is some finite quantity depending only on $\alpha_{\mathsf{ndg}}$, $r$, $d$, and $\ell$. For this choice of $\tau$, we conclude that if a point $(g_1, ..., g_r) \in \mathbb{R}^r$

satisfies $|p^*(g_1,...,g_r)|>\tau$ and $|g_i|\leq 1$ for all $i\in[\ell]$, then it must lie outside $K$. We conclude that

$$\sum_{i=\ell+1}^{r} Z_i \geq (r-\ell)\cdot \Pr\left[\{|p^*(g_1,...,g_r)|>\tau\} \wedge \{g\notin K\}\right].$$

In particular, there exists some $i>\ell$ for which $Z_i \geq \Pr\left[\{|p^*(g_1,...,g_r)|>\tau\} \wedge \{g\notin K\}\right]$. The right-hand side is a continuous function in $p$, call it $A_p$. For any $p$, there must exist some point $x\notin K$ for which $p^*(x)>\tau$, so again by compactness of $\mathcal{P}^{\alpha_{\mathsf{ndg}}}_{r,d}$, we see that $Z_i\geq\lambda$ for some strictly positive constant $\lambda$ depending only on $\alpha_{\mathsf{ndg}}, r, d, \ell$. ■

Henceforth, for brevity, we will denote the constants $\tau_{r,d,\alpha_{\mathsf{ndg}}}$ and $\lambda_{r,d,\alpha_{\mathsf{ndg}}}$ from Lemma 31 by $\tau$ and $\lambda$ respectively.

We next show that the above lemma implies Lemma 30 for the case when $V\subseteq U^*$.

**Lemma 32** *Given orthonormal vectors $V^*=\{v_1^*,...,v_\ell^*\}\subseteq U^*$, and a matrix $\mathbf{M}$ for which $\|\mathbf{M}-\mathbf{M}^\tau_{V^*}\|_2\leq\rho$, the top eigenvector $v$ of $\mathbf{M}$ satisfies*

$$\|\Pi^* v\|_2 \geq \left(1-\frac{\rho}{\lambda-\rho}\right)^{1/2} \geq 1-(2/\lambda)\rho.$$

**Proof** By Lemma 33, the top eigenvalue of $\mathbf{M}^\tau_{V^*}$ is at least that of $\mathbf{M}^\tau_{V^*}$ minus $\rho$. Let $V^*\in \mathrm{St}^n_{r-\ell}$ be the matrix whose columns consist of $v^*_{\ell+1},...,v^*_r$. Invoking the first part of Lemma 31, let $B$ be the matrix whose columns consist of a basis $(v_1^*,...,v_\ell^*,w_{\ell+1},...,w_n)$ for the kernel of $\mathbf{M}^\tau_{V^*}$, so that

$$V^*V^{*\top}+BB^\top=\mathbf{I}_n. \tag{6}$$

By applying Lemma 28 to $\mathbf{M}$ and $\mathbf{M}^\tau_{V^*}$ with $\mu=\gamma=\lambda-\rho$, we get that $\|v^\top\cdot B\|_2\leq\frac{\rho}{\lambda-\rho}$. By (6),

$$\|v^\top\cdot V^*\|_2=\left(1-\|v^\top\cdot B\|_2^2\right)^{1/2}\geq\left(1-\frac{\rho}{\lambda-\rho}\right)^{1/2}.$$

Note that $\|\Pi^*v\|_2=\|v^\top\cdot V^*\|_2$. The lemma now follows. ■

Finally, we show that for orthonormal vectors $V=\{v_1,...,v_\ell\}$ which all have large component in $U^*$, the matrix $\mathbf{M}^\tau_V$ is spectrally close to some $\mathbf{M}^\tau_{V^*}$ for $V^*=\{v_1^*,\dots v_\ell^*\}$ in $U^*$.

**Lemma 33** *There is an absolute constant $c_2>0$ for which the following holds. Given orthonormal vectors $V=\{v_1,...,v_\ell\}$ for which $\|\Pi^* v_i\|_2\geq 1-\varepsilon$ for some $0\leq\varepsilon<1$ for all $i\in[\ell]$, there exist orthonormal vectors $V^*=\{v_1^*,\dots,v_\ell^*\}\subset U^*$ such that $\|\mathbf{M}^\tau_V-\mathbf{M}^\tau_{V^*}\|_2\leq c_2\varepsilon\ell^2$.*

**Proof** Let $V^*=\{v_1^*,\dots,v_\ell^*\}$ be orthonormal vectors in $U^*$ guaranteed by Lemma 29 such that $\langle v_i,v_i^*\rangle\geq 1-\varepsilon^2\ell^2/2$.

For each $0\leq a\leq\ell$, define the hybrid collections of vectors $V^{(a)}\triangleq\{v_1^*,...,v^*_{\ell-1},v_\ell,...,v_r\}$, and also define the hybrid matrices

$$\mathbf{M}^{(a)}\triangleq\left(\Pi^\perp_{\{v_i\}}\right)^\top\cdot\left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\left[\{|y|>\tau\}\wedge\{|\langle v_i^{(a)},x\rangle|\leq 1\;\forall\, i\in[\ell]\}\right]\cdot(xx^\top-\mathbf{I})\right]\right)\cdot\Pi^\perp_{\{v_i\}}.$$

Note that $V^{(0)}=V$ and $V^{(\ell)}=V^*$, and similarly $\mathbf{M}^{(0)}=\mathbf{M}^\tau_V$.

We will bound $\|\mathbf{M}^{(a+1)}-\mathbf{M}^{(a)}\|_2$ for every $0\leq a<\ell$, and then bound $\|\mathbf{M}^{(\ell)}-\mathbf{M}^\tau_{V^*}\|_2$. The lemma will then follow by triangle inequality.

**Claim 2** *For any* $0 \le a < \ell$, $\|\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)}\|_2 \le O(\varepsilon\ell)$.

**Proof** We will bound $v^\top(\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)})v$ for any $v \in \mathbb{R}^n$; without loss of generality, we may assume $v$ is orthogonal to $v_1, ..., v_\ell$.

Let $\mathcal{E}$ denote the event that $|\langle v_a, x\rangle| > 1$ and $\{|\langle v_a^*, x\rangle \le 1\}$ or vice-versa. Now, note that the indicator events in the definitions of $M^{(a)}$ and $M^{(a+1)}$ only differ when $\mathcal{E}$ occurs. Therefore,

$$\begin{aligned}\left|v^\top(\mathbf{M}^{(a+1)} - \mathbf{M}^{(a)})v\right| &\le E[1(\mathcal{E}) \cdot (\langle v, x\rangle^2 + 1)] \\ &\le E[1(\mathcal{E})] \cdot (1 + E[\langle v, x\rangle^2 | \mathcal{E}]) \\ &= O(\Pr[\mathcal{E}]),\end{aligned}$$

where the last inequality follows by Fact 9.

Finally note that by Fact 8,

$$Pr[\mathcal{E}] \le O(\sqrt{1 - \langle v_i, v_i^*\rangle}) = O(\rho\ell).$$

The claim now follows. ■

To bound $\|\mathbf{M}^{(\ell)} - \mathbf{M}^\tau_{V^*}\|_2$, we will use Claim 1. We note that the matrix

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\left[\{|y| > \tau\} \wedge \{|\langle v_i, x\rangle| \le 1 \; \forall \, i \in [\ell]\}\right] \cdot (xx^\top - \mathbf{I})\right]$$

has spectral norm at most $\|\mathbb{E}[xx^\top]\|_2 + 1 = 2$. So if $U \triangleq \text{span}(v_1, ..., v_\ell)$ and $U' \triangleq \text{span}(v_1^*, ..., v_\ell^*)$, then by Claim 1,

$$\|\mathbf{M}^{(\ell)} - \mathbf{M}^\tau_{V^*}\|_2 \le O(d_C(U, U')) \le O(\varepsilon \cdot \ell),$$

where the last step follows by Lemma 27.

Lemma 33 follows by applying the above inequality, Claim 2 for all $0 \le a < \ell$, and triangle inequality. ■

We now put Lemmas 32, 33 together to prove Lemma 30.

**Proof** [Proof of Lemma 30] Choose $\tau$ to be as in Lemma 31. We will choose $C = C\ell^2/\lambda$ for $\lambda$ as in the lemma and $C$ a universal constant.

Let $V^*$ be the set of $\ell$ orthonormal vectors in $U^*$ as in Lemma 33 so that

$$\|M^\tau_V - M^\tau_{V^*}\| \le O(\rho\ell^2).$$

Thus, we have $\|M - M^\tau_{V^*}\| \le O(\rho\ell^2)$. The lemma now follows by applying Lemma 32. ■

## Appendix C. Boosting via Stochastic Riemannian Optimization

In this section we describe our algorithm for boosting a warm start to arbitrary accuracy and defer the details of its analysis to Sections E and D.

**Theorem 34 (Error Guarantee for GEOSGD)** *There is an absolute constant $c_3 > 0$ such that the following holds. Let $U^*$ be the true subspace of $\mathcal{D}$. Given $V^{(0)} \in St_r^n$ spanning a subspace $U$ for which $d_P(U, U^*) \leq (c_3 \cdot dr^3)^{-d-2}$, if in the specification of* GEOSGD *we take*

$$T = \frac{n}{\alpha_{\mathsf{ndg}}} \cdot \log(1/\epsilon) \cdot \mathit{poly}(\ln(1/\alpha_{\mathsf{ndg}}), r, d, \ln(1/\delta), \ln(n))^d, \tag{7}$$

*then* GEOSGD*$(\mathcal{D}, V^{(0)}, \epsilon, \delta)$ returns $(\mathbf{c}^{(T)}, V^{(T)})$ for which there exists a realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$ such that $d_P(V^{(T)}, V^*) \leq \epsilon$ and $\|\mathbf{c}^{(T)} - \mathbf{c}^*\|_2 \leq \epsilon$.*

**Theorem 35 (Complexity of GEOSGD)** *Let $T_1 \triangleq O(rd^4)^{d+1} \cdot \log(1/\epsilon)$, $B \triangleq O(\log(T_1 \cdot T/\delta))^{2d}$, and $T_2 \triangleq (r/\alpha_{\mathsf{ndg}})^2 \cdot O(d \cdot \log(T/\delta))^{2c_1 d}$. Then* GEOSGD *draws*

$$N \triangleq T \cdot (B \cdot T_1 + T_2) = \tilde{O}\left(\frac{n \log^2(1/\epsilon)}{\alpha_{\mathsf{ndg}}^3} \cdot \mathit{poly}(\ln(1/\alpha_{\mathsf{ndg}}), r, d, \ln(1/\delta), \ln(n))^d\right)$$

*samples and runs in time $n \cdot r^{O(d)} \cdot N$ time.*

### C.1. Preliminaries

Let $M = r^{O(d)}$ be the dimension of the linear space of polynomials of polynomials of degree $d$ over $r$ variables. For $\mathbf{c} = \{c_I\} \in \mathbb{R}^M$, where $I$ ranges over multisets of size at most $d$ consisting of elements of $[r]$, and $V \in \mathrm{St}_r^n$, let parameters $\Theta = (\mathbf{c}, V)$ correspond to a rank-$r$ polynomial $F_x(\Theta) \triangleq \sum_I c_I \phi_I(V^\top x)$ in the variable $x$. Given a sample $(x, y) \sim \mathcal{D}$, let $L_x(\Theta) \triangleq (F_x(\Theta) - y)^2$ denote the empirical risk of a single sample.

We will often regard $F_x$ and $L_x$ as functions solely in $\mathbf{c}$ (resp. $V$) for a fixed choice of $V$ (resp. $\mathbf{c}$): given a fixed $V$ (resp. a fixed $\mathbf{c}$), define $F_x^V(\mathbf{c})$ and $L_x^V(\mathbf{c})$ (resp. $F_x^{\mathbf{c}}(V)$ and $L_x^{\mathbf{c}}(V)$) in the obvious way.

Let $\nabla F_x(\Theta)$ denote the gradient of $F_x$ as a function on Euclidean space, and let $\nabla^{\mathsf{vec}} F_x(\Theta) \triangleq \nabla F_x^c(V)$ and $\nabla^{\mathsf{coef}} F_x(\Theta) \triangleq \nabla F_x^V(\mathbf{c})$ denote its components corresponding to $V$ and $\mathbf{c}$ respectively. We can compute their gradients, indeed all of their higher derivative tensors, explicitly:

**Proposition 36** *For any $x \in \mathbb{R}^n$, $a, b \in \mathbb{Z}_{\geq 0}$, and $\Theta = (c, V)$,*

$$\frac{\partial^{a+b}}{\partial c_{I^{(1)}} \cdots \partial c_{I^{(a)}} \partial V_{i_1, j_1} \cdots \partial V_{i_b, j_b}} F_x(\Theta) = \begin{cases} \left(\prod_{\nu=1}^b x_{i_\nu}\right) \cdot p^{[b]}(V^\top x) & \text{if } a = 0 \\ \left(\prod_{\nu=1}^b x_{i_\nu}\right) \cdot \phi_I^{[b]}(V^\top x) & \text{if } a = 1 \\ 0 & \text{otherwise} \end{cases}$$

From Proposition 36 we conclude that

$$\nabla^{\mathsf{vec}} F_x(\Theta) = x \cdot (\nabla p(V^\top x))^\top \quad \text{and} \quad \nabla^{\mathsf{coef}} F_x(\Theta) = \{\phi_I(V^\top x)\}_I.$$

It will be important to consider $\overline{\nabla}^{\mathsf{vec}} F_x(\Theta) \triangleq \Pi_V^\perp \nabla^{\mathsf{vec}} F_x(\Theta)$ the projection of $\nabla^{\mathsf{vec}} F_x(\Theta)$, to the tangent space of $\mathrm{G}(n, r)$ at the point $[V]$.

Lastly, we record here an elementary estimate which will be used repeatedly in the proceeding sections and defer its proof to Appendix I.1.

**Lemma 37** *For any integer $m \geq 1$ and $\ell = (\ell_1, ..., \ell_m) \in [d+1]^m$,*

$$\left| \mathbb{E}\left[\prod_{\nu=1}^m \left\langle \nabla^{[\ell_\nu]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell_\nu} \right\rangle\right] \right| \leq 2^m \cdot \left(2mdr^2\right)^{m(d+1)/2} \cdot \|V^* - V\|_F^{\sum_\nu \ell_\nu} \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^m$$

### C.2. Gradient Updates: Vanilla and Geodesic

GEOSGD alternates between one of two phases: updating $\mathbf{c}$ or updating $V$. Our updates for $\mathbf{c}$ are straightforward: at iterate $\Theta = (\mathbf{c}, V)$ and given a batch of samples $(x_0, y_0), ..., (x_{B-1}, y_{B-1}) \sim \mathcal{D}$, we fix $V$ and take a vanilla gradient descent step using $\frac{1}{B}\sum_{i=0}^{B-1} L_{x_i}^V(\mathbf{c})$. For learning rate $\eta_{\mathsf{coef}}$, this leads to the update

$$c_I' = c_I - 2\eta_{\mathsf{coef}} \cdot \frac{1}{B}\sum_{i=0}^{B-1}(F_{x_t}(\Theta) - F_{x_i}(\Theta^*)) \cdot \phi_I(V^\top x_i) \triangleq c_I - \frac{1}{B}\sum_{i=0}^{B-1}\left(\Delta_{\mathsf{coef}}^{\Theta,x_i}\right)_I \ \forall\, I. \quad (8)$$

The updates for $V$ will be less standard. At iterate $\Theta = (\mathbf{c}, V)$, and given a sample $(x, y) \sim \mathcal{D}$, consider the geodesic $\Gamma$ on $\mathrm{G}(n, r)$ with initial point $[V] \in \mathrm{G}(n, r)$ and initial velocity $\dot{\Gamma}(0) \triangleq \Pi_V^{\perp}\nabla L_x^{\mathbf{c}}(V)$, where $L_x^{\mathbf{c}}(V) \triangleq L_x(\Theta)$.[3]

Define the vectors $h^{\Theta,x} \in \mathbb{R}^n, \nabla^{\Theta,x} \in \mathbb{R}^r$ by

$$h^{\Theta,x} \triangleq 2(F_x(\Theta) - F_x(\Theta^*)) \cdot \Pi_V^{\perp} \cdot x \quad \text{and} \quad \nabla^{\Theta,x} \triangleq \nabla p(V^\top x) \quad (9)$$

so that $\dot{\Gamma}(0) = h^{\Theta,x} \cdot (\nabla^{\Theta,x})^\top$. Geodesics on $\mathrm{G}(n, r)$ are determined by the SVD of the initial velocity $\dot{\Gamma}(0)$, which is simply given by

$$\dot{\Gamma}(0) = \sigma \cdot \widehat{h}^{\Theta,x} \cdot (\widehat{\nabla}^{\Theta,x})^\top,$$

where

$$\widehat{h}^{\Theta,x} \triangleq \frac{h^{\Theta,x}}{\|h^{\Theta,x}\|} \qquad \widehat{\nabla}^{\Theta,x} \triangleq \frac{\nabla^{\Theta,x}}{\|\nabla^{\Theta,x}\|} \qquad \sigma^{\Theta,x} \triangleq \|h^{\Theta,x}\| \cdot \|\nabla^{\Theta,x}\|.$$

Walking along the geodesic with initial velocity $\dot{\Gamma}(0)$ for time $\eta_{\mathsf{vec}}$ then yields the following update rule (for the details, see the derivation of equation (2.65) in Edelman et al. (1998)),

$$V' \triangleq V - \left(\cos\left(\sigma^{\Theta,x}\eta_{\mathsf{vec}}\right) - 1\right) \cdot V \cdot \widehat{\nabla}^{\Theta,x}(\widehat{\nabla}^{\Theta,x})^\top - \sin\left(\sigma\eta_{\mathsf{vec}}\right) \cdot \widehat{h}^{\Theta,x}\left(\widehat{\nabla}^{\Theta,x}\right)^\top \triangleq V - \Delta_{\mathsf{vec}}^{\Theta,x}. \quad (10)$$

One readily checks that the columns of $V'$ are orthonormal.

We are now ready to state our boosting algorithm GEOSGD, which is composed of two alternating phases, SUBSPACEDESCENT and REALIGNPOLYNOMIAL which execute the updates (8) and (10) respectively. In the next two sections, we will analyze these two phases.

## Appendix D. Guarantees for REALIGNPOLYNOMIAL

Before we can describe our main result of this section, we require some setup.

Henceforth, fix a frame $V \in \mathrm{St}_r^n$. The aim of REALIGNPOLYNOMIAL is to approximately find the $r$-variate, degree-$d$ polynomial $p$ for which $p(V^\top x)$ is closest to the true low-rank polynomial. Suppose $V$ was $\beta$-far in subspace distance from the true subspace for some $\beta$, or equivalently, that there was some frame $V^* \in \mathrm{St}_r^n$ for the true subspace for which $\|V - V^*\|_F = \beta$. By working

3. We emphasize that technically this is not well-defined as this velocity depends on the choice of representative $V$; indeed, $F_x^{\mathbf{c}}(V)$ cannot be regarded as a function on $\mathrm{G}(n, r)$, as $\mathbf{c}$ is fixed so that different rotations of $V$ will actually yield different values. But as our goal is simply to produce an update rule, we can freely ignore this point and see where this line of reasoning leads.

**Algorithm 3:** SUBSPACEDESCENT($\mathcal{D}, V^{(0)}, \mathbf{c}\delta$)

**Input:** Sample access to $\mathcal{D}$; frame $V^{(0)} \in \mathrm{St}_r^n$; coefficients $\mathbf{c} \in \mathbb{R}^M$, failure probability $\delta$

**Output:** $V^{(T)} \in \mathrm{St}_r^n$ which is slightly closer to the true subspace than $V$, provided $(\mathbf{c}, V^{(0)})$ satisfies certain conditions (see Theorem 52 for formal guarantees)

Define iteration count $T$ according to (24).

Define learning rate $\eta_{\mathsf{vec}}$ according to (23).

$\Theta^{(0)} \leftarrow (\mathbf{c}, V^{(0)})$

**for** $0 \le t < T$ **do**

  Sample $(x^t, y^t) \sim \mathcal{D}$ $\widehat{h} \leftarrow \frac{h^{\Theta^{(t)},x^t}}{\|h^{\Theta^{(t)},x^t}\|}$ and $\widehat{\nabla} \leftarrow \frac{\nabla^{\Theta^{(t)},x^t}}{\|\nabla^{\Theta^{(t)},x^t}\|}$ // equation (9)

  $\sigma \leftarrow \|h^{\Theta^{(t)},x^t}\| \cdot \|\nabla^{\Theta^{(t)},x^t}\|$.; $V^{(t+1)} \leftarrow V^{(t)} - \Delta_{\mathsf{vec}}^{\Theta^{(t)},x^t}$ // equation (10)

  $\Theta^{(t+1)} \leftarrow (\mathbf{c}, V^{(t+1)})$

**end**

Output $V^{(T)}$.

**Algorithm 4:** REALIGNPOLYNOMIAL($\mathcal{D}, V, \underline{\epsilon}, \delta$)

**Input:** Sample access to $\mathcal{D}$; $V \in \mathrm{St}_r^n$; target error $\underline{\epsilon}$; failure probability $\delta$

**Output:** $\mathbf{c} \in \mathbb{R}^M$ for which $(\mathbf{c}^{(T)}, V)$ is close to a realization of $\mathcal{D}$ (see Section D for details)

Define batch size $B$ according to (13).

Define iteration count $T$ according to (12).

Define learning rate $\eta_{\mathsf{coef}}$ according to (11).

$\mathbf{c}^{(0)} \leftarrow \mathbf{0}$.

$\Theta^{(0)} \leftarrow (\mathbf{c}^{(0)}, V)$.

**for** $0 \le t < T$ **do**

  Sample $(x_1^t, y_1^t), ..., (x_B^t, y_B^t) \sim \mathcal{D}$.

  For every $I$, $c_I^{(t+1)} \leftarrow c_I^{(t)} - \frac{1}{B}\sum_{i=0}^{B-1} \left(\Delta_{\mathsf{coef}}^{\Theta,x_i^t}\right)_I$ // equation (8)

  $\mathbf{c}^{(t+1)} \leftarrow \left\{c_I^{(t+1)}\right\}_I$ and $\Theta^{(t)} \leftarrow (\mathbf{c}^{(t+1)}, V)$

**end**

Output $\mathbf{c}^{(T)}$.

with $V$ instead of $V^*$, we obviously cannot hope to produce $p$ for which $p(V^\top x)$ is exactly equal to the true low-rank polynomial $p_*(V^{*\top} x)$. But it is reasonable to hope for a $p$ for which the error incurred by $p$ is comparable to the inherent error $\beta$ contributed by the misspecified frame $V$. The main result of this section is to show that REALIGNPOLYNOMIAL can find such a $p$ given $V$:

**Theorem 38** *There are absolute constants $c_4, c_5, c_6, c_7, c_8 > 0$ such that the following holds for any $\underline{\epsilon}, \delta > 0$. Let $V \in St_r^n$, and let $(\mathbf{c}^*, V^*)$ be the realization of $\mathcal{D}$ for which $d_P(V, V^*) = \|V - V^*\|_F$. Suppose $d_P(V, V^*) \le (c_9 \cdot dr^3)^{-(d+1)/2}$.*

*Define $\mathbf{c}^{(T)}$ = REALIGNPOLYNOMIAL($\mathcal{D}, V, \underline{\epsilon}, \delta$), where in the specification of* REALIGNPOLYNOMIAL *we take*

$$\eta_{\mathsf{coef}} \triangleq \left(c_6 r d^4\right)^{d+1} \tag{11}$$

**Algorithm 5:** GEOSGD($\mathcal{D}, V^{(0)}, \epsilon, \delta$)

**Input:** Sample access to $\mathcal{D}$, $V^{(0)} \in \mathrm{St}_r^n$, target error $\epsilon$, failure probability $\delta$
**Output:** $\Theta = (\mathbf{c}^{(T)}, V^{(T)}) \in \mathcal{M}$ for which $d_P(V^{(T)}, V^*) \leq \epsilon$ and $\|\mathbf{c} - \mathbf{c}^*\|_2 \leq \epsilon$ for some realization $(\mathbf{c}^*, V^*)$ of $\mathcal{D}$

Define iteration count $T$ according to (7)
$\delta' \leftarrow \delta/(2T+1)$
**for** $0 \leq t < T$ **do**
  $\mathbf{c}^{(t)} \leftarrow$ REALIGNPOLYNOMIAL$(\mathcal{D}, V^{(t)}, \epsilon/2, \delta')$
  $V^{(t+1)} \leftarrow$ SUBSPACEDESCENT$(\mathcal{D}, V^{(t)}, \mathbf{c}^{(t)}, \delta')$
**end**
$\mathbf{c}^{(T)} \leftarrow$ REALIGNPOLYNOMIAL$(\mathcal{D}, V^{(T)}, \epsilon/2, \delta')$
Output $\Theta \triangleq (\mathbf{c}^{(T)}, V^{(T)})$.

$$T \triangleq c_5 \cdot \left(c_6 r d^4\right)^{d+1} \cdot \log(1/\underline{\epsilon}). \tag{12}$$

$$B \triangleq (c_8 \cdot \log(T/\delta))^{2d}. \tag{13}$$

*Then with probability at least* $1-\delta$*, we have that*

$$\|\mathbf{c}^{(T)} - \mathbf{c}^*\|_2 \leq \left(1 + c_7 \cdot (c_6 d r^4)^{-(d+1)/2}\right) \cdot \{\underline{\epsilon} \vee d_P(V, V^*)\}. \tag{14}$$

*Furthermore,* REALIGNPOLYNOMIAL *requires sample complexity*

$$N \triangleq O(B \cdot T) = \mathit{poly}\left(\log(1/\delta), r, d, \log\log(1/\underline{\epsilon})\right)^d \cdot \log(1/\underline{\epsilon})$$

*and runs in time* $n \cdot r^{O(d)} \cdot N$.

Before turning to the proof, we set some conventions. Henceforth, fix any $V, V^*$ satisfying the hypotheses of Theorem 38. Given coefficients $\mathbf{c}$ corresponding to the $r$-variate polynomial $p$, define $\delta_{\mathbf{c}} \triangleq p_* - p$. In light of (14), it will be convenient in our analysis to quantify, for an iterate $\mathbf{c}^{(t)}$, the extent to which $\|\mathbf{c}^{(t)} - \mathbf{c}^*\|_2$ differs from $d_P(V, V^*)$ via the (unknown) parameter

$$\rho_{\mathbf{c}^{(t)}} \triangleq \frac{d_P(V, V^*)}{\|\mathbf{c}^{(t)} - \mathbf{c}^*\|_2}.$$

For both $\delta_{\mathbf{c}}$ and $\rho_{\mathbf{c}}$, we will sometimes omit the subscript when the context is clear.

Note that we would like the eventual output $\mathbf{c}^{(T)}$ of REALIGNPOLYNOMIAL to have large $\rho$. The proof of Theorem 38 thus comes in two parts: 1) when $\rho_{\mathbf{c}^{(t)}}$ is small, the next $\rho_{\mathbf{c}^{(t+1)}}$ is larger by some margin, 2) when $\rho_{\mathbf{c}^{(t)}}$ is large, $\rho_{\mathbf{c}^{(t+1)}}$ may be smaller but will still be no smaller than the bound we are targeting in (14). Formally:

**Theorem 39** *Suppose* $d_P(V, V^*) \leq O(dr^3)^{-(d+1)/2}$. *For any* $\delta > 0$, *let* $\mathbf{c}$ *be an iterate in the execution of* REALIGNPOLYNOMIAL, *and let* $\mathbf{c}'$ *be the next iterate, given by*

$$c' \triangleq c - \frac{1}{B} \sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{\Theta, x_i}$$

*as defined in* (8) *for iid samples* $(x^0, y^0), \ldots, (x^{B-1}, y^{B-1}) \sim \mathcal{D}$. *If* $\eta_{\mathsf{coef}} \triangleq \Theta(dr^4)^{-d-1}$, *then with probability at least* $1-\delta$ *over the samples* $\{(x^i, y^i)\}_{i \in [B]}$,

1. *If* $\rho_{\mathbf{c}} \leq 1$, *then* $\rho_{\mathbf{c}'} \geq (1 + \Omega(dr^4)^{-d-1}) \cdot \rho_{\mathbf{c}}$.
2. *If* $\rho_{\mathbf{c}} \geq 1$ *then* $\rho_{\mathbf{c}'} \geq 1 - O(dr^4)^{-(d+1)/2}$.

We quickly verify that Theorem 39 implies Theorem 38.

**Proof** [Proof of Theorem 38] Take any iterate $\mathbf{c}^{(t)}$ in the execution of REALIGNPOLYNOMIAL. Taking $\delta$ to be $1/T$ times the error probability in Theorem 39, we have by a union bound over all $T$ iterations of REALIGNPOLYNOMIAL that with probability at least $1 - \delta$,

$$\rho_{\mathbf{c}^{(t+1)}} \geq \left\{1 - O(dr^4)^{-(d+1)/2}\right\} \wedge \left\{\rho_{\mathbf{c}^{(t)}} \cdot (1 + \Omega(dr^4)^{-d-1})\right\},$$

for every $0 \leq t < T$, which can be unrolled to give

$$\rho_{\mathbf{c}^{(T)}} \geq \left\{1 - O(dr^4)^{-(d+1)/2}\right\} \wedge \left\{\rho_{\mathbf{c}^{(0)}} \cdot (1 + \Omega(dr^4)^{-d-1})^T\right\}.$$

We can rewrite this inequality as

$$\|\mathbf{c}^{(t)} - \mathbf{c}^*\|_2 \leq \left\{\frac{d_P(V, V^*)}{1 - O(dr^4)^{-(d+1)/2}}\right\} \vee \left\{\|\mathbf{c}^{(0)} - \mathbf{c}^*\|_2 \cdot (1 + \Omega(dr^4)^{-d-1})^{-T}\right\}.$$

As we are initializing $\mathbf{c}^{(0)} = \mathbf{0}$, we have that $\|\mathbf{c}^{(0)} - \mathbf{c}^*\|_2 = \|\mathbf{c}^*\|_2 \leq r$. The theorem follows from taking $T = \Theta(dr^4)^{d+1} \cdot \log(r/\underline{\epsilon}) = \Theta(dr^4)^{d+1} \cdot \log(1/\underline{\epsilon})$. ■

As Theorem 39 suggests, we just need to analyze REALIGNPOLYNOMIAL on a per-iterate basis. Henceforth, fix an iterate $\mathbf{c}$; we will sometimes refer to the pair $(\mathbf{c}, V)$ as $\Theta$. Let $(x^0, y^0), ..., (x^{B-1}, y^{B-1}) \sim \mathcal{D}$ be the batch of samples drawn for the next iteration of REALIGNPOLYNOMIAL.

We first show that it suffices to prove that with high probability, the step $-\frac{1}{B}\sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x^i}$ is both 1) correlated with the direction $\mathbf{c} - \mathbf{c}^*$ in which we want to move, and 2) not too large. 1) and 2) can be interpreted respectively as curvature and smoothness of the gradient of the empirical risk in a neighborhood of our current iterate. Quantitatively, we claim that it suffices to show

**Lemma 40 (Local Curvature with High Probability)** *For any* $\delta > 0$ *and* $\gamma > 0$, *if* $B = \Omega(\log(1/\delta))^{2d} \cdot \gamma^{-2}$, *then we have that*

$$\frac{1}{B}\sum_{i=0}^{B-1} \left\langle \Delta_{\mathsf{coef}}^{x^t}, \mathbf{c} - \mathbf{c}^* \right\rangle \geq \upsilon_{\mathbf{c}}^{\mathsf{cu}} \cdot \eta_{\mathsf{coef}} \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \tag{15}$$

*for*

$$\upsilon_{\mathbf{c}}^{\mathsf{cu}} \triangleq 1 - \gamma\rho_{\mathbf{c}} - \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \left(O(r^{3/2}d) \cdot \rho_{\mathbf{c}}^2 + O(dr^3)^{(d+1)/2} \cdot \rho_{\mathbf{c}}(1 + \rho_{\mathbf{c}})\right)$$

*with probability at least* $1 - \delta$.

**Lemma 41 (Local Smoothness With High Probability)** *For any* $\delta > 0$, *if* $B = \Omega(\log(1/\delta))^{2d}$, *then we have that*

$$\left\|\frac{1}{B}\sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x_i}\right\|_2^2 \leq \upsilon_{\mathbf{c}}^{\mathsf{sm}} \cdot \eta_{\mathsf{coef}}^2 \|\mathbf{c} - \mathbf{c}^*\|_2^2 \quad \textit{for} \quad \upsilon_{\mathbf{c}}^{\mathsf{sm}} \triangleq O(dr^4)^{d+1} \cdot (1 \vee \rho_{\mathbf{c}}^2).$$

*with probability at least* $1 - \delta$.

We verify that Lemmas 40 and 41 are enough to prove Theorem 39.

**Proof** [Proof of Theorem 39] (Part 1) By (8) we have

$$\|\mathbf{c}' - \mathbf{c}^*\|_2^2 - \|\mathbf{c} - \mathbf{c}^*\|_2^2 = \left\| \frac{1}{B} \sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x^i} \right\|_2^2 - 2 \left\langle \frac{1}{B} \sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x^i}, \mathbf{c} - \mathbf{c}^* \right\rangle.$$

If the events of Lemmas 40 and 41 occur, then we get that

$$\|\mathbf{c}' - \mathbf{c}^*\|_2^2 - \|\mathbf{c} - \mathbf{c}^*\|_2^2 \leq \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot \left( \eta_{\mathsf{coef}} \upsilon_{\mathbf{c}}^{\mathsf{cu}} - \eta_{\mathsf{coef}}^2 \upsilon_{\mathbf{c}}^{\mathsf{sm}} \right),$$

If $\rho_{\mathbf{c}} \leq 1$, then we have that

$$\upsilon_{\mathbf{c}}^{\mathsf{cu}} \geq 1 - \gamma - \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot O(\rho_{\mathbf{c}}) \cdot O(dr^3)^{(d+1)/2} = 1 - \gamma - O(d_P(V, V^*)) \cdot O(dr^3)^{(d+1)/2},$$

so if we take $\gamma = 1/4$ and $d_P(V, V^*)_2 \leq O(dr^3)^{-(d+1)/2}$, then we ensure that $\upsilon_{\mathbf{c}}^{\mathsf{cu}} \geq 1/2$. Additionally, $\rho_{\mathbf{c}} \leq 1$ implies that $\upsilon_{\mathbf{c}}^{\mathsf{sm}} = O(dr^4)^{d+1}$. So if we take $\eta_{\mathsf{coef}} = \Theta(dr^4)^{-d-1}$, we conclude that

$$\|\mathbf{c}' - \mathbf{c}^*\|_2^2 \leq (1 - \eta_{\mathsf{coef}}/3) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \iff \rho_{\mathbf{c}'} \geq \rho_{\mathbf{c}} \cdot (1 - \eta_{\mathsf{coef}}/3)^{-1/2}$$

(Part 2) By triangle inequality,

$$\|\mathbf{c}' - \mathbf{c}^*\|_2 \leq \|\mathbf{c} - \mathbf{c}^*\|_2 + \left\| \frac{1}{B} \sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x^i} \right\|_2.$$

If Lemma 41 occurs, then we get that

$$\|\mathbf{c}' - \mathbf{c}^*\|_2 \leq \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot (1 + \eta_{\mathsf{coef}} \cdot \sqrt{\upsilon_{\mathbf{c}}^{\mathsf{sm}}}) = \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \left( 1 + \eta_{\mathsf{coef}} \cdot O(dr^4)^{(d+1)/2} \cdot \rho_{\mathbf{c}} \right),$$

or equivalently,

$$\rho_{\mathbf{c}'} \geq \rho_{\mathbf{c}} \cdot \left( 1 + \eta_{\mathsf{coef}} \cdot O(dr^4)^{(d+1)/2} \cdot \rho_{\mathbf{c}} \right)^{-1}. \tag{16}$$

For our choice of $\eta_{\mathsf{coef}} = \Theta(dr^4)^{-d-1}$, note that the quantity on the right-hand side of (16), as a function of $\rho_{\mathbf{c}}$, has minimum value $\left(1 + O(dr^4)^{-(d+1)/2}\right)^{-1}$ over $\rho_{\mathbf{c}} \in [1, \infty)$, attained by $\rho_{\mathbf{c}} = 1$, from which Part 2 of the theorem follows. ■

We now proceed to show local curvature and smoothness.

### D.1. Local Smoothness

In this section we show Lemma 41.

First, by Jensen's,

$$\left\| \frac{1}{B} \sum_{i=0}^{B-1} \Delta_{\mathsf{coef}}^{x_i} \right\|_2^2 \leq \frac{1}{B} \sum_{i=0}^{B-1} \|\Delta_{\mathsf{coef}}^{x^i}\|_2^2,$$

so to show Lemma 41 it suffices to bound the expectation and variance of the random variable $\|\Delta_{\mathsf{coef}}^{x}\|_2^2$ with respect to $x \sim \mathcal{N}(0, \mathbf{I}_n)$ and invoke Lemma 11.

We will need the following helper lemma which is a straightforward consequence of Lemma 37 and whose proof we defer to Appendix H.1.

**Lemma 42** *For any* $\Theta = (\mathbf{c}, V)$ *and* $\Theta^* = (\mathbf{c}^*, V^*)$, $\mathbb{E}[(F_x(\Theta) - F_x(\Theta^*))^4]^{1/2} \leq O(dr^3)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$.

We now use this to bound the expectation and variance of $\|\Delta^x_{\mathsf{coef}}\|_2^2$.

**Lemma 43** $\mathbb{E}[\|\Delta^x_{\mathsf{coef}}\|_2^2] \leq \eta^2_{\mathsf{coef}} \cdot O(dr^4)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$.

**Proof** By Cauchy-Schwarz,

$$\begin{aligned}\frac{1}{4\eta^2_{\mathsf{coef}}}\mathbb{E}\left[\|\Delta^x_{\mathsf{coef}}\|_2^2\right] &\leq \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right]^{1/2} \cdot \mathbb{E}\left[\left(\sum_I \phi_I(V^\top x)^2\right)^2\right]^{1/2} \\ &\leq O(dr^4)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2,\end{aligned}$$

where the second step follows by Lemma 42 and Lemma 21. ■

**Lemma 44** $\mathbb{E}[\|\Delta^x_{\mathsf{coef}}\|_2^4] \leq \eta^4_{\mathsf{coef}} \cdot O(dr^4)^{2d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^4$.

**Proof** Note that $(F_x(\Theta) - F_x(\Theta^*))^2$ and $\sum_I \phi_I(V^\top x)^2$ are degree-$2d$ polynomials in $x$. So by Cauchy-Schwarz,

$$\begin{aligned}\frac{1}{16\eta^4_{\mathsf{coef}}}\mathbb{E}\left[\|\Delta^x_{\mathsf{coef}}\|_2^4\right] &\leq \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^8\right]^{1/2} \cdot \mathbb{E}\left[\left(\sum_I \phi_I(V^\top x)^2\right)^4\right]^{1/2} \\ &\leq 3^{4d} \cdot \mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right] \cdot \mathbb{E}\left[\left(\sum_I \phi_I(V^\top x)^2\right)^2\right] \\ &\leq O(dr^4)^{2d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^4,\end{aligned}$$

where the second step follows by Fact 6, and the third step follows by Lemmas 42 and 21. ■

We are now ready to prove Lemma 41.

**Proof** [Proof of Lemma 41] Note that $\|\Delta^x_{\mathsf{coef}}\|_2^2$ is a polynomial of degree $2d$ in $x$. So by Lemma 11, Lemma 43, and Lemma 44, we see that

$$\frac{1}{B}\sum_{i=0}^{B-1}\|\Delta^{x^i}\|_2^2 \leq \eta^2_{\mathsf{coef}} \cdot O(dr^4)^{d+1} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \cdot \left(1 + \frac{1}{\sqrt{B}} \cdot O(\log(1/\delta))^d\right),$$

so the lemma follows by recalling that $\|V - V^*\|_F = d_P(V, V^*)$ so that

$$(\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \leq 4\|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot (1 \vee \rho_{\mathbf{c}}^2)$$

and taking $B = \Omega(\log(1/\delta))^{2d}$. ■

We note that this is one of the first of many places where the fact that one cannot obtain a $\mathbf{c}$ whose error is much smaller than the "misspecification error" $d_P(V, V^*)$ incurred by the subspace $V$ manifests: here, our bounds on the magnitudes of the gradient steps $\|\Delta^x_{\mathsf{coef}}\|$ inherently depend on $d_P(V, V^*)$, yet we require that the gradient steps have norm bounded by $\|\mathbf{c} - \mathbf{c}^*\|$.

### D.2. Local Curvature

We begin by outlining our argument for proving Lemma 40. It will be helpful to first decompose $\langle \Delta_{\mathsf{coef}}, \mathbf{c} - \mathbf{c}^* \rangle$ into "dominant" and "non-dominant" terms.

**Proposition 45** *For every monomial index $I$ and any $x \in \mathbb{R}^n$, let*

$$({\Delta'_{\mathsf{coef}}}^x)_I \triangleq -2\eta_{\mathsf{coef}} \cdot \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot \phi_I(V^\top x) \quad \textit{and} \quad ({\Delta''_{\mathsf{coef}}}^x)_I \triangleq -2\eta_{\mathsf{coef}} \cdot \mathfrak{R}^x \cdot \phi_I(V^\top x) \, \forall \, I.$$

*Then* $\Delta^x_{\mathsf{coef}} = {\Delta'_{\mathsf{coef}}}^x + {\Delta''_{\mathsf{coef}}}^x$.

**Proof** ${\Delta'_{\mathsf{coef}}}^x$ and ${\Delta''_{\mathsf{coef}}}^x$ correspond to the first-order and higher-order terms in the Taylor expansion of $\Delta^x_{\mathsf{coef}}$. Concretely, recall that

$$(\Delta^x_{\mathsf{coef}})_I = 2\eta_{\mathsf{coef}} \cdot (F_x(\Theta) - F_x(\Theta^*)) \cdot \phi_I(V^\top x).$$

We can decompose $\Delta^x_{\mathsf{coef}}$ by Taylor expanding the factor $F_x(\Theta) - F_x(\Theta^*)$ around $\Theta^* = \Theta$ to get

$$F_x(\Theta^*) - F_x(\Theta) = \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle + \mathfrak{R}^{\Theta, x} \text{ for } \mathfrak{R}^{\Theta,x} \triangleq \sum_{\ell=2}^{d+1} \frac{1}{\ell!} \left\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell} \right\rangle, \tag{17}$$

from which the proposition follows. ■

Motivated by Proposition 45, for any $x \in \mathbb{R}^n$ define

$$Y^x \triangleq \langle {\Delta'_{\mathsf{coef}}}^x, \mathbf{c} - \mathbf{c}^* \rangle, \quad \text{and} \quad E^x \triangleq \langle {\Delta''_{\mathsf{coef}}}^x, \mathbf{c} - \mathbf{c}^* \rangle.$$

To show Lemma 40, we will show that the random variables $\frac{1}{B}\sum_{i=0}^{B-1} Y^{x^i}$ and $\frac{1}{B}\sum_{i=0}^{B-1} E^{x^i}$ are respectively large and negligible with high probability. Eventually we will invoke the concentration inequalities of Lemmas 11 and 12 to control them, so we will compute the expectations (Section D.2.1) and variances (Section D.2.2) of their summands next.

#### D.2.1. LOCAL CURVATURE IN EXPECTATION

In this section we give bounds for $\mu_Y \triangleq \mathbb{E}_x[Y^x]$ and $\mu_E \triangleq \mathbb{E}_x[E^x]$ in the following two lemmas. Throughout this section, we will omit the superscript $x$ when the context is clear.

**Lemma 46** $\mu_Y \geq 2\eta_{\mathsf{coef}} \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \left(\|\mathbf{c} - \mathbf{c}^*\|_2 - O(r^{3/2}d) \cdot d_P(V, V^*)^2\right)$.

**Lemma 47** $|\mu_E| \leq 2\eta_{\mathsf{coef}} \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|_2)$.

In this section we will give the proof of Lemma 46; we will defer the proof of Lemma 47 to Appendix H.2.

**Proof** [Proof of Lemma 46] We have that

$$\langle \Delta'_{\mathsf{coef}}, \mathbf{c} - \mathbf{c}^* \rangle = -2\eta_{\mathsf{coef}} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot \delta(V^\top x) \tag{18}$$

Writing

$$\begin{aligned}\langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle &= \langle \nabla^{\mathsf{vec}} F_x(\Theta), V^* - V \rangle + \langle \nabla^{\mathsf{coef}} F_x(\Theta), \mathbf{c}^* - \mathbf{c} \rangle \\ &= x^\top (V^* - V)\nabla + \delta(V^\top x) \\ &= x^\top \Pi_V^\perp (V^* - V)\nabla + x^\top \Pi_V (V^* - V)\nabla + \delta(V^\top x) \\ &= x^\top \Pi_V^\perp V^* \nabla + x^\top \Pi_V \cdot (V^* - V)\nabla + \delta(V^\top x), \end{aligned} \tag{19}$$

we see that (18) is given by $2\eta_{\mathsf{coef}}$ times

$$\underbrace{\Big(\delta(V^\top x)\Big)^2}_{\text{(A)}} + \underbrace{\delta(V^\top x) \cdot \Big(x^\top \Pi_V (V^* - V)\nabla\Big)}_{\text{(B)}} + \underbrace{\delta(V^\top x) \cdot \Big(x^\top \Pi_V^\perp V^* \nabla\Big)}_{\text{(C)}} \tag{20}$$

Note that $x^\top \Pi_V$ and $x^\top \Pi_V^\perp$ are independent Gaussian vectors with mean zero and covariances $\Pi_V$ and $\Pi_V^\perp$ respectively. So we readily conclude that

**Observation 1** *For any $V$, the expectation of (C) with respect to $x$ vanishes.*

The following is also immediate:

**Observation 2** $\mathbb{E}[\text{(A)}] = \mathbb{E}_{g \sim \mathcal{N}(0, \mathbf{I}_r)}[\delta(g)^2] = \|\mathbf{c} - \mathbf{c}^*\|_2^2$.

We now turn to bounding $\mathbb{E}[\text{(B)}]$. We will make use of the following helper bound whose proof we defer to Appendix H.3

**Proposition 48** *If $\|V - V^*\| = d_P(V, V^*)$, then*

$$\mathbb{E}_g \left[ \Big( x^\top \Pi_V (V^* - V) \nabla p(V^\top x) \Big)^2 \right]^{1/2} \leq d_P(V, V^*)^2 \cdot O(r^{3/2} d).$$

**Lemma 49** $\mathbb{E}[\text{(B)}] \leq O(r^{3/2} d) \cdot d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2$.

**Proof** Note that

$$\begin{aligned} |\mathbb{E}[\text{(B)}]| &= \left| \mathbb{E} \left[ \delta(V^\top x) \cdot \Big( x^\top \Pi_V (V^* - V) \nabla p(V^\top x) \Big) \right] \right| \\ &\leq \mathbb{E}_g \left[ \delta(g)^2 \right]^{1/2} \cdot \mathbb{E}_g \left[ \Big( g^\top V^\top (V^* - V) \nabla p(g) \Big)^2 \right]^{1/2} \\ &\leq \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot d_P(V, V^*)^2 \cdot O(r^{3/2} d), \end{aligned}$$

where the second step follows by Cauchy-Schwarz, and the third by Proposition 48. ■

Lemma 46 now follows from (20), Observations 1 and 2, and Lemma 49. ■

### D.2.2. LOCAL CURVATURE WITH HIGH PROBABILITY

In this section, we complete the proof of Lemma 40 by establishing high-probability bounds for $Y^x$ and $E^x$. That is, we argue that with high probability, the dominant term given by $Y$ is large and the error from Taylor approximation is small. Specifically, we will show:

**Lemma 50** *For any $\delta > 0$ and $\gamma > 0$, if $B = \Omega(\log(1/\delta))^d \cdot O(\gamma^{-2})$, then*

$$\frac{1}{B}\sum_{i=1}^{B-1} Y^{x^i} \geq \eta_{\mathsf{coef}} \left( \|\mathbf{c} - \mathbf{c}^*\|_2^2 - O(r^{3/2}d) \cdot d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 - \gamma \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \right)$$

**Lemma 51** *For any $\delta > 0$, if $B = \Omega(\log(1/\delta))^{2d}$, then*

$$\left| \frac{1}{B}\sum_{i=0}^{B-1} E^{x^i} \right| \leq \eta_{\mathsf{coef}} \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|_2)$$

We defer their proofs to Appendices H.4 and H.5 respectively. We can finally deduce Lemma 54, completing the proof of Theorem 39 and thus Theorem 38.

**Proof** [Proof of Lemma 54] By Lemmas 50 and 51, and the earlier calculation showing that for any $x$, $\langle \Delta^x_{\mathsf{coef}}, V - V^* \rangle = Y^x + E^x$, we see that under our choice of $B$, (15) holds with probability $1 - 3\delta$. By replacing $3\delta$ with $\delta$, and absorbing the constant factors, the lemma follows. ■

## Appendix E. Guarantees for SUBSPACEDESCENT

Henceforth, fix a set of coefficients $\mathbf{c} \in \mathbb{R}^M$. In contrast with REALIGNPOLYNOMIAL, the aim of SUBSPACEDESCENT is to take a frame $V^{(0)}$ of a subspace which is somewhat close to the true subspace and refine it to some $V^{(T)}$ which is slightly closer, using only the misspecified coefficients $\mathbf{c}$. It turns out that if the misspecification error of $\mathbf{c}$ is comparable to the subspace distance from $V^{(0)}$ to the true subspace, SUBSPACEDESCENT can indeed accomplish this, and this is the main result of this section.

**Theorem 52** *There are absolute constants $c_{10}, c_{11} > 0$ and $c_{12} < 1/10$ such that the following holds for any $\delta > 0$. Let $V^{(0)} \in St_r^n$, and let $(\mathbf{c}^*, V^*)$ be the realization of $\mathcal{D}$ for which $d_P(V, V^*) = \|V - V^*\|_F$. Suppose*

$$d_P(V^{(0)}, V^*) \leq c_{12} \cdot \alpha_{\mathsf{ndg}} \cdot O(dr^3)^{-d-2}, \tag{21}$$

*Let $\mathbf{c}$ be a set of coefficients satisfying*

$$d_P(V^{(0)}, V^*) \geq \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2 \tag{22}$$

*Define $V^{(T)} = \text{SUBSPACEDESCENT}(\mathcal{D}, V^{(0)}, \mathbf{c}, \delta)$, where in the specification of SUBSPACEDESCENT we take*

$$\eta_{\mathsf{vec}} \triangleq \frac{\alpha_{\mathsf{ndg}}}{T \cdot n} \left( c_{11} \cdot dr^3 \ln(T/\delta) \right)^{-d-2} \tag{23}$$

$$T \triangleq \left( \frac{r}{\alpha_{\mathsf{ndg}}} \right)^2 \cdot (c_{10} \cdot d \cdot \log(1/\delta))^{2c_1 d}. \tag{24}$$

*Then with probability at least* $1-\delta$*, we have that*

$$1 - \frac{d_P(V^{(T)}, V^*)^2}{d_P(V^{(0)}, V^*)^2} \geq \frac{\alpha_{\mathsf{ndg}}}{n} \cdot \mathit{poly}(\ln(1/\alpha_{\mathsf{ndg}}), r, d, \ln(1/\delta))^{-d}.$$

*Furthermore,* SUBSPACEDESCENT *draws* $N \triangleq O(T)$ *samples and runs in time* $n \cdot r^{O(d)} \cdot N$*.*

Henceforth, let $\delta, V^{(0)}, V^*, \mathbf{c}, \mathbf{c}^*, T, \eta_{\mathsf{vec}}$ satisfy the hypotheses of Theorem 52.

As discussed in Section 2.2.3, a single execution of SUBSPACEDESCENT should be thought of as a single step of stochastic gradient descent over a batch of size $T$. The only difference lies in the fact that the empirical risk we work with in each iteration of SUBSPACEDESCENT is slightly different, as our subspace estimate $V^{(t)}$ continues to update by a small amount. So just as we analyzed the individual steps of REALIGNPOLYNOMIAL in Lemma 39 via local curvature and smoothness estimates, we would like to do the same for an entire execution of SUBSPACEDESCENT. That is, we want to show that with high probability, the steps $-\Delta_{\mathsf{vec}}^{\Theta^t, x^t}$ are 1) bounded, and 2) each correlated with the direction $V^* - V^{(t)}$ in which we want to move. Quantitatively, we claim that it suffices to show

**Lemma 53 (Local Smoothness With High Probability)**

$$\|V^{(0)} - V^{(T)}\|_F^2 \leq \eta_{\mathsf{vec}}^2 \cdot O(dr^3 \ln(T/\delta))^{d+2} \cdot O(n) \cdot d_P(V^{(0)}, V^*)^2.$$

*with probability at least* $1-\delta$*.*

**Lemma 54 (Local Curvature with High Probability)**

$$\sum_{t=0}^{T-1} \left\langle \Delta_{\mathsf{vec}}^{\Theta^{(t)}, x^t}, V^{(t)} - V^* \right\rangle \geq T \cdot \eta_{\mathsf{vec}} \cdot (\alpha_{\mathsf{ndg}}/4) \cdot d_P(V^{(0)}, V^*)^2$$

*with probability at least* $1-\delta$*.*

We verify that Lemmas 54 and 53 are enough to prove Theorem 52.

**Proof** [Proof of Theorem 52] For every $0 \leq t < T$, we have

$$\|V^{(t+1)} - V^*\|_F^2 - \|V^{(t)} - V^*\|_F^2 = \|\Delta_{\mathsf{vec}}^{\Theta^{(t)}, x^t}\|_F^2 - 2\left\langle \Delta_{\mathsf{vec}}^{\Theta^{(t)}, x^t}, V^{(t)} - V^* \right\rangle. \tag{25}$$

If the event of Lemma 54 holds, then

$$\sum_{t=0}^{T-1} \left\langle \Delta_{\mathsf{vec}}^{\Theta^{(t)}, x^t}, V^{(t)} - V^* \right\rangle \geq T \cdot (\alpha_{\mathsf{ndg}}/4) \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2.$$

If the event of Lemma 53 holds, then

$$\begin{aligned}\sum_{t=0}^{T-1} \|\Delta_{\mathsf{vec}}^{\Theta^{(t)}, x^t}\|_F^2 &\leq T \cdot \eta_{\mathsf{vec}}^2 \cdot O(dr^3 \ln(T/\delta))^{d+2} \cdot O(n) \cdot d_P(V^{(0)}, V^*)^2 \\ &\leq O\left(\alpha_{\mathsf{ndg}} \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2\right).\end{aligned}$$

where the last step follows by the choice of $\eta_{\mathsf{vec}}$ in (23), and the constant factor in the last expression can be made arbitrarily small. By summing (25) over $t$, telescoping, and recalling that $\|V^{(0)} - V^*\|_F^2 = d_P(V^{(0)}, V^*)^2$, we conclude that

$$\|V^{(T)} - V^*\|_2^2 - d_P(V^{(0)}, V^*)^2 \leq -T \cdot (\alpha_{\mathsf{ndg}}/5) \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2,$$

from which we get, because $d_P(V^{(T)}, V^*) \leq \|V^{(T)} - V^*\|_F$, that

$$1 - \frac{d_P(V^{(T)}, V^*)^2}{d_P(V^{(0)}, V^*)^2} \geq T \cdot (\alpha_{\mathsf{ndg}}/5) \cdot \eta_{\mathsf{vec}}.$$

The claim follows by substituting the choice of $\eta_{\mathsf{vec}}$ and $T$ in (23) and (24). ■

We now proceed to show Lemma 53 and 54.

### E.1. Local Smoothness

In this section we establish Lemma 53. We also show that $d_P(V^{(t)}, V^*)$ does not change much, both in expectation (Lemma 58) and with high probability (Lemma 57), as $t$ varies. While we have already seen that Lemma 53 is needed to prove Theorem 52, Lemmas 57 and 58 will be crucial to our arguments in later sections, where we argue that at each step $t$ we make progress scaling with the distance $d_P(V^{(t)}, V^*)$ and thus need that this distance is comparable to the initial distance $d_P(V^{(0)}, V^*)$.

For a fixed $\Theta$, we will first show a high-probability bound on the norm of $\Delta_{\mathsf{vec}}^{\Theta,x}$, that is, we bound the size of the step made in a single iteration inside SUBSPACEDESCENT.

Where the context is clear, we will suppress superscript $\Theta, x$. Then very naively, using the inequalities $1 - \cos(x) \leq x$ and $|\sin(x)| \leq x$ for all $x \geq 0$, we have

$$\|\Delta_{\mathsf{vec}}\|_F \leq (1 - \cos(\sigma\eta_{\mathsf{vec}}))(2\sqrt{r}) + |\sin(\sigma\eta_{\mathsf{vec}})| \leq 2\sqrt{r} \cdot \sigma\eta_{\mathsf{vec}} + \sigma\eta_{\mathsf{vec}} \leq 3\sqrt{r} \cdot \sigma\eta_{\mathsf{vec}}. \quad (26)$$

We first bound the moments of $\sigma^2$.

**Lemma 55** *For all integers $q \geq 1$, $\mathbb{E}[\sigma^{2q}]^{1/q} \leq O(nrd) \cdot O(q^2 dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2$.*

**Proof** Recall that $\sigma = 2(F_x(\Theta) - F_x(\Theta^*)) \cdot \|\Pi_V^\perp x\|_2 \cdot \|\nabla p(V^\top x)\|_2$. So by Cauchy-Schwarz,

$$\mathbb{E}[\sigma^{2q}]^{1/q} \leq 4\mathbb{E}[(F_x(\Theta) - F_x(\Theta^*))^{4q}]^{1/2q} \cdot \mathbb{E}\left[\|\Pi_V^\perp x\|_2^{4q} \cdot \|\nabla p(V^\top x)\|_2^{4q}\right]^{1/2q}. \quad (27)$$

The second factor in (27) is simply

$$\begin{aligned}
&\mathbb{E}_{g' \sim \mathcal{N}(0, \Pi_V^\perp)}[\|g'\|_2^{4q}]^{1/2q} \cdot \mathbb{E}_{g \sim \mathcal{N}(0, Id_r)}[\|\nabla p(g)\|_2^{4q}]^{1/2q} \\
&\quad \leq ((2q-1) \cdot (n - r + 1)) \cdot \left(rd \cdot (4q-1)^d \cdot \mathrm{Var}[p]\right) \\
&\quad \leq O(n) \cdot qrd \cdot (4q)^d \cdot \mathrm{Var}[p] \\
&\quad \leq O(n) \cdot rd \cdot (4q)^{d+1},
\end{aligned}$$

where in the first step we used Corollary 16 and Lemma 18, and in the last step we used Fact 1 and triangle inequality to bound $\mathrm{Var}[p] = O(1)$.

For the first factor in (27), we have that

$$\begin{aligned}\mathbb{E}\left[(F_x(\Theta)-F_x(\Theta^*))^{4q}\right]^{1/2q} &\le (2q-1)^d\cdot\mathbb{E}\left[(F_x(\Theta)-F_x(\Theta^*))^4\right]^{1/2}\\ &\le O(qdr^3)^{d+1}\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)^2\end{aligned}$$

by Fact 6 and Lemma 42 respectively, from which the claim follows. ■

As a result, the random variable $\sigma^2$ enjoys sub-Weibull-type concentration.

**Corollary 56** *For any $0<\delta'<1$, let $\tau=\Omega(\ln(1/\delta'))^{d+2}$. Then*

$$\Pr\left[\sigma^2\ge\tau\cdot\Omega(n)\cdot\Omega(dr^3)^{d+2}\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)^2\right]\le\delta'.$$

**Proof** Let $\gamma\triangleq n\cdot O(rd^3)^{d+2}\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)^2$. We wish to apply Lemma 11 to $\sigma^2$, which is a degree-$4d$ polynomial in $x$. By Lemma 55 above, $\mathbb{E}[\sigma^2]\le O(\gamma)$ and $\mathrm{Var}[\sigma^2]\le\mathbb{E}[\sigma^4]\le O(\gamma^2)$. By Lemma 11 specialized to $T=1$,

$$\Pr\left[\sigma^2\ge O(\log(1/\delta'))^{2d}\cdot\gamma\right]\le\delta',$$

from which the lemma follows. ■

From (26) we conclude that for any $0<\delta<1$,

$$\|\Delta_{\mathsf{vec}}\|_F\le 3\sqrt{r}\cdot\eta_{\mathsf{vec}}\cdot O(\ln(1/\delta))^{(d+2)/2}\cdot O(\sqrt{n})\cdot O(dr^3)^{(d+2)/2}\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2) \quad (28)$$

with probability at least $1-\delta$.

Now consider the sequence of iterates $\{\Theta^{(t)}\}_{0\le t\le T}$ in SUBSPACEDESCENT. In this subsection alone, for convenience define

$$\alpha\triangleq 3\sqrt{r}\cdot\eta_{\mathsf{vec}}\cdot O(\ln(1/\delta))^{(d+2)/2}\cdot O(\sqrt{n})\cdot O(dr^3)^{(d+2)/2}$$

For every $0\le t<T$, let $\mathcal{E}_t$ be the event that (28) holds for $\Delta_{\mathsf{vec}}^{\Theta^{(t)},x^t}$, that is, that $\|\Delta_{\mathsf{vec}}^{\Theta^{(t)},x^t}\|_F\le\alpha(\|V^{(t)}-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)$. If $\mathcal{E}_t$ held for every $t$, then by triangle inequality and induction, we would have that for every $0\le t<T$,

$$\begin{aligned}\|\Delta_{\mathsf{vec}}^{\Theta^{(t)},x^t}\|_F &\le \alpha\left(\|V^{(0)}-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2+\sum_{s=0}^{t-1}\|\Delta_{\mathsf{vec}}^{\Theta^{(s)},x^s}\|_F\right)\\ &\le \alpha(1+\alpha)^t\left(\|V^{(0)}-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2\right)\\ &= \alpha(1+\alpha)^t\left(d_P(V^{(0)},V^*)+\|\mathbf{c}-\mathbf{c}^*\|_2\right)\\ &\le 3\alpha(1+\alpha)^t\cdot d_P(V^{(0)},V^*),\end{aligned}$$

where the last step follows by (22). So

$$\sum_{t=0}^{T-1}\|\Delta_{\mathsf{vec}}^{\Theta^{(t)},x^t}\|_F\le 3\left((1+\alpha)^T-1\right)\cdot d_P(V^{(0)},V^*). \quad (29)$$

Taking $\delta'$ in Corollary 56 to be $\delta/T$ and applying a union bound, we deduce by monotonicity of $L_p$ norms that Lemma 53 holds for our choice of $\eta_{\mathsf{vec}},T$. We also deduce the following crude bound.

**Lemma 57** *$\|V^{(t)} - V^*\|_F \in [0.9, 1.1] \cdot d_P(V^{(0)}, V^*)$ for every $0 \le t \le T$ with probability at least $1 - \delta$.*

This modest level of control over how much the distance to the true subspace fluctuates over the course of SUBSPACEDESCENT will be sufficient for our subsequent analysis.

We pause to note that the assumption that the "misspecification error" $\|\mathbf{c} - \mathbf{c}^*\|_2$ incurred by the coefficients $\mathbf{c}$ must, by (22), be small relative to the subspace distance error incurred by the initial subspace $V^{(0)}$ is crucial here. Indeed, our bounds for the moments of $\sigma^2$, i.e. the moments of the size of the gradient steps, inherently scale with $\|\mathbf{c} - \mathbf{c}^*\|$, yet we need local smoothness in the sense that the gradient steps have norm comparable to $d_P(V^{(0)}, V^*)$.

Lastly, it will be useful to establish bounds on the moments of $\|V^{(t)} - V^*\|_F$ for each $t$.

**Lemma 58** *For any absolute, integer-valued constant $q \ge 1$, $\mathbb{E}\left[\|V^{(t)} - V^*\|_F^q\right] \le 1.1 \cdot d_P(V^{(0)}, V^*)^q$ for every $0 \le t < T$, where the expectation is in the randomness of the samples $x^0, ..., x^{T-1}$ drawn in* SUBSPACEDESCENT.

We defer the proof of this to Appendix I.2.

### E.2. Local Curvature

We begin by outlining our argument for proving Lemma 54. As with the proof of Lemma 40 for REALIGNPOLYNOMIAL, it will be helpful to first decompose $\langle \Delta_{\mathsf{vec}}, V - V^* \rangle$ into "dominant" and "non-dominant" terms. Here the "non-dominant" terms will be more complicated because of the trigonometric corrections associated with geodesic gradient descent.

**Proposition 59** *For any $\Theta, x$, define*

$$\Delta'_{\mathsf{vec}}{}^{\Theta,x} \triangleq -2\eta_{\mathsf{vec}} \cdot \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot \Pi_V^{\perp} \cdot x \cdot (\nabla^{\Theta,x})^\top \quad \textit{and} \quad \Delta''_{\mathsf{vec}}{}^{\Theta,x} \triangleq -2\eta_{\mathsf{vec}} \cdot \mathfrak{R}^{\Theta,x} \cdot \Pi_V^{\perp} \cdot x \cdot (\nabla^{\Theta,x})^\top$$

*and also*

$$\begin{aligned}\mathcal{E}^{\Theta,x} &\triangleq \Delta_{\mathsf{vec}}^{\Theta,x} - \Delta'_{\mathsf{vec}}{}^{\Theta,x} - \Delta''_{\mathsf{vec}}{}^{\Theta,x} \\ &= \left(\cos(\sigma^{\Theta,x}\eta_{\mathsf{vec}}) - 1\right) V \cdot \widehat{\nabla}^{\Theta,x} (\widehat{\nabla}^{\Theta,x})^\top + \left(\sin(\sigma^{\Theta,x}\eta_{\mathsf{vec}}) - \sigma^{\Theta,x}\eta_{\mathsf{vec}}\right) \widehat{h}^{\Theta,x} (\widehat{\nabla}^{\Theta,x})^\top.\end{aligned}$$

*Then $\Delta_{\mathsf{vec}}^{\Theta,x} = \Delta'_{\mathsf{vec}}{}^{\Theta,x} + \Delta''_{\mathsf{vec}}{}^{\Theta,x} = \mathcal{E}^{\Theta,x}$.*

**Proof** $\widehat{\Delta}_{\mathsf{vec}}^{\Theta,x} \triangleq \Delta'_{\mathsf{vec}}{}^{\Theta,x} + \Delta''_{\mathsf{vec}}{}^{\Theta,x}$ is the lowest-order term in the Taylor expansion of $\Delta_{\mathsf{vec}}^{\Theta,x}$ around $\eta_{\mathsf{vec}} = 0$, given by

$$\widehat{\Delta}_{\mathsf{vec}}^{\Theta,x} \triangleq \eta_{\mathsf{vec}} \cdot h^{\Theta,x} (\nabla^{\Theta,x})^\top.$$

Recalling the factor $F_x(\Theta) - F_x(\Theta^*)$ in the definition of $h$ in (9), we Taylor expand around $\Theta^* = \Theta$ to get (17) from Section D and therefore the decomposition of $\widehat{\Delta}_{\mathsf{vec}}^{\Theta,x}$ into $\Delta'_{\mathsf{vec}}{}^{\Theta,x}$ and $\Delta''_{\mathsf{vec}}{}^{\Theta,x}$. ■

$\widehat{\Delta}'_v$ Motivated by Proposition 59, for any $x \in \mathbb{R}^n$ and $\Theta = (\mathbf{c}, V)$ define

$$X^{\Theta,x} \triangleq \langle (\widehat{\Delta}'_{\mathsf{vec}})^{\Theta,x}, V - V^* \rangle, \quad E_1^{\Theta,x} \triangleq \langle (\widehat{\Delta}''_{\mathsf{vec}})^{\Theta,x}, V - V^* \rangle, \quad E_2^{\Theta,x} \triangleq \langle \mathcal{E}^{\Theta,x}, V - V^* \rangle.$$

Consider a sequence of iid samples $(x^0, y^0), ..., (x^{T-1}, y^{T-1}) \sim \mathcal{D}$ and iterates $\Theta^{(0)}, ..., \Theta^{(T-1)}$ in the execution of SUBSPACEDESCENT, where each $\Theta^{(t)}$ is given by $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$. To show

Lemma 54, we will show that the random variable $\sum_{t=0}^{T-1} X^{\Theta^{(t)},x^t}$ is large with high probability, while the random variables $\sum_{t=0}^{T-1} E_1^{\Theta^{(t)},x^t}$, and $\sum_{t=0}^{T-1} E_2^{\Theta^{(t)},x^t}$ are negligible with high probability. Eventually, we will invoke the martingale concentration inequalities of Lemmas 13 and 14 to control them. Before that, we first need to compute their expectations.

### E.2.1. LOCAL CURVATURE IN EXPECTATION- SINGLE STEP

In this section we give bounds on the *expected* correlation between the direction in which we would like to move, and a step taken in a *single* iteration in SUBSPACEDESCENT.

Given an iterate $\Theta = (\mathbf{c}, V)$, let $\mu_X(\Theta), \mu_{E_1}(\Theta), \mu_{E_2}(\Theta)$ be the expectations $\mathbb{E}[X^{\Theta,x}]$, $\mathbb{E}[E_1^{\Theta,x}]$, $\mathbb{E}[E_2^{\Theta,x}]$ with respect to $x \sim \mathcal{N}(0, \mathbf{I}_n)$. In this section we will bound these quantities in terms of the distance between $\Theta$ and $(\mathbf{c}^*, V^*)$. As usual, we will omit the superscript $\Theta, x$ when the context is clear.

**Lemma 60** $\mu_X(\Theta) \geq 2\eta_{\mathsf{vec}} \cdot (\alpha_{\mathsf{ndg}}/4) \cdot d_P(V, V^*)^2$.

**Lemma 61**

$$|\mu_{E_1}(\Theta)| \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \|V - V^*\|_F \cdot d_P(V, V^*) \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)\,.$$

**Lemma 62** *If* $\eta_{\mathsf{vec}} \leq O(1/n)$, *then*

$$|\mu_{E_2}(\Theta)| \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2\,.$$

At this point we pause to emphasize that Lemma 60 is the key reason why we must work with $\mathrm{G}(n, r)$ and not simply with the Euclidean space of $n \times r$ matrices, as Lemma 60 says that the local curvature with respect to the empirical risk in a neighborhood of a subspace $V$ is dictated solely by its Procrustes distance to $V^*$ rather than by $\|V - V^*\|_F$.

Additionally, note that once again, (22) is essential here, to ensure that the expectations from Lemmas 61 and 62 of the "non-dominant" terms do not overwhelm the expectation from Lemma 68 of the "dominant" term, which only depends on $d_P(V, V^*) \sim d_P(V^{(0)}, V^*)$.

We now turn to proving Lemma 60.

**Proof** [Proof of Lemma 60] Fix a sample $(x, y) \sim \mathcal{D}$. We have that

$$\begin{aligned}\langle \widehat{\Delta}'_{\mathsf{vec}}, V - V^* \rangle &= -2\eta_{\mathsf{vec}} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot x^\top \Pi_V^\perp (V - V^*) \nabla \\ &= 2\eta_{\mathsf{vec}} \langle \nabla F_x(\Theta), \Theta^* - \Theta \rangle \cdot x^\top \cdot \Pi_V^\perp V^* \cdot \nabla \end{aligned} \tag{30}$$

By (19) we see that (30) is given by $2\eta_{\mathsf{vec}}$ times

$$\underbrace{\left(x^\top \Pi_V^\perp V^* \nabla\right)^2}_{\text{(A')}} + \underbrace{\left(x^\top \Pi_V (V^* - V) \nabla\right) \cdot \left(x^\top \Pi_V^\perp V^* \nabla\right)}_{\text{(B')}} + \underbrace{\delta(V^\top x) \cdot \left(x^\top \Pi_V^\perp V^* \nabla\right)}_{\text{(C')}}\,. \tag{31}$$

As in the proof of Lemma 46, note that $x^\top \Pi_V$ and $x^\top \Pi_V^\perp$ are independent Gaussan random vectors with mean zero and covariances $\Pi_V$ and $\Pi_V^\perp$ respectively. So we immediately conclude that

**Observation 3** *For any* $V$, *the expectations of* (B') *and* (C') *with respect to* $x$ *vanish.*

We next bound $\mathbb{E}[\text{Ⓐ'}]$.

**Lemma 63** $(\alpha_{\mathsf{ndg}}/4) \cdot d_P(V, V^*)^2 \leq \mathbb{E}[\text{Ⓐ'}] \leq 4d_P(V, V^*)^2$.

**Proof** Note that

$$\begin{aligned}
\mathbb{E}[\text{Ⓐ'}] &= \mathbb{E}\left[\left(x^\top \Pi_V^\perp V^* \nabla\right)^2\right] \\
&= \mathbb{E}_{\substack{h\sim\mathcal{N}(0,\Pi_V)\\ h_\perp\sim\mathcal{N}(0,\Pi_V^\perp)}}\left[\nabla p(V^\top h)^\top V^{*\top} h_\perp h_\perp^\top V^* \nabla p(V^\top h)\right] \\
&= \mathbb{E}_{h\sim\mathcal{N}(0,\Pi_V)}\left[\nabla p(V^\top h)^\top V^{*\top} \Pi_V^\perp V^* \nabla p(V^\top h)\right] \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}\left[\nabla p(g)^\top \cdot \left(\mathbf{I} - V^{*\top} V V^\top V^*\right) \cdot \nabla p(g)\right] \\
&= \left\langle \mathbb{E}_g\left[\nabla p(g) \nabla p(g)^\top\right], \mathbf{I} - V^{*\top} V V^\top V^* \right\rangle \qquad (32)
\end{aligned}$$

where we used independence of $h, h_\perp$ in the third step. We wil need the following bound.

**Lemma 64** *If* $\|\mathbf{c} - \mathbf{c}^*\|_2 \leq O(r^{-3/2}d^{-1})$*, then we have that*

$$(\alpha_{\mathsf{ndg}}/2) \cdot \mathbf{I}_r \preceq \mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}\left[\nabla p(g) \nabla p(g)^\top\right] \preceq 2 \cdot \mathbf{I}_r.$$

**Proof** For convenience, let $M$ and $M_*$ denote $\mathbb{E}\left[\nabla p(g) \nabla p(g)^\top\right]$ and $\mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}\left[\nabla p_*(g) \nabla p_*(g)^\top\right]$ respectively. For any $v \in \mathbb{S}^{r-1}$, we have that

$$\begin{aligned}
|v^\top M_* v - v^\top M v| &= \left|\mathbb{E}\left[\langle v, \nabla p_*(g)\rangle^2 - \langle v, \nabla p(g)\rangle^2\right]\right| \\
&= \left|\mathbb{E}\left[\langle v, \nabla \delta(g)\rangle \cdot \langle v, \nabla(p + p_*)(g)\rangle\right]\right| \\
&\leq \mathbb{E}\left[\|\nabla\delta(g)\|_2^2\right]^{1/2} \cdot \left(\mathbb{E}\left[\|\nabla p(g)\|_2^2\right]^{1/2} + \mathbb{E}\left[\|\nabla p_*(g)\|_2^2\right]^{1/2}\right) \\
&\leq rd \cdot \mathrm{Var}[\delta]^{1/2} \cdot (\mathrm{Var}[p]^{1/2} + \mathrm{Var}[p_*]^{1/2}) \\
&< O(r^{3/2} d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2),
\end{aligned}$$

where in the third step we used Cauchy-Schwarz, in the fourth step we used Lemma 18, and in the last step we upper bounded $\mathrm{Var}[p]$ and $\mathrm{Var}[p_*]$ by $O(r)$ using Corollary 1 and the fact that $\|\mathbf{c} - \mathbf{c}^*\|_2 = O(1)$. ■

To conclude the proof of Lemma 63, we see that

$$\begin{aligned}
\mathbb{E}[\text{Ⓐ'}] &\in [\alpha_{\mathsf{ndg}}/2, 2] \cdot \mathrm{Tr}(\mathbf{I} - V^{*\top} V V^\top V^*) \\
&= [\alpha_{\mathsf{ndg}}/2, 2] \cdot d_C(V, V^*)^2 \\
&\in [\alpha_{\mathsf{ndg}}/4, 4] \cdot d_P(V, V^*)^2, \qquad (33)
\end{aligned}$$

where the first step follows by (32) and Lemma 64, the second step follows by the fact that $\mathrm{Tr}(\mathbf{I} - V^{*\top} V V^\top V^*) = d - \|V^{*\top} V\|_F^2$, and the last step follows by Lemma 24. ■

Lemma 60 now follows from (31), Observation 3, and Lemma 63. ■

We defer the proofs of Lemmas 61 and 62, to Appendix I.

### E.2.2. LOCAL CURVATURE IN EXPECTATION- ALL ITERATIONS

In this section we extend the results of the previous section to give bounds on the sum *over all* $t$ of the expected correlations between the direction in which we would like to move at time $t$, and the step we actually take at time $t$.

Specifically, for the sequence of iterates $\{\Theta^{(t)}\}_{0\le t\le T}$ in SUBSPACEDESCENT, we would like to bound $\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_X(\Theta^{(t)})\right]$, $\left|\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_{E_1}(\Theta^{(t)})\right]\right|$, and $\left|\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_{E_2}(\Theta^{(t)})\right]\right|$. We emphasize that the expectation here is over the randomness of the samples $x^0,...,x^{T-1}$, so e.g. $\mu_X(\Theta^{(t)})$ is a random variable depending on $x^0,...,x^{t-1}$ and is itself an expectation over the next sample $x^t$.

Intuitively, for our choice (23) of small step size $\eta_{\mathsf{vec}}$ which scales with $O(1/T)$, Lemma 58 suggests that the expected behavior of the corresponding martingales should not be very different from that of a sum of iid random variables. That is, these expected sums should be not much different than $T$ times the expectation of their *first* summand, corresponding to the first iteration which takes a step from $\Theta^{(0)}$. In Lemmas 65, 66, and 67, we show that this is indeed the case:

**Lemma 65** $\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_X(\Theta^{(t)})\right] \le T\cdot\eta_{\mathsf{vec}}\cdot(\alpha_{\mathsf{ndg}}/2.2)\cdot d_P(V^{(0)},V^*)^2.$

**Lemma 66** $\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_{E_1}(\Theta^{(t)})\right] \le T\cdot O(\eta_{\mathsf{vec}})\cdot O(dr^3)^{(d+1)/2}\cdot d_P(V^{(0)},V^*)^3.$

**Lemma 67** $\mathbb{E}\left[\sum_{t=0}^{T-1}\mu_{E_2}(\Theta^{(t)})\right] \le T\cdot O(\eta_{\mathsf{vec}})\cdot O(dr^3)^{d+2}\cdot d_P(V^{(0)},V^*)^3.$

We defer their proofs to Appendices I.5, I.6, and I.7 respectively.

### E.2.3. LOCAL CURVATURE WITH HIGH PROBABILITY

In this section, we complete the proof of Lemma 54 by establishing high-probability bounds for the MDS's corresponding to $X$, $E_1$, and $E_2$. That is, we argue that with high probability, the dominant term given by $X$ is large, while the error terms from Taylor approximation and from the trigonometric corrections are small. Specifically, we show:

**Lemma 68**

$$\sum_{t=0}^{T-1} X^{\Theta^{(t)},x^t} \ge T\cdot\eta_{\mathsf{vec}}\cdot(\alpha_{\mathsf{ndg}}/3)\cdot d_P(V^{(0)},V^*)^2$$

*with probabiliy at least* $1-\delta$.

**Lemma 69**

$$\left|\sum_{t=0}^{T-1} E_1^{\Theta^{(t)},x^t}\right| \le T\cdot\eta_{\mathsf{vec}}\cdot(c_{12}\cdot\alpha_{\mathsf{ndg}})\cdot d_P(V^{(0)},V^*)^2$$

*with probability at least* $1-\delta$.

**Lemma 70**

$$\left|\sum_{t=0}^{T-1} E_2^{\Theta^{(t)},x^t}\right| \le T\cdot\eta_{\mathsf{vec}}\cdot(c_{12}\cdot\alpha_{\mathsf{ndg}})\cdot d_P(V^{(0)},V^*)^2$$

*with probability at least* $1-\delta$.

We defer their proofs to Appendix I.8.4. The key technical step in all three proofs is to upper bound the variance of the martingale differences, after which one can invoke the corresponding expectation bounds from Section E.2.2 together with the martingale concentration inequalities of Lemma 14 for Lemma I.8 and Lemma 13 for Lemmas 69 and 70. We emphasize that here we must again crucially use (22), this time to ensure that the variances of the martingale differences, which depend in part on $\|\mathbf{c} - \mathbf{c}^*\|_2$, do not swamp the expectation $\mu_X(\Theta)$ of the dominant term.

Also, we remark that it is in the proof of Lemma 69 and Lemma 70 that we finally use the assumption (21) that $d_P(V^{(0)}, V^*)$ is somewhat small.

Finally, we can deduce Lemma 54, completing the proof of Theorem 52.

**Proof** [Proof of Lemma 54] By Lemmas 68, 69, and 70, and the earlier calculation showing that for any $\Theta = (\mathbf{c}, V)$, $\langle \Delta_{\mathsf{vec}}^{\Theta,x}, V - V^* \rangle = X^{\Theta,x} + E_1^{\Theta,x} + E_2^{\Theta,x}$, we see that under our choice of $T, \eta_{\mathsf{vec}}$,

$$\sum_{t=0}^{T-1} \left\langle \Delta_{\mathsf{vec}}^{\Theta^{(t)},x^t}, V^{(t)} - V^* \right\rangle \geq \alpha_{\mathsf{ndg}} \left( \frac{1}{3} - 2c_{12} \right) \cdot T \cdot \eta_{\mathsf{vec}} \cdot d_P(V^{(0)}, V^*)^2$$

with probability $1 - 3\delta$. By replacing $3\delta$ with $\delta$, and absorbing the constant factors, the lemma follows. ■

## Appendix F. Putting Everything Together for GEOSGD

In this section we conclude the proof of Theorem 34 using Theorems 38 and 52.

There is one last subtlety we must address. In Theorem 38 on the distance $\|\mathbf{c} - \mathbf{c}^*\|$ between the coefficients $\mathbf{c}$ output by REALIGNPOLYNOMIAL and the true coefficients $\mathbf{c}^*$, the upper bound is at best only in terms of the known parameter $\underline{\epsilon}$. On the other hand, in Theorem 52 on the error $d_P(V^{(T)}, V^*)$ incurred by the subspace $V^{(T)}$ output by SUBSPACEDESCENT when initialized to $V^{(0)}$, the upper bound we can show only applies when (22) holds.

The scenario that these guarantees do not account for is when at some point in the middle of GEOSGD, we arrive upon a subspace $V^{(0)}$ for which $d_P(V^{(0)}, V^*) \ll \underline{\epsilon}/2$, in which case running REALIGNPOLYNOMIAL with $V^{(0)}$ gives coefficients $\mathbf{c}$ for which (22) fails to hold. Intuitively, this should be fine because $d_P(V^{(0)}, V^*) < \underline{\epsilon}$, so GEOSGD has already produced a good enough estimate for the true subspace and we could just terminate. Unfortunately, it is not immediately obvious how to tell when this has happened and terminate accordingly.

Instead, we argue that local smoothness for SUBSPACEDESCENT (Lemma 53), implies that in this case, running SUBSPACEDESCENT initialized to $V^{(0)}$ will produce a subspace $V^{(T)}$ whose error is still good enough:

**Lemma 71** *Suppose all of the assumptions of Theorem 52 hold except for (22). Then we still have that $d_P(V^{(T)}, V^*) \leq \|\mathbf{c} - \mathbf{c}^*\|_2$ with probability at least $1 - \delta$.*

**Proof** Suppose the event of Lemma 53 occurs. We have that

$$\begin{aligned} d_P(V^{(T)}, V^*) &\leq d_P(V^{(0)}, V^*) + d_P(V^{(0)}, V^{(T)}) \\ &\leq d_P(V^{(0)}, V^*) \cdot \left(1 + \eta_{\mathsf{vec}} \cdot O(dr^3 \ln(T/\delta))^{(d+2)/2} \cdot O(\sqrt{n})\right) \\ &\leq \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2 \cdot \left(1 + \eta_{\mathsf{vec}} \cdot O(dr^3 \ln(T/\delta))^{(d+2)/2} \cdot O(\sqrt{n})\right) \\ &= \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\| \cdot \left(1 + O\left(\frac{\alpha_{\mathsf{ndg}}}{T\sqrt{n}}\right)\right) \\ &< \|\mathbf{c} - \mathbf{c}^*\|_2, \end{aligned}$$

where the first step follows by triangle inequality for Procrustes distance (Fact 10), the second by the assumption that the event of Lemma 53 holds, the third by the assumption that (22) does not hold, and the fourth by the definition of $\eta_{\mathsf{vec}}$ in (23). ■

We can now complete the proof of Theorem 34.

**Proof** [Proof of Theorem 34] Let $\mathbf{c}^{(t)}$ and $V^{(t)}$ be the iterates of GEOSGD. Suppose for $0 \leq t < T$ we had $d_P(V^{(t)}, V^*) \leq c_{12} \cdot \alpha_{\mathsf{ndg}} \cdot O(dr^3)^{-d-2}$. By Theorem 38, we have that

$$\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|_2 < 2 \cdot \epsilon/2 \vee d_P(V^{(t)}, V^*).$$

If $\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|_2 < \epsilon$, then by Lemma 71, $d_P(V^{(t+1)}, V^*) < \epsilon$. Otherwise, if $\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|_2 \leq 2d_P(V^{(t)}, V^*)$, then (22) in Theorem 52 holds and we get that

$$d_P(V^{(t+1)}, V^*) \leq (1 - \alpha) \cdot d_P(V^{(t)}, V^*),$$

where

$$\alpha \triangleq \frac{\alpha_{\mathsf{ndg}}}{n} \cdot \mathrm{poly}(\ln(1/\alpha_{\mathsf{ndg}}), r, d, \ln(1/\delta'))^{-d}$$

for $\delta' = \delta/(2T+1)$ as defined in GEOSGD.

In either case, $d_P(V^{(t+1)}, V^*) \leq c_{12} \cdot \alpha_{\mathsf{ndg}} \cdot O(dr^3)^{-d-2}$. And furthermore, if we unroll this recurrence, we conclude that

$$d_P(V^{(T)}, V^*) \leq \epsilon \vee (1 - \alpha)^T \cdot d_P(V^{(0)}, V^*).$$

So by taking $T = \alpha^{-1} \cdot \log(1/\epsilon)$, we get that $d_P(V^{(T)}, V^*) \leq \epsilon$ as desired. This corresponds to the choice of $T$ in (7). Lastly, we get that $\|\mathbf{c}^{(T)} - \mathbf{c}\|_2 \leq \epsilon$ by one last application of Theorem 38. ■

**Proof** [Proof of Theorem 35] This follows from the runtime and sample complexity guarantees of Theorems 38 and 52. ■

## Appendix G. Martingale Concentration Inequalities

In this section we prove the two martingale concentration inequalities from Section A.2.2 that are needed for the analysis of the boosting phase of our algorithm.

### G.1. Proof of Lemma 13

We first prove the following more general statement.

**Lemma 72** *Let $\sigma > 0$ and $0 < \alpha \leq 2$ be constants, and let $\mathcal{E}_i$ be the event that $\mathbb{E}[|Z_i|^q | \xi_1, ..., \xi_{i-1}] \leq \sigma^q \cdot q^{q/\alpha}$ for all $q \geq 1$.*

*If $\Pr[\mathcal{E}_i | \xi_1, ..., \xi_{i-1}] \geq 1 - \beta$ for each $i \in [T]$, then for any $t > 0$,*

$$\Pr\left[\max_{\ell \in [T]} \left|\sum_{i=1}^{\ell} Z_i\right| \geq t \cdot \sqrt{T} \cdot \sigma\right] \leq O\left(1 + t^2 (1/\alpha)^{O(1/\alpha)}\right) \cdot \exp\left(-\left(t^2/32\right)^{\frac{\alpha}{2+\alpha}}\right) + T \cdot \beta. \quad (34)$$

*In particular, there is an absolute constant $c_1 > 0$ such that for any $\delta > 0$,*

$$\Pr\left[\max_{\ell \in [T]} \left|\sum_{i=1}^{\ell} Z_i\right| \geq (\log(1/\delta)/\alpha)^{2c_1/\alpha} \cdot \sqrt{T} \cdot \sigma\right] \leq \delta + T \cdot \beta.$$

We first show that this implies Lemma 13.

**Proof** [Proof of Lemma 13] This is an immediate consequence of Lemma 72 together with Fact 6, which implies the requisite moment bounds for Lemma 72 for $\alpha = d/2$. ■

To show Lemma 72, we require the following theorem on the concentration of martingales with sub-Weibull differences, which is a consequence of the main result of Li (2018a).

**Theorem 73 (Li (2018a))** *Let $\sigma > 0$ and $0 < \alpha \leq 2$ be constants. Suppose that for every $i \in [T]$, we have that with probability one, $\mathbb{E}[|Z_i|^q | \xi_1, ..., \xi_{i-1}] \leq \sigma^q \cdot q^{q/\alpha}$ holds for all $q \geq 1$. Then for any $z > 0$,*

$$\Pr\left[\max_{\ell \in [T]} \left|\sum_{i=1}^{\ell} Z_i\right| \geq t \cdot \sqrt{T} \cdot \sigma\right] \leq O\left(1 + t^2 (1/\alpha)^{O(1/\alpha)}\right) \cdot \exp\left(-\left(t^2/32\right)^{\frac{\alpha}{2+\alpha}}\right) \quad (35)$$

We use a standard trick, see e.g. Lemma 3.1 of Vu (2002), to relax the assumption that the differences are sub-Weibull almost surely to the assumption that they are sub-Weibull with high probability. It will also be more convenient for us to state the inequality in terms of moment bounds rather than Orlicz norm bounds.

**Proof** [Proof of Lemma 72] Given a realization $\xi$ of the random variables $(\xi_1, ..., \xi_T)$, let $i_\xi$ be the first index $i$, if any, for which $\mathcal{E}_i$ does not hold. Define $B_i \triangleq \{\xi : i_\xi = i\}$ and note that these sets are disjoint for different $i$. Let $Y'(\xi)$ be the function which agrees with $Y(\xi)$ for $\xi \in (\cup B_i)^c$ and which is equal to $\mathbb{E}_{B_i}[Y]$ for $\xi \in B_i$. $Y'$ and $Y$ have the same mean, so the lemma follows by union bounding over the events $\cup B_i$ together with the probability that the martingale $Y'$ fails to concentrate. For the former probabilities, by definition $\Pr[B_i] \leq \beta$. And for the latter, because the martingale differences for $Y'$ satisfy the assumptions of Theorem 73, $Y'$ fails to concentrate with probability at most the right-hand side of (35). This yields (34). ■

### G.2. Proof of Lemma 14

To show Lemma 14, we require the following theorem due to Bentkus (2003), which controls the tails of martingales whose differences are only bounded on one side.

**Theorem 74 (Bentkus (2003))** *Let $\{c_i\}_{i\in[T]}$ and $\{s_i\}_{i\in[T]}$ be collections of positive constants for which $Z_i \le c_i$ and $\mathbb{E}[Z_i^2|\xi_1,...,\xi_{i-1}] \le s_i^2$ with probability one for every $i \in [T]$. Let $\sigma_i = c_i \vee s_i$, and define $\sigma^2 = \sum_i \sigma_i^2$. Then*

$$\Pr\left[\sum_{i=1}^{T} Z_i \ge t\cdot\sigma\right] \le \exp(-t^2/2).$$

**Proof** [Proof of Lemma 14] The proof is identical to that of Lemma 72, except instead of applying Theorem 73 to the auxiliary martingale, we apply Theorem 74 to get that for any $t > 0$,

$$\Pr\left[\sum_{i=1}^{T} Z_i \ge t\cdot\sigma\right] \le \exp(-t^2/2) + T\cdot\beta.$$

The lemma follows by taking $t = \sqrt{2}\log(1/\delta)$. ■

## Appendix H. Deferred Proofs from Section D

### H.1. Proof of Lemma 42

**Proof**

$$\begin{aligned}
&\mathbb{E}\left[(F_x(\Theta) - F_x(\Theta^*))^4\right] \\
&\quad\le \sum_{\ell_1,\ldots,\ell_4\in[d+1]} \frac{1}{\prod_{\nu=1}^4 \ell_\nu!}\mathbb{E}\left[\prod_{\nu=1}^4 \left\langle \nabla^{[\ell_\nu]}F_x(\Theta), (\Theta^*-\Theta)^{\otimes\ell_\nu}\right\rangle\right] \\
&\quad\le \sum_{\ell_1,\ldots,\ell_4\in[d+1]} \frac{1}{\prod_{\nu=1}^4 \ell_\nu!}\cdot 16\cdot(8dr^2)^{2(d+1)}\cdot\|V-V^*\|_F^{\sum_\nu \ell_\nu}\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^4 \\
&\quad\le 16\cdot(8dr^2)^{2(d+1)}\left(\sum_{\ell=1}^{d+1}\frac{1}{\ell!}\cdot\|V-V^*\|_F^{\ell}\right)^4\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^4 \\
&\quad\le 16\cdot(8dr^2)^{2(d+1)}\cdot(e\cdot(4r)^{d/2}\|V-V^*\|_F)^4\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^4 \\
&\quad\le (2e)^4\cdot(32dr^3)^{2(d+1)}\cdot\left(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2\right)^4,
\end{aligned}$$

where the second step follows by Lemma 37, the fourth by the fact that $\|V-V^*\|_F \le 2\sqrt{r}$ and the fact that $\sum_{\ell=1}^{d+1}\frac{1}{\ell!}\cdot x^\ell \le e\cdot(4r)^{d/2}\cdot x$ for $x\in[0,2\sqrt{r}]$. ■

### H.2. Proof of Lemma 47

**Proof** We have that

$$\begin{aligned}\frac{1}{2\eta_{\mathsf{coef}}}\left|\langle \Delta''_{\mathsf{coef}}, \mathbf{c}-\mathbf{c}^*\rangle\right| &= \left|\mathbb{E}\left[\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle \nabla^{[\ell]}F_x(\Theta), (\Theta^*-\Theta)^{\otimes \ell}\right\rangle \cdot \delta(V^\top x)\right]\right| \\ &\le \mathbb{E}\left[\left(\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle \nabla^{[\ell]}F_x(\Theta), (\Theta^*-\Theta)^{\otimes \ell}\right\rangle\right)^2\right]^{1/2} \cdot \mathbb{E}\left[\delta(V^\top x)^2\right]^{1/2} \\ &\le O(dr^3)^{(d+1)/2}\cdot \|V-V^*\|_F\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2 \\ &= O(dr^3)^{(d+1)/2}\cdot d_P(V,V^*)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\cdot(d_P(V,V^*)+\|\mathbf{c}-\mathbf{c}^*\|_2),\end{aligned}$$

where the second step follows by Cauchy-Schwarz, the third step follows by Lemma 83, and the last step follows by the assumption that $\|V-V^*\|_F = d_P(V,V^*)$. ■

### H.3. Proof of Proposition 48

**Proof** Note that

$$\begin{aligned}\mathbb{E}_g\left[\left(x^\top \Pi_V(V^*-V)\nabla p(V^\top x)\right)^2\right]^{1/2} &\le \|\mathbf{I}-V^\top V^*\|_2\cdot\mathbb{E}_g\left[\|g\|_2^2\cdot\|\nabla p(g)\|_2^2\right]^{1/2} \\ &\le d_P(V,V^*)^2\cdot O(r^{3/2}d),\end{aligned}$$

where the second step follows by the second part of Lemma 26, Lemma 20, and the fact that

$$\mathrm{Var}[p]^{1/2} \le \|\mathbf{c}-\mathbf{c}^*\|_2 + \mathrm{Var}[p^*]^{1/2} \le O(r)$$

because $\|\mathbf{c}-\mathbf{c}^*\|_2 \le 1$ by assumption and because of Corollary 1. ■

### H.4. Proof of Lemma 50

We will split up $\frac{1}{B}\sum_{i=0}^{B-1} Y^{x^i}$ according to the decomposition (20). That is, define

$$\begin{aligned}\text{Ⓐ}^x &\triangleq \left(\delta(V^\top x)\right)^2 \\ \text{Ⓑ}^x &\triangleq \delta(V^\top x)\cdot\left(x^\top \Pi_V(V^*-V)\nabla\right) \\ \text{Ⓒ}^x &\triangleq \delta(V^\top x)\cdot\left(x^\top \Pi_V^{\perp} V^*\nabla\right)\end{aligned}$$

so that for any $x$,

$$\frac{1}{2\eta_{\mathsf{coef}}}Y^x = \text{Ⓐ}^x + \text{Ⓑ}^x + \text{Ⓒ}^x. \tag{36}$$

We will show concentration for these three random variables separately.

**Lemma 75** *For any $\delta > 0$, if $B = \Omega(\log(1/\delta)^2 \cdot 9^d)$, then*

$$\frac{1}{B}\sum_{i=0}^{B-1} \text{Ⓐ}^{x^i} \geq \frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2^2$$

*with probability at least $1 - \delta$.*

**Lemma 76** *For any $\delta > 0$, if $B = \Omega(\log(1/\delta))^{2d}$, then*

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} \text{Ⓑ}^{x^i}\right| \leq O(r^{3/2}d) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot d_P(V, V^*)^2$$

*with probability at least $1 - \delta$.*

**Lemma 77** *For any $\delta > 0$ and $\gamma > 0$, if $B = \Omega(\log(1/\delta))^{2d} \cdot \gamma^{-2}$, then*

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} \text{Ⓒ}^{x^i}\right| \leq \gamma \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2$$

*with probability at least $1 - \delta$.*

We prove these in the subsequent Appendices H.4.1, H.4.2, and H.4.3. Note that Lemma 50 immediately follows from these lemmas.

**Proof** [Proof of Lemma 50] By a union bound over the failure probabilities of Lemmas 75, 76, and 77, we see by triangle inequality and (36) that

$$\frac{1}{B}\sum_{i=0}^{B-1} Y^{x^i} \geq 2\eta_{\mathsf{coef}} \cdot \left(\frac{1}{2}\|\mathbf{c} - \mathbf{c}^*\|_2^2 - O(r^{3/2}d) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 \cdot d_P(V, V^*)^2 - \gamma \cdot d_P(V, V^*) \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right)$$

with probability at least $1 - 3\delta$, provided $B = \Omega(\log(1/\delta))^d \cdot \gamma^{-2}$. The result follows by replacing $3\delta$ with $\delta$ and absorbing constants. ■

### H.4.1. PROOF OF LEMMA 75

**Proof** Observe that $\frac{1}{B}\sum_{i=0}^{B-1}\left(\mathbb{E}_x[\text{Ⓐ}^x] - \text{Ⓐ}^{x^i}\right)$ is an average of $B$ iid copies of a mean-zero random variable satisfying one-sided bounds, so we wish to apply Lemma 12.

To do so, we just need to bound the variances of the summands.

**Lemma 78** *$Var_x[\text{Ⓐ}^x] \leq 9^d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^4$.*

**Proof** Clearly $\text{Var}[\text{Ⓐ}^x] \leq \mathbb{E}[(\text{Ⓐ}^x)^2]$, so it suffices to bound the latter. By Fact 6 applied to the degree-$d$ polynomial $\delta$,

$$\mathbb{E}[(\text{Ⓐ}^x)^2] = \mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}[\delta(g)^4] \leq 9^d \cdot \mathbb{E}[\delta(g)^2]^2 = 9^d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^4 \tag{37}$$

as claimed. ■

We can now complete the proof of Lemma 75.
By Lemma 12, Observation 2, and Lemma 78,

$$\frac{1}{B}\sum_{i=0}^{B-1} \text{Ⓐ}^{x^i} \geq \|\mathbf{c}-\mathbf{c}^*\|_2^2 - \frac{1}{\sqrt{B}} \cdot \sqrt{2}\log(1/\delta)\cdot 3^d \cdot \|\mathbf{c}-\mathbf{c}^*\|^2$$

with probability at least $1-\delta$. The lemma follows by taking $B = \Omega(\log(1/\delta)^2$. ■

### H.4.2. PROOF OF LEMMA 76

**Proof** Note that $\text{Ⓑ}^x$ is a polynomial of degree $2d$ in $x$, so by Lemma 11, we just need to upper bound its variance.

**Lemma 79** *$Var_x[\text{Ⓑ}^x] \leq 9^d \cdot O(r^{3/2}d) \cdot \|\mathbf{c}-\mathbf{c}^*\|_2^2 \cdot d_P(V,V^*)^4$.*

**Proof** We will upper bound $\mathbb{E}_x[(\text{Ⓑ}^x)^2]$ via

$$\begin{aligned}
\mathbb{E}\left[\text{Ⓑ}^2\right] &\leq \mathbb{E}\left[\delta(V^\top x)^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \Pi_V(V^*-V)\nabla\right)^4\right]^{1/2} \\
&\leq \mathbb{E}[\text{Ⓐ}^2]\cdot 3^d \cdot \mathbb{E}\left[\left(x^\top \Pi_V(V^*-V)\nabla\right)^2\right] \\
&\leq 9^d \cdot O(r^3d^2)\cdot \|\mathbf{c}-\mathbf{c}^*\|_2^2 \cdot d_P(V,V^*)^4,
\end{aligned}$$

where in the first step we used Cauchy-Schwarz, in the second we used Proposition 48, and in the third we used (37). ■

We can now complete the proof of Lemma 75.
By Lemma 11, Lemma 49, and Lemma 79,

$$\left|\frac{1}{B}\sum_{i=0}^{B-1} \text{Ⓑ}^{x^i}\right| \leq O(r^{3/2}d)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2 \cdot d_P(V,V^*)^2 \cdot \left(1+\frac{1}{\sqrt{B}}\cdot O(\log(1/\delta))^d \cdot 3^d\right),$$

with probability at least $1-\delta$. The lemma follows by taking $B = \Omega(\log(1/\delta))^{2d}\cdot\Omega(9^d)$. ■

### H.4.3. PROOF OF LEMMA 77

**Proof** Note that $\text{Ⓒ}^x$ is a polynomial of degree $2d$ in $x$, so by Lemma 11, we just need to upper bound its variance.

**Lemma 80** *For any $\Theta$, $\mathbb{E}_x[(\text{Ⓒ}^{\Theta,x})^2] \leq d_P(V,V^*)^2 \cdot \|\mathbf{c}-\mathbf{c}^*\|_2^2 \cdot \exp(O(d))$.*

**Proof** This is shown in Lemma 91 below. The proof involves calculations which are more pertinent to the behavior of SUBSPACEDESCENT, so we defer the details to there. ■

We can now complete the proof of Lemma 77. By Lemma 11, Observation 1, and Lemma 80,

$$\left|\frac{1}{B}\sum_{i=1}^{B-1} \text{Ⓒ}^{x^i}\right| \leq \frac{1}{\sqrt{B}}\cdot O(\log(1/\delta))^d \cdot d_P(V,V^*)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2 \cdot \exp(O(d))$$

with probability at least $1-\delta$. The lemma follows by taking $B = \Omega(\log(1/\delta))^{2d}\cdot\gamma^{-2}$. ■

### H.5. Proof of Lemma 51

**Proof** Note that $E^x$ is a polynomial of degree $2d$ in $x$, so by Lemma 11, we just need to upper bound its variance.

To do so, we will need the following helper lemma, which like Lemma 42 is a straightforward consequence of Lemma 37.

**Lemma 81** $\mathbb{E}[(\mathfrak{R}^{\Theta,x})^4]^{1/2} \le O(dr^3)^{d+1} \cdot \|V - V^*\|_F^2 \cdot (\|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^* - V\|_F)^2$

**Proof** We have that

$$\begin{aligned}
\mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^4\right]^{1/2} &= \left(\sum_{\ell_1,\dots,\ell_4>1} \frac{1}{\prod_{\nu=1}^4 \ell_\nu!} \mathbb{E}\left[\prod_{\nu=1}^4 \left\langle \nabla^{[\ell_\nu]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell_\nu} \right\rangle\right]\right)^{1/2} \\
&\le \left(\sum_{\ell_1,\dots,\ell_4>1} \frac{1}{\prod_{\nu=1}^4 \ell_\nu!} 16 \cdot (8dr^2)^{2(d+1)} \cdot \|V^* - V\|_F^{\sum_\nu \ell_\nu} \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^4\right)^{1/2} \\
&= 4(8dr^2)^{d+1} \left(\sum_{\ell=2}^{d+1} \frac{1}{\ell!} \|V^* - V\|_F^\ell\right)^2 \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^2 \\
&\le 4(8dr^2)^{d+1} \cdot \left(e^2 \cdot (4r)^{d-1} \|V^* - V\|_F^4\right) \cdot \left(1 + \frac{\|\mathbf{c} - \mathbf{c}^*\|_2}{\|V^* - V\|_F}\right)^2 \\
&= 4e^2 \cdot (32dr^3)^{d+1} \cdot \|V - V^*\|_F^2 \cdot (\|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^* - V\|_F)^2,
\end{aligned}$$

where the second step follows by Lemma 37, and the fourth step follows by the fact that we always have $\|V - V^*\|_F \le 2\sqrt{r}$, and $\sum_{\ell=2}^{d+1} \frac{1}{\ell!} x^\ell < e \cdot (4r)^{(d-1)/2} \cdot x^2$ for $x \in [0, 2\sqrt{r}]$. ■

We can now show the variance bound.

**Lemma 82** $\mathbb{E}_x[(E^x)^2] \le \eta_{\mathsf{coef}}^2 \cdot O(dr^3)^{d+1} \cdot d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|)^2$.

**Proof** By Cauchy-Schwarz,

$$\begin{aligned}
\frac{1}{4\eta_{\mathsf{coef}}^2} \mathbb{E}\left[(E^x)^2\right] &\le \mathbb{E}[(\mathfrak{R}^{\Theta,x})^4]^{1/2} \cdot \mathbb{E}[\delta(g)^4]^{1/2} \\
&\le 4e^2 \cdot (32dr^3)^{d+1} \cdot \|V - V^*\|_F^2 \cdot (\|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^* - V\|_F)^2 \cdot 3^d \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \\
&= O(dr^3)^{d+1} \cdot d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|)^2
\end{aligned}$$

where the second step follows by Lemma 81 and the third step follows by the assumption that $\|V - V^*\|_F = d_P(V, V^*)$. ■

Finally, by Lemma 11, Lemma 47, and Lemma 82,

$$\left|\frac{1}{B} \sum_{i=0}^{B-1} E^{x^i}\right| \le O(dr^3)^{(d+1)/2} \, d_P(V, V^*) \, \|\mathbf{c} - \mathbf{c}^*\|_2 \, (d_P(V, V^*) + \|\mathbf{c} - \mathbf{c}^*\|_2) \cdot \left(1 + \frac{1}{\sqrt{B}} \cdot O(\log(1/\delta))^d\right).$$

The lemma follows by taking $B = O(\log(1/\delta))^{2d}$. ■

## Appendix I. Deferred Proofs from Section E

### I.1. Proof of Lemma 37

**Proof** We begin by explicitly computing the higher-order terms in the Taylor-expansion of $F_x(\Theta) - F_x(\Theta^*)$. For any $\ell \in [d+1]$, recalling the notation of (3) and (4),

$$\begin{aligned}
&\left\langle \nabla^{[\ell]} F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell} \right\rangle \\
&\quad = \sum_{\mathbf{i} \in [n]^\ell, \mathbf{j} \in [r]^\ell} \prod_{a=1}^{\ell} (V^*_{i_a, j_a} - V_{i_a, j_a}) \cdot \mathrm{D}_{\mathbf{i},\mathbf{j}} F_x(\Theta) + \sum_{I, \mathbf{i} \in [n]^\ell, \mathbf{j} \in [r]^{\ell-1}} \prod_{a=1}^{\ell-1} (V^*_{i_a, j_a} - V_{i_a, j_a}) \cdot (c^*_I - c_I) \cdot \mathrm{D}_{\mathbf{i},\mathbf{j}} F_x(\Theta) \\
&\quad = \sum_{\mathbf{i} \in [n]^\ell, \mathbf{j} \in [r]^\ell} \prod_{a=1}^{\ell} (V^*_{i_a, j_a} - V_{i_a, j_a}) \cdot x_{i_a} \cdot \mathrm{D}_{\mathbf{j}}\, p(V^\top x) + \sum_{\mathbf{i} \in [n]^\ell, \mathbf{j} \in [r]^{\ell-1}} \prod_{a=1}^{\ell-1} (V^*_{i_a, j_a} - V_{i_a, j_a}) \cdot x_{i_a} \cdot \mathrm{D}_{\mathbf{j}}\, \delta(V^\top x) \\
&\quad = \sum_{\mathbf{j} \in [r]^\ell} \prod_{a=1}^{\ell} \langle (V^* - V)_{j_a}, x \rangle \cdot \mathrm{D}_{\mathbf{j}}\, p(V^\top x) + \sum_{\mathbf{j} \in [r]^{\ell-1}} \prod_{a=1}^{\ell-1} \langle (V^* - V)_{j_a}, x \rangle \cdot \mathrm{D}_{\mathbf{j}}\, \delta(V^\top x) \qquad (38)
\end{aligned}$$

From (38), we can rewrite the quantity in the expectation as

$$\sum_{\substack{\mathbf{b} \in \{0,1\}^m \\ \{\mathbf{j}^{(\nu)}\}_{\nu \in [m]}}} \prod_{\nu=1}^{m} \left( \prod_{a=1}^{\ell_\nu - b_\nu} \left\langle (V^* - V)_{j_a^{(\nu)}}, x \right\rangle \right) \left( \mathbb{1}\left[b_\nu = 0\right] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, p(V^\top x) + \mathbb{1}\left[b_\nu = 1\right] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, \delta(V^\top x) \right).$$

We will bound the expected absolute values of each of these summands individually, so henceforth fix an arbitrary $\mathbf{b}$, $\{\mathbf{j}^{(\nu)}\}$. For convenience, define $C_\nu \triangleq \left( \mathbb{1}\left[b_\nu = 0\right] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, p(V^\top x) + \mathbb{1}\left[b_\nu = 1\right] \cdot \mathrm{D}_{\mathbf{j}^{(\nu)}}\, \delta(V^\top x) \right)$.

By AM-GM, we have that

$$\begin{aligned}
&\mathbb{E}\left[ \left( \prod_{\nu=1}^{m} |C_\nu| \right) \cdot \left( \prod_{\nu=1}^{m} \prod_{a=1}^{\ell_\nu - b_\nu} \left| \left\langle (V^* - V)_{j_a^{(\nu)}}, x \right\rangle \right| \right) \right] \\
&\quad \le \mathbb{E}\left[ \left( \prod_{\nu=1}^{m} C_\nu \right) \cdot \left( \prod_{\nu=1}^{m} \frac{1}{\ell_\nu - b_\nu} \sum_{a=1}^{\ell_\nu - b_\nu} \left| \left\langle (V^* - V)_{j_a^{(\nu)}}, x \right\rangle \right|^{\ell_\nu - b_\nu} \right) \right] \\
&\quad \le \mathbb{E}\left[ \prod_{\nu=1}^{m} \frac{C_\nu^2}{(\ell_\nu - b_\nu)^2} \right]^{1/2} \cdot \mathbb{E}\left[ \left( \sum_{\mathbf{a} \in \prod_\nu [\ell_\nu - b_\nu]} \prod_{\nu=1}^{m} \left| \left\langle (V^* - V)_{j^{(\nu)}_{a_\nu}}, x \right\rangle \right|^{\ell_\nu - b_\nu} \right)^2 \right]^{1/2} \qquad (39)
\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz.

Defining $w_{\mathbf{b}} = \sum_\nu \ell_\nu - b_\nu$, we may write the second factor in (39) as

$$\begin{aligned}
&\mathbb{E}\left[ \sum_{\mathbf{a}^1, \mathbf{a}^2} \prod_{\nu=1}^{m} \left| \left\langle (V^* - V)_{j^{(\nu)}_{a^1_\nu}}, x \right\rangle \right|^{\ell_\nu - b_\nu} \cdot \prod_{\nu=1}^{m} \left| \left\langle (V^* - V)_{j^{(\nu)}_{a^2_\nu}}, x \right\rangle \right|^{\ell_\nu - b_\nu} \right]^{1/2} \\
&\qquad \le (2w_{\mathbf{b}})^{w_{\mathbf{b}}/2} \|V^* - V\|_F^{w_{\mathbf{b}}} \cdot \prod_\nu (\ell_\nu - b_\nu) \le (2m)^{m/2} \|V^* - V\|_F^{w_{\mathbf{b}}} \cdot \prod_\nu (\ell_\nu - b_\nu),
\end{aligned}$$

where we used the standard bound for moments of a univariate Gaussian, the fact that there are $\prod_\nu(\ell_\nu - b_\nu)^2$ pairs of summands $\mathbf{a}^1, \mathbf{a}^2$, and the fact that any column of $V^* - V$ has $L_2$ norm at most $\|V^* - V\|_F$.

By Holder's, we may upper bound the first factor in (39) by $\prod_{\nu=1}^m \frac{1}{\ell_\nu - b_\nu}\mathbb{E}\left[C_\nu^{2m}\right]^{1/2m}$.
By Corollary 17,

$$\mathbb{E}\left[\left(\mathrm{D}_{\mathbf{j}^{(\nu)}}\,\delta(V^\top x)\right)^{2m}\right]^{1/2m} \leq (2m)^{d/2}d^{(\ell_\nu-1)/2}\cdot \mathrm{Var}[\delta]^{1/2} \leq (2m)^{d/2}d^{\ell_\nu/2}\cdot\|\mathbf{c}-\mathbf{c}_*\|_2.$$

$$\mathbb{E}\left[\left(\mathrm{D}_{\mathbf{j}^{(\nu)}}\,p(V^\top x)\right)^{2m}\right]^{1/2m} \leq (2m)^{d/2}d^{\ell_\nu/2}\cdot \mathrm{Var}[p]^{1/2} \leq 2\cdot(2m)^{d/2}d^{\ell_\nu/2},$$

where in the last step we used that $\mathrm{Var}[p]^{1/2} \leq \mathrm{Var}[p*]^{1/2} + \mathrm{Var}[\delta]^{1/2} \leq 2$. So the first factor in (39) is at most

$$\left(\prod_{\nu=1}^m \frac{1}{\ell_\nu - b_\nu}\right)\cdot 2^m\cdot(2m)^{md/2}d^{\sum_\nu \ell_\nu/2}\|\mathbf{c}-\mathbf{c}_*\|_2^{\sum_\nu b_\nu},$$

so (39) is at most $2^m\cdot(2m)^{m(d+1)/2}d^{m(d+1)/2}\cdot\|V^*-V\|_F^{w_\mathbf{b}}\cdot\|\mathbf{c}-\mathbf{c}_*\|_2^{\sum_\nu b_\nu}$. The proof follows by noting that

$$\sum_{\mathbf{b}}\|V^*-V\|_F^{w_\mathbf{b}}\cdot\|\mathbf{c}-\mathbf{c}_*\|_2^{\sum_\nu b_\nu} = \|V^*-V\|_F^{\sum_\nu \ell_\nu}\cdot\sum_{\mathbf{b}}\left(\frac{\|\mathbf{c}-\mathbf{c}_*\|_2}{\|V^*-V\|_F}\right)^{\sum_\nu b_\nu}$$

and summing (39) over all choices of $\mathbf{b}$ and all $\prod_\nu r^{\ell_\nu} \leq r^{m(d+1)}$ choices of $\{\mathbf{j}^{(\nu)}\}$. ■

### I.2. Proof of Lemma 58

**Proof** Let

$$\alpha_q \triangleq 3\sqrt{r}\cdot\eta_{\mathsf{vec}}\cdot O(\sqrt{n})\cdot O(dr^3)^{(d+2)/2}.$$

($q = 1$). Analogous to the derivation of (29), we have that

$$\begin{aligned}
\mathbb{E}\left[\|V^{(t)}-V^*\|_F\right] &\leq \mathbb{E}\left[\|V^{(t-1)}-V^*\|_F\right] + \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F\right]\\
&\leq \mathbb{E}\left[\|V^{(t-1)}-V^*\|_F\right] + 3\sqrt{r}\cdot\eta_{\mathsf{vec}}\mathbb{E}\left[(\sigma^{\Theta^{(t-1)},x^{t-1}})^2\right]^{1/2}\\
&\leq (1+\alpha_1)\mathbb{E}\left[\|V^{(t-1)}-V^*\|_F\right] + \alpha_1\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\\
&\leq (1+\alpha_1)^t\cdot\|V^{(0)}-V^*\|_F + \left((1+\alpha_1)^t-1\right)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2\\
&= (1+\alpha_1)^t\cdot d_P(V^{(0)},V^*) + \left((1+\alpha_1)^t-1\right)\cdot\|\mathbf{c}-\mathbf{c}^*\|_2
\end{aligned}$$

where in the second step we used Cauchy-Schwarz and (26), in the third step we used Lemma 55, in the fourth step we unrolled the recurrence, and in the last step we used the assumption that $\|V^{(0)}-V^*\|_F = d_P(V^{(0)},V^*)$. The proof follows by taking $\eta_{\mathsf{vec}}$ small enough that

$$(1+\alpha_1)^t + \left((1+\alpha_1)^t-1\right)\cdot\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{d_P(V^{(0)},V^*)} \leq 1.1.$$

$\eta_{\mathsf{vec}}$ given by (23) will easily satisfy this.

(Larger $q$) We have that

$$\begin{aligned}
\mathbb{E}\left[\|V^{(t)}-V^*\|_F^q\right]^{1/q} &\le \mathbb{E}\left[\|V^{(t-1)}-V^*\|_F^q\right]^{1/q} + \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F^q\right]^{1/q} \\
&\le \mathbb{E}\left[\|V^{(t-1)}-V^*\|_F^q\right]^{1/q} + \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)},x^{t-1}}\|_F^{2q}\right]^{1/2q} \\
&\le \mathbb{E}\left[\|V^{(t-1)}-V^*\|_F^q\right]^{1/q} + \alpha_q\cdot\left(\mathbb{E}\left[\|V^{(t-1)}-V^*\|_F\right] + \|\mathbf{c}-\mathbf{c}^*\|_2\right) \\
&\le \mathbb{E}\left[\|V^{(t-1)}-V^*\|_F^2\right]^{1/q} + 1.1\alpha_q\cdot\left(d_P(V^{(0)},V^*) + \cdot\|\mathbf{c}-\mathbf{c}^*\|_2\right) \\
&\le d_P(V^{(0)},V^*) + 1.1t\cdot\alpha_q\cdot\left(d_P(V^{(0)},V^*) + \cdot\|\mathbf{c}-\mathbf{c}^*\|_2\right)
\end{aligned}$$

where the first step follows by triangle inequality, the second by monotonicity of $L_p$ norms, the third by Lemma 55, the fourth by Lemma 58, and the fifth by unrolling the recurrence and using the assumption that assumption that $\|V^{(0)}-V^*\|_F = d_P(V^{(0)},V^*)$.

The proof follows by taking $\eta_{\mathsf{vec}}$ small enough that $1.1T\cdot\alpha_q\cdot\|\mathbf{c}-\mathbf{c}^*\|_2 \le O(\alpha_q\cdot T)$ is a negligible constant, which is certainly the case if $\eta_{\mathsf{vec}}$ satisfies (23) (with hidden constant factors there depending on $q$). ■

### I.3. Proof of Lemma 61

We first prove the following basic consequence of Lemma 37:

**Lemma 83**

$$\mathbb{E}\left[\left(\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle\nabla^{[\ell]}F_x(\Theta),(\Theta^*-\Theta)^{\otimes\ell}\right\rangle\right)^2\right]^{1/2} \le O(dr^3)^{(d+1)/2}\cdot\|V-V^*\|_F\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2) \quad (40)$$

**Proof** The left-hand side of (40) can be rewritten as

$$\begin{aligned}
&\left(\sum_{\ell_1,\ell_2>1}\frac{1}{\ell_1!\,\ell_2!}\mathbb{E}\left[\prod_{\nu=1}^{2}\left\langle\nabla^{[\ell_\nu]}F_x(\Theta),(\Theta^*-\Theta)^{\otimes\ell_\nu}\right\rangle\right]\right)^{1/2}\cdot\mathbb{E}\left[(x^\top\cdot\Pi_V^\perp V^*\cdot\Delta)^2\right]^{1/2} \\
&\quad\le\left(\sum_{\ell_1,\ell_2>1}\frac{1}{\ell_1!\,\ell_2!}4\cdot(4dr^2)^{d+1}\cdot\|V-V^*\|_F^{\ell_1+\ell_2}\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right)^2\right)^{1/2} \\
&\quad= 2(4dr^2)^{(d+1)/2}\sum_{\ell>1}\frac{1}{\ell!}\|V-V^*\|_F^\ell\cdot\left(1+\frac{\|\mathbf{c}-\mathbf{c}^*\|_2}{\|V^*-V\|_F}\right) \\
&\quad\le 2e(16dr^3)^{(d+1)/2}\cdot\|V-V^*\|_F\cdot(\|V-V^*\|_F+\|\mathbf{c}-\mathbf{c}^*\|_2)\,,
\end{aligned}$$

where the first step follows by Lemma 37, and the last step follows by the fact that $\|V-V^*\|_F\le 2\sqrt{r}$ and the fact that $\sum_{\ell=2}^{d+1}\frac{1}{\ell!}x^\ell < e\cdot(4r)^{(d-1)/2}\cdot x^2$ for $x\in[0,2\sqrt{r}]$. ■

**Proof** [Proof of Lemma 61] We have that

$$\begin{aligned}
&\frac{1}{2\eta_{\mathsf{vec}}}\left|\langle \tilde{\Delta}''_V, V - V^*\rangle\right| \\
&\quad= \left|\mathbb{E}\left[\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle \nabla^{[\ell]}F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell}\right\rangle \cdot x^\top \cdot \Pi_V^\perp V^* \cdot \Delta\right]\right| \\
&\quad\le \mathbb{E}\left[\left(\sum_{\ell=2}^{d+1}\frac{1}{\ell!}\left\langle \nabla^{[\ell]}F_x(\Theta), (\Theta^* - \Theta)^{\otimes \ell}\right\rangle\right)^2\right]^{1/2} \cdot \mathbb{E}\left[(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta)^2\right]^{1/2} \\
&\quad\le O(dr^3)^{(d+1)/2}\cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)\cdot \mathbb{E}[\text{(A')}]^{1/2} \\
&\quad\le O(dr^3)^{(d+1)/2}\cdot \|V - V^*\|_F \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)\cdot (2d_P(V, V^*)),
\end{aligned}$$

where the second step follows by Cauchy-Schwarz, the third by Lemma 40 and the definition of (A'), the fourth by the upper bound in (33). ■

### I.4. Proof of Lemma 62

By Holder's,

$$\begin{aligned}
&|\mathbb{E}\left[\langle \mathcal{E}, V - V^*\rangle\right]| \\
&\quad\le \mathbb{E}\left[|\cos(\sigma\eta_{\mathsf{vec}}) - 1|\right]\cdot \sup_{\widehat{\nabla}}\left|\langle V\cdot\widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle\right| + \mathbb{E}\left[|\sin(\sigma\eta_{\mathsf{vec}}) - \sigma\eta_{\mathsf{vec}}|\right]\cdot \sup_{\widehat{h},\widehat{\nabla}}\left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^*\rangle\right| \\
&\quad\le O(\eta_{\mathsf{vec}}^2)\cdot \mathbb{E}[\sigma^2]\cdot\left(\sup_{\widehat{\nabla}}\left|\langle V\cdot\widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle\right| + \sup_{\widehat{h},\widehat{\nabla}}\left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^*\rangle\right|\right), \qquad (41)
\end{aligned}$$

where in the second step we used that $|\cos(x) - 1| \le x^2/2$ and $|\sin(x) - x| \le x^2/\pi$ for all $x \ge 0$, and in the third step we invoked Lemmas 84 and 85 below.

**Lemma 84** *For any $\widehat{\nabla} \in \mathbb{S}^{r-1}$, $\left|\langle V\cdot\widehat{\nabla}\widehat{\nabla}^\top, V - V^*\rangle\right| \le \|V - V^*\|_F$.*

**Proof** We may write the quantity on the left-hand side as

$$\widehat{\nabla}^\top\cdot\left((V - V^*)^\top V\right)\cdot\widehat{\nabla} = \widehat{\nabla}^\top\left(\mathbf{I} - V^{*\top}V\right)\widehat{\nabla} \le \|\mathbf{I} - V^{*\top}V\|_2 \le \|V - V^*\|_F,$$

where the last step follows by the first part of Lemma 26. ■

**Lemma 85** *For any $\widehat{\nabla} \in \mathbb{S}^{r-1}$ and $\widehat{h} \in \mathbb{S}^{n-1}$ for which $\widehat{h}$ lies in the orthogonal complement of the column span of $V$, $\left|\left\langle \widehat{h}\widehat{\nabla}^\top, V - V^*\right\rangle\right| \le d_P(V, V^*)$.*

**Proof** Because $\Pi_V^{\perp}\widehat{h} = \widehat{h}$, The left-hand side can be rewritten as

$$\widehat{h}^{\top}(V - V^*)\widehat{\nabla} = \widehat{h}^{\top}\Pi_V^{\perp}(V - V^*)\widehat{\nabla},$$

it is upper-bounded by

$$\begin{aligned}\sigma_{\max}(\Pi^{\perp}(V - V^*)) &\leq \mathrm{Tr}((V - V^*)^{\top}(\mathbf{I} - VV^{\top})(V - V^*)^{1/2} \\ &= \mathrm{Tr}(\mathbf{I} - V^*V^{\top}VV^{*\top})^{1/2} \\ &= d_C(V, V^*) \leq d_P(V, V^*),\end{aligned}$$

where the last step follows by Lemma 24. ■

**Proof** [Proof of Lemma 62] We have

$$|\mathbb{E}\left[\langle \mathcal{E}, V - V^*\rangle\right]| \leq O(\eta_{\mathsf{vec}}^2) \cdot O(n) \cdot O(dr^3)^{d+2} \cdot \|V - V^*\|_F \cdot \left(\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}_*\|_2\right)^2,$$

by (41), Lemmas 84, 85, and 55. The lemma follows by taking $\eta_{\mathsf{vec}} \leq O(1/n)$. ■

### I.5. Proof of Lemma 65

**Proof** We will bound each $\mathbb{E}_{x^0,\ldots,x^{t-1}}[\mu_X(\Theta^{(t)}]$ individually. By Lemma 60, for any realization of $x^0, \ldots, x^{t-1}$ giving rise to iterate $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$, $\mu_X(\Theta^{(t)}) \geq (\alpha_{\mathsf{ndg}}/4) \cdot d_P(V^{(t)}, V^*)^2$. We have that

$$\begin{aligned}&\mathbb{E}\left[d_P(V^{(t)}, V^*)^2\right] \\ &\quad\geq \mathbb{E}\left[\left(d_P(V^{(t-1)}, V^*) - d_P(V^{(t)}, V^{(t-1)})\right)^2\right] \\ &\quad\geq \mathbb{E}\left[d_P(V^{(t-1}, V^*)^2\right] - 2\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right]^{1/2} \cdot \mathbb{E}\left[d_P(V^{(t)}, V^{(t-1)})^2\right]^{1/2} \\ &\quad\geq \mathbb{E}\left[d_P(V^{(t-1}, V^*)^2\right] - 2\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right]^{1/2} \cdot \mathbb{E}\left[\|\Delta_{\mathsf{vec}}^{\Theta^{(t-1)}, x^{t-1}}\|_F^2\right]^{1/2} \\ &\quad\geq \mathbb{E}\left[d_P(V^{(t-1}, V^*)^2\right] - 6\sqrt{r} \cdot \eta_{\mathsf{vec}}\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right]^{1/2} \cdot \mathbb{E}\left[(\sigma^{\Theta^{(t-1)}, x^{t-1}})^2\right]^{1/2} \quad (42)\end{aligned}$$

where the first step follows by triangle inequality (Fact 10), the second by Cauchy-Schwarz, the third by the definition of Procrustes distance, and the fourth by (26). By Lemma 55 and Lemma 58,

$$\begin{aligned}&6\sqrt{r} \cdot \eta_{\mathsf{vec}}\mathbb{E}\left[(\sigma^{\Theta^{(t-1)}, x^{t-1}})^2\right]^{1/2} \\ &\quad\leq 6\sqrt{r} \cdot \eta_{\mathsf{vec}} \cdot O(\sqrt{n}) \cdot (dr^3)^{(d+2)/2} \cdot \left(\mathbb{E}\left[\|V^{(t-1)} - V^*\|_F\right] + \|\mathbf{c} - \mathbf{c}^*\|_2\right) \\ &\quad\leq 6\sqrt{r} \cdot O(\sqrt{n}) \cdot (dr^3)^{(d+2)/2} \cdot \left(1.1d_P(V^{(0)}, V^*) + \|\mathbf{c} - \mathbf{c}^*\|_2\right) \\ &\quad\leq \frac{1}{100T}d_P(V^{(0)}, V^*),\end{aligned}$$

where the last step follows by our choice of $\eta_{\mathsf{vec}}$ in (23). So by (42) we conclude that as long as $\mathbb{E}[d_P(V^{(s)}, V^*)^2] > d_P(V^{(0)}, V^*)^2/1.1$ for all $s < t$,

$$\begin{aligned}\mathbb{E}\left[d_P(V^{(t)}, V^*)^2\right] &\geq \left(1 - \frac{\sqrt{1.1}}{100T}\right)\mathbb{E}\left[d_P(V^{(t-1)}, V^*)^2\right] \\ &\geq \left(1 - \frac{\sqrt{1.1}}{100T}\right)^t d_P(V^{(0)}, V^*)^2 \\ &\geq d_P(V^{(0)}, V^*)^2/1.1.\end{aligned}$$

By induction, $d_P(V^{(t)}, V^*)^2 \geq d_P(V^{(0)}, V^*)^2/1.1$ for all $0 \leq t < T$. Recalling that $\mu_X(\Theta^{(t)}) \geq (\alpha_{\mathsf{ndg}}/4) \cdot d_P(V^{(t)}, V^*)^2$, we conclude that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \mu_X(\Theta^{(t)})\right] \geq T \cdot (\alpha_{\mathsf{ndg}}/4) \cdot \left(d_P(V^{(0)}, V^*)^2/1.1\right)$$

as desired. ■

### I.6. Proof of Lemma 66

**Proof** We will bound each $\mathbb{E}_{x^0,\ldots,x^{t-1}}\left[\left|\mu_{E_1}(\Theta^{(t)}\right|\right]$ individually and apply triangle inequality.

By Lemma 61, for any realization of $x^0, \ldots, x^{t-1}$ giving rise to iterate $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$,

$$\left|\mu_{E_1}(\Theta^{(t)})\right| \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \|V^{(t)} - V^*\|_F \cdot d_P(V^{(t)}, V^*) \cdot \left(\|V^{(t)} - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2\right).$$

$$\leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \left(\|V^{(t)} - V^*\|_F^3 + \|V^{(t)} - V^*\|_F^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right).$$

By Lemma 58 and (22), we conclude that

$$\mathbb{E}\left[\left|\mu_{E_1}(\Theta^{(t)})\right|\right] \leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot \left(1.1 d_P(V^{(0)}, V^*)^3 + 1.1 d_P(V^{(0)}, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right)$$

$$\leq O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3.$$

The claim follows by summing over $t$. ■

### I.7. Proof of Lemma 67

**Proof** We will bound each $\mathbb{E}_{x^0,\ldots,x^{t-1}}\left[\left|\mu_{E_2}(\Theta^{(t)}\right|\right]$ individually and apply triangle inequality.

By Lemma 62, for any realization of $x^0, \ldots, x^{t-1}$ giving rise to iterate $\Theta^{(t)} = (\mathbf{c}, V^{(t)})$,

$$\begin{aligned}&\left|\mu_{E_2}(\Theta^{(t)})\right| \\ &\quad\leq \eta_{\mathsf{vec}} \cdot O(dr^3)^{d+2} \cdot \|V - V^*\|_F \cdot \left(\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2\right)^2. \\ &\quad\leq \eta_{\mathsf{vec}} \cdot O(dr^3)^{d+2} \cdot \left(\|V^{(t)} - V^*\|_F^3 + 2\|V^{(t)} - V^*\|_F^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2 + \|V^{(t)} - V^*\|_F \cdot \|\mathbf{c} - \mathbf{c}^*\|_2\right).\end{aligned}$$

By Lemma 58 and (22), we conclude that

$$\mathbb{E}\left[\left|\mu_{E_2}(\Theta^{(t)})\right|\right] \le O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot d_P(V^{(0)}, V^*)^3$$

The claim follows by summing over $t$. ■

### I.8. Proof of Lemma 68

Analogous to the proof of Lemma 50 in Appendix H.4, we will prove concentration by decomposing the MDS $\{\mu_X(\Theta^{(t)}) - X^{\Theta^{(t)},x^t}\}_{0\le t<T}$ into components corresponding to the decomposition (31). That is, define $\text{(A')}^{\Theta,x}$, $\text{(B')}^{\Theta,x}$, $\text{(C')}^{\Theta,x}$ to be the quantities in (31) for an iterate $\Theta$ and sample $x$. So by Observation 3, $\left\{\frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) - \text{(A')}^{\Theta^{(t)},x^t}\right\}$, $\{\text{(B')}^{\Theta^{(t)},x^t}\}$, and $\{\text{(C')}^{\Theta^{(t)},x^t}\}$ are MDS's, and for any $\Theta, x$,

$$\frac{1}{2\eta_{\mathsf{vec}}} X^{\Theta,x} = \text{(A')}^{\Theta,x} + \text{(B')}^{\Theta,x} + \text{(C')}^{\Theta,x}$$

by (31). We will show concentration for these MDS's separately.

**Lemma 86**

$$\sum_{t=0}^{T-1} \text{(A')}^{\Theta^{(t)},x^t} \ge T \cdot (\alpha_{\mathsf{ndg}}/5) \cdot d_P(V^{(0)}, V^*)^2$$

*with probability at least* $1-\delta$.

**Lemma 87**

$$\left|\sum_{t=0}^{T-1} \text{(B')}^{\Theta^{(t)},x^t}\right| \le T \cdot (\alpha_{\mathsf{ndg}}/60) \cdot d_P(V^{(0)}, V^*)^2$$

*with probability at least* $1-\delta$.

**Lemma 88**

$$\left|\sum_{t=0}^{T-1} \text{(C')}^{\Theta^{(t)},x^t}\right| \le T \cdot (\alpha_{\mathsf{ndg}}/60) \cdot d_P(V^{(0)}, V^*)^2$$

*with probability at least* $1-\delta$.

We prove these in the subsequent Appendices I.8.1, I.8.2, and I.8.3. Note that Lemma 68 follows easily from these three lemmas:

**Proof** [Proof of Lemma 68] The claim follows immediately from Lemmas 86, 87, and 88; triangle inequality; replacing $3\delta$ in the resulting union bound with $\delta$; and absorbing constant factors. ■

### I.8.1. PROOF OF LEMMA 86

**Proof** Observe that $\left\{\frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) - \text{(A')}^{\Theta^{(t)},x^t}\right\}$ is an MDS which satisfies one-sided bounds, as $\text{(A')}^{\Theta,x} \geq 0$ with probability one for any $\Theta, x$, so we wish to apply Lemma 14. To do so, we just need to bound the variances of the differences.

**Lemma 89** *For any $\Theta$, $Var_x[\text{(A')}^{\Theta,x}] \leq 2^{4d+4} \cdot d_P(V, V^*)^4$.*

**Proof** We will suppress superscripts $\Theta, x$ in this proof. $\mathrm{Var}[\text{(A')}] \leq \mathbb{E}[\text{(A')}^2]$, so it suffices to bound the latter. But note that $x^\top \Pi_V^\perp V^* \nabla p(V^\top x)$ is a polynomial, call it $f(x)$, of degree $d$ in the Gaussians $x_1, ..., x_n$. By Fact 6,

$$\mathbb{E}[\text{(A')}^2] = \mathbb{E}[f(x)^4] \leq \left(4^{d/2} \cdot \mathbb{E}[f(x)^2]^{1/2}\right)^4 \leq 2^{4d} \cdot \mathbb{E}[f(x)^2]^2 = 2^{4d} \cdot \mathbb{E}[\text{(A')}]^2 \leq 2^{4d+4} \cdot d_P(V, V^*)^4, \tag{43}$$

where the last step is by Lemma 63. ■

We can now complete the proof of Lemma 86. By Lemma 57 and Lemma 89, if $\eta_{\mathsf{vec}}$ satisfies (23), then with probability $1-\delta$ we have that for all $0 \leq t < T$,

$$\frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) - \text{(A')}^{\Theta^{(t)},x^t} \leq \frac{1}{2\eta_{\mathsf{vec}}}\mu_X(\Theta^{(t)}) \leq 4d_P(V^{(t)}, V^*)^2 \leq 4.84 d_P(V^{(0)}, V^*)^2.$$

$$\operatorname*{Var}_{x^t}[\text{(A')}^{\Theta^{(t)},x^t}] \leq 1.1^4 \cdot 2^{4d+4} \cdot d_P(V^{(0)}, V^*)^4$$

Applying Lemma 14 with the parameter $\sigma^2$ taken to be $T \cdot 1.1^4 \cdot 2^{4d+4} \cdot d_P(V^{(0)}, V^*)^4$, we get

$$\Pr\left[\sum_{t=0}^{T-1} \text{(A')}^{\Theta^{(t)},x^t} \geq \frac{1}{2\eta_{\mathsf{vec}}} \sum_{t=0}^{T-1} \mathbb{E}\left[\mu_X\left(\Theta^{(t)}\right)\right] - O\left(4^d \log(1/\delta)\sqrt{T} \cdot d_P(V^{(0)}, V^*)^2\right)\right] \geq 1 - 2\delta,$$

where the expectation in $\mathbb{E}\left[X\left(\Theta^{(t)}\right)\right]$ is over the randomness of the samples $x^0, ..., x^{t-1}$.

By Lemma 65, we conclude that

$$\sum_{t=0}^{T-1} \text{(A')}^{\Theta^{(t)},x^t} \geq T \cdot (\alpha_{\mathsf{ndg}}/4.4) \cdot d_P(V^{(0)}, V^*)^2 - O\left(4^d \log(1/\delta)\sqrt{T} \cdot d_P(V^{(0)}, V^*)^2\right) \tag{44}$$

with probability at least $1-2\delta$. Taking $T$ according to (24) will certainly ensure the right-hand side of (44) is at least $T \cdot (\alpha_{\mathsf{ndg}}/5) \cdot d_P(V^{(0)}, V^*)^2$. The proof is completed by replacing $2\delta$ in the above with $\delta$ and absorbing the resulting constant factors. ■

### I.8.2. PROOF OF LEMMA 87

**Proof** For fixed $x^1, ..., x^{t-1}$, the martingale difference $\text{(B')}^{\Theta^{(t)},x^t}$ is a polynomial of degree $2d$ in $x^t$, so by Lemma 13 we just need to upper bound the second moments of the differences, which we do in the following lemma.

**Lemma 90** *For any $\Theta$, $\mathbb{E}_x[(\text{(B')}^{\Theta,x})^2] \leq d_P(V, V^*)^2 \cdot \|V - V^*\|_F^2 \cdot O(r^2) \cdot \exp(O(d))$.*

**Proof** By Cauchy-Schwarz,

$$\begin{aligned}\mathbb{E}\left[\text{(B')}^2\right] &\le \mathbb{E}\left[\left(x^\top \Pi_V (V^* - V)\nabla\right)^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \Pi_V^\perp V^* \nabla\right)^4\right]^{1/2} \\ &= \mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}\left[\left(g^\top V^\top (V^* - V)\nabla p(g)\right)^4\right]^{1/2} \cdot \mathbb{E}\left[\text{(A')}^2\right]^{1/2} \\ &\le \mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}\left[\left(g^\top (\mathbf{I} - V^\top V^*)\nabla p(g)\right)^4\right]^{1/2} \cdot 2^{2d+2} \cdot d_P(V, V^*)^2, \end{aligned} \tag{45}$$

where the third step follows by (43). It remains to bound the first factor in (45). As this factor is independent of $n$, we do not need a particularly sharp bound. We have

$$\begin{aligned}\mathbb{E}_g\left[\left(g^\top (\mathbf{I} - V^\top V^*)\nabla p(g)\right)^4\right]^{1/2} &\le \|\mathbf{I} - V^\top V^*\|_2^2 \cdot \mathbb{E}_g[\|g\|_2^4 \cdot \|\nabla p(g)\|_2^4]^{1/2} \\ &\le \|V - V^*\|_F^2 \mathbb{E}_g[\|g\|_2^8]^{1/4} \cdot \mathbb{E}_g[\|\nabla p(g)\|_2^8]^{1/4} \\ &\le \|V - V^*\|_F^2 \cdot 3(r+1) \cdot (rd \cdot 7^d \cdot \mathrm{Var}[p]) \\ &\le \|V - V^*\|_F^2 \cdot O(r^2 d \cdot 7^d),\end{aligned}$$

where the second step follows by Lemma 26, the third step follows by Corollary 16 and Lemma 18 applied to $q = 4$, and the last step follows by noting that $\mathrm{Var}[p] = O(1)$ by triangle inequality and absorbing constant factors. The claimed bound follows. ■

We now complete the proof of Lemma 87. By Lemma 57, $d_P(V^{(t)}, V^*) \le \|V^{(t)} - V^*\|_F \le 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \le t \le T$ with probability at least $1 - \delta$, in which case Lemma 90 implies that for every $0 \le t < T$,

$$\mathbb{E}\left[\left(\text{(B')}^{\Theta^{(t)}, x^t}\right)^2 \,\middle|\, x^1, ..., x^{t-1}\right] \le d_P(V^{(0)}, V^*)^4 \cdot O(r^2) \cdot \exp(O(d))$$

with probability at least $1 - \delta$. So by Lemma 13,

$$\left|\sum_{t=0}^{T-1} \text{(B')}^{\Theta^{(t)}, x^t}\right| \le (\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot d_P(V^{(0)}, V^*)^2 \cdot O(r) \cdot \exp(O(d))$$

with probability at least $1 - 2\delta$. By taking $T$ according to (24), we ensure that this quantity is upper bounded by a negligible multiple of $T \cdot (\alpha_{\mathsf{ndg}}/5) \cdot d_P(V^{(0)}, V^*)^2$ as desired. The proof is completed by replacing $2\delta$ in the above with $\delta$ and absorbing the resulting constant factors. ■

### I.8.3. PROOF OF LEMMA 88

**Proof** As in the proof of Lemma 87, for fixed $x^1, ..., x^{t-1}$, the martingale difference $\text{(C')}^{\Theta^{(t)}, x^t}$ is a polynomial of degree $2d$ in $x^t$, so by Lemma 13 we just need to upper bound the second moments of the differences, which we do in the following lemma.

**Lemma 91** *For any* $\Theta$, $\mathbb{E}_x[(\text{(C')}^{\Theta, x})^2] \le d_P(V, V^*)^2 \cdot \|\mathbf{c} - \mathbf{c}^*\|_2^2 \cdot \exp(O(d))$.

**Proof** By Cauchy-Schwarz,

$$\begin{aligned}\mathbb{E}\left[\text{Ⓒ'}^2\right] &\le \mathbb{E}\left[(\delta(V^\top x)^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \Pi_V^\perp V^* \nabla\right)^4\right]^{1/2} \\ &= \mathbb{E}_{g\sim\mathcal{N}(0,\mathbf{I}_r)}\left[\delta(g)^4\right]^{1/2} \cdot \mathbb{E}\left[\text{Ⓐ'}^2\right]^{1/2} \\ &\le \left(3^d \cdot \|\mathbf{c}-\mathbf{c}^*\|_2^2\right) \cdot \left(2^{2d+2} \cdot d_P(V,V^*)^2\right),\end{aligned}$$

where the third step follows by Fact 6 and (43). ■

We now complete the proof of Lemma 88. By Lemma 57, $d_P(V^{(t)},V^*) \le \|V^{(t)}-V^*\|_F \le 1.1 \cdot d_P(V^{(0)},V^*)$ for every $0 \le t \le T$ with probability at least $1-\delta$, in which case Lemma 91 implies that for every $0 \le t < T$,

$$\mathbb{E}[(\text{Ⓒ'}^{\Theta^{(t)},x^t})^2 | x^1,...,x^{t-1}] \le d_P(V^{(0)},V^*) \cdot \|\mathbf{c}-\mathbf{c}^*\|_2 \cdot \exp(O(d))$$

with probability at least $1-\delta$. So by Lemma 13,

$$\left|\sum_{t=0}^{T-1} \text{Ⓒ'}^{\Theta^{(t)},x^t}\right| \le (\log(1/\delta)\cdot d)^{c_1 d} \cdot \sqrt{T} \cdot d_P(V^{(0)},V^*) \cdot \|\mathbf{c}-\mathbf{c}^*\|_2 \cdot \exp(O(d))$$

with probability at least $1-2\delta$. By taking $T$ satisfying the bound in the lemma statement and invoking (22), we ensure that this quantity is upper bounded by a negligible multiple of $T\cdot(\alpha_{\mathsf{ndg}}/5)\cdot d_P(V^{(0)},V^*)^2$ as desired. As in the proof of Lemma 86, the proof is completed by replacing $2\delta$ in the above with $\delta$ and absorbing the resulting constant factors. ■

### I.8.4. PROOF OF LEMMAS 69 AND 70

We will apply Lemma 13 to the MDS's $\left\{E_1^{\Theta^{(t)},x^t} - \mu_{E_1}(\Theta^{(t)})\right\}$ and $\left\{E_2^{\Theta^{(t)},x^t} - \mu_{E_2}(\Theta^{(t)})\right\}$. As in the analysis of the MDS's for Lemmas 87 and 88, the differences in these MDS's are polynomials of degree at most $2d$, so we just need to bound the second moments of their differences. We do so in the following two lemmas.

**Lemma 92** *For any* $\Theta$,

$$\mathbb{E}_x[(E_1^{\Theta,x})^2] \le O(\eta_{\mathsf{vec}}^2)\cdot O(dr^3)^{d+1}\cdot\|V-V^*\|_F^2 \cdot d_P(V,V^*)^2 \cdot (\|\mathbf{c}-\mathbf{c}^*\|_2+\|V^*-V\|_F)^2$$

**Proof** We have that

$$\begin{aligned}\frac{1}{4\eta_{\mathsf{vec}}^2}\mathbb{E}\left[\left(E_1^{\Theta,x}\right)^2\right] &= \mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^2 \cdot \left(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta\right)^2\right] \\ &\le \mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^4\right]^{1/2} \cdot \mathbb{E}\left[\left(x^\top \cdot \Pi_V^\perp V^* \cdot \Delta\right)^4\right]^{1/2} \\ &= \mathbb{E}\left[(\mathfrak{R}^{\Theta,x})^4\right]^{1/2} \cdot \mathbb{E}[\text{Ⓐ'}^2]^{1/2} \\ &\le O(dr^3)^{d+1}\cdot\|V-V^*\|_F^2 \cdot d_P(V,V^*)^2 \cdot (\|\mathbf{c}-\mathbf{c}^*\|_2+\|V^*-V\|_F)^2,\end{aligned}$$

where the second step follows by Cauchy-Schwarz, the third step follows by definition of Ⓐ', and the fourth step follows by Lemma 42. ■

**Lemma 93** *For any $\Theta$, if $\eta_{\mathsf{vec}} \leq O(1/n)$, then*

$$\mathbb{E}_x[(E_2^{\Theta,x})^2] \leq O(\eta_{\mathsf{vec}}^2) \cdot (64dr^3)^{2d+4} \cdot \|V - V^*\|_F^2 \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^4$$

**Proof** By triangle inequality and Jensen's, $\mathbb{E}[(E_2^{\Theta,x})^2]^{1/2} = \mathbb{E}[\langle \mathcal{E}, V - V^* \rangle^2]^{1/2}$ is at most

$$\mathbb{E}\left[(\cos(\sigma\eta_{\mathsf{vec}}) - 1)^2 \cdot \langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^* \rangle^2\right]^{1/2} + \mathbb{E}\left[(\sin(\sigma\eta_{\mathsf{vec}}) - \sigma\eta_{\mathsf{vec}})^2 \cdot \langle \widehat{h}\widehat{\nabla}^\top, V - V^* \rangle^2\right]^{1/2}$$

By Holder's and the fact that $|\cos(x) - 1| \leq x^2/2$ and $|\sin(x) - x| \leq x^2/\pi$ for all $x \geq 0$, we may upper bound the first term by

$$\mathbb{E}\left[(\cos(\sigma\eta_{\mathsf{vec}}) - 1)^2\right]^{1/2} \cdot \max_{\widehat{\nabla}} \left|\langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^* \rangle\right| \leq O(\eta_{\mathsf{vec}}^2) \cdot \mathbb{E}[\sigma^4]^{1/2} \cdot \max_{\widehat{\nabla}} \left|\langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^* \rangle\right|$$

and the second term by

$$\mathbb{E}\left[(\sin(\sigma\eta_{\mathsf{vec}}) - \sigma\eta_{\mathsf{vec}})^2\right]^{1/2} \cdot \max_{\widehat{h},\widehat{\nabla}} \left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^* \rangle\right| \leq O(\eta_{\mathsf{vec}}^2) \cdot \max_{\widehat{h},\widehat{\nabla}} \left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^* \rangle\right|.$$

So $\mathbb{E}[(E_2^{\Theta,x})^2]^{1/2}$ is at most

$$\begin{aligned}
&O(\eta_{\mathsf{vec}}^2) \cdot \mathbb{E}[\sigma^4]^{1/2} \cdot \left(\max_{\widehat{\nabla}} \left|\langle V \cdot \widehat{\nabla}\widehat{\nabla}^\top, V - V^* \rangle\right| + \max_{\widehat{h},\widehat{\nabla}} \left|\langle \widehat{h}\widehat{\nabla}^\top, V - V^* \rangle\right|\right) \\
&\quad \leq O(\eta_{\mathsf{vec}}^2) \cdot O(n) \cdot (64dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \cdot \|V - V^*\|_F \\
&\quad \leq O(\eta_{\mathsf{vec}}) \cdot (64dr^3)^{d+2} \cdot (\|V - V^*\|_F + \|\mathbf{c} - \mathbf{c}^*\|_2)^2 \cdot \|V - V^*\|_F,
\end{aligned}$$

where the first step follows by Lemma 55, Lemma 84, and Lemma 85, and the sixth follows by the assumption that $\eta_{\mathsf{vec}} \leq O(1/n)$. ■

We are now ready to complete the proofs of Lemma 69 and 70.

**Proof** [Proof of Lemma 69] By Lemma 57, $d_P(V^{(t)}, V^*) \leq \|V^{(t)} - V^*\|_F \leq 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \leq t \leq T$ with probability at least $1 - \delta$, in which case Lemma 92 implies that for every $0 \leq t < T$,

$$\mathbb{E}[(E_1^{\Theta^{(t)},x^t})^2 | x^1, ..., x^{t-1}] \leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{d+1} \cdot d_P(V^{(0)}, V^*)^4 \cdot \left(\|\mathbf{c} - \mathbf{c}^*\|_2 + d_P(V^{(0)}, V^*)\right)^2$$

$$\leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{d+1} \cdot d_P(V^{(0)}, V^*)^6$$

with probability at least $1 - \delta$. So by Lemma 13,

$$\begin{aligned}
&\left|\sum_{t=0}^{T-1} \left(E_1^{\Theta^{(t)},x^t} - \mathbb{E}\left[\mu_{E_1}(\Theta^{(t)})\right]\right)\right| \\
&\qquad \leq (\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3
\end{aligned}$$

with probability at least $1 - 2\delta$. By Lemma 66, we conclude that

$$\left|\sum_{t=0}^{T-1} E_1^{\Theta^{(t)},x^t}\right| \leq O(\sqrt{T} \cdot \eta_{\mathsf{vec}}) \cdot O(dr^3)^{(d+1)/2} \cdot d_P(V^{(0)}, V^*)^3 \cdot \left((\log(1/\delta) \cdot d)^{c_1 d} + \sqrt{T}\right)$$

By taking $T$ according to (24) and using the bound (21), we ensure that this quantity is upper bounded by a negligible multiple of $T \cdot \eta_{\mathsf{vec}} \cdot (\alpha_{\mathsf{ndg}}/3) \cdot d_P(V^{(0)}, V^*)^2$ as desired. As usual, the proof is completed by replacing $2\delta$ in the above with $\delta$ and absorbing the resulting constant factors. ■

**Proof** [Proof of Lemma 70] By Lemma 57, $\|V^{(t)} - V^*\|_F \leq 1.1 \cdot d_P(V^{(0)}, V^*)$ for every $0 \leq t \leq T$ with probability at least $1 - \delta$, in which case Lemma 93 implies that for every $0 \leq t < T$,

$$\mathbb{E}[(E_2^{\Theta^{(t)},x^t})^2 | x^1, ..., x^{t-1}] \leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{2d+4} \cdot d_P(V^{(0)}, V^*)^2 \cdot \left(\|\mathbf{c} - \mathbf{c}^*\|_2 + d_P(V^{(0)}, V^*)\right)^4$$

$$\leq O(\eta_{\mathsf{vec}}^2) \cdot O(dr^3)^{2d+4} \cdot d_P(V^{(0)}, V^*)^6$$

with probability at least $1 - \delta$. So by Lemma 13,

$$\left|\sum_{t=0}^{T-1} \left( E_2^{\Theta^{(t)},x^t} \sum_{t=0}^{T-1} \mathbb{E}\left[\mu_{E_2}(\Theta^{(t)})\right]\right)\right| \leq (\log(1/\delta) \cdot d)^{c_1 d} \cdot \sqrt{T} \cdot O(\eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot d_P(V^{(0)}, V^*)^3$$

with probability at least $1 - 2\delta$. By (67), we conclude that

$$\left|\sum_{t=0}^{T-1} E_2^{\Theta^{(t)},x^t}\right| \leq O(\sqrt{T} \cdot \eta_{\mathsf{vec}}) \cdot O(dr^3)^{d+2} \cdot d_P(V^{(0)}, V^*)^3 \cdot \left((\log(1/\delta) \cdot d)^{c_1 d} + \sqrt{T}\right)$$

By taking $T$ according to (24) and using the bound (21), we ensure that this quantity is upper bounded by a negligible multiple of $T \cdot \eta_{\mathsf{vec}} \cdot (\alpha_{\mathsf{ndg}}/3) \cdot d_P(V^{(0)}, V^*)^2$ as desired. As usual, the proof is completed by replacing $2\delta$ in the above with $\delta$ and absorbing the resulting constant factors. ■