Robust estimation via generalized quasi-gradients

Banghua Zhu, Jiantao Jiao, Jacob Steinhardt*

May 29, 2020

Abstract

We explore why many recently proposed robust estimation problems are efficiently solvable, even though the underlying optimization problems are non-convex. We study the loss landscape of these robust estimation problems, and identify the existence of "generalized quasi-gradients". Whenever these quasi-gradients exist, a large family of low-regret algorithms are guaranteed to approximate the global minimum; this includes the commonly-used filtering algorithm.

For robust mean estimation of distributions under bounded covariance, we show that any first-order stationary point of the associated optimization problem is an approximate global minimum if and only if the corruption level $\epsilon < 1/3$. Consequently, any optimization algorithm that aproaches a stationary point yields an efficient robust estimator with breakdown point $1/3$. With careful initialization and step size, we improve this to $1/2$, which is optimal.

For other tasks, including linear regression and joint mean and covariance estimation, the loss landscape is more rugged: there are stationary points arbitrarily far from the global minimum. Nevertheless, we show that generalized quasi-gradients exist and construct efficient algorithms. These algorithms are simpler than previous ones in the literature, and for linear regression we improve the estimation error from $O(\sqrt{\epsilon})$ to the optimal rate of $O(\epsilon)$ for small ϵ assuming certified hypercontractivity. For mean estimation with near-identity covariance, we show that a simple gradient descent algorithm achieves breakdown point $1/3$ and iteration complexity $\tilde{O}(d/\epsilon^2)$.

Contents

*Banghua Zhu is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Jiantao Jiao is with the Department of Electrical Engineering and Computer Sciences and the Department of Statistics, University of California, Berkeley. Jacob Steinhardt is with the Department of Statistics and the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Email: {banghua, jiantao,jsteinhardt}@berkeley.edu.

1 Introduction and main results

We study the problem of robust estimation in the presence of outliers. In general, this means that we observe a dataset of n points, and an adversary can corrupt (via additions or deletions) any subset of ϵn of the points. Our goal is to estimate some property of the original points (such as the mean) under some assumptions (such as the good points having bounded covariance). In addition to mean estimation (Huber, 1973; Donoho, 1982; Beran, 1977; Davies et al., 1992; Adrover and Yohai, 2002; Hubert and Debruyne, 2010; Diakonikolas et al., 2016), we are interested linear regression (Diakonikolas et al., 2019c; Klivans et al., 2018) and covariance estimation (Kothari and Steurer, 2017; Diakonikolas et al., 2017).

Robust estimation has been extensively studied, and a general issue is how to design computationally efficient estimators. Recent papers have provided general (inefficient) recipes for solving these problems, showing that it suffices to solve an optimization problem that removes outliers to obtain a nice distribution—where "nice" can be formalized and is problem-dependent (Steinhardt, 2018; Zhu et al., 2019). Although this recipe in general leads to non-convex or otherwise seemingly intractable estimators, a variety of efficient algorithms have been proposed for many problems (Lai et al., 2016; Diakonikolas et al., 2019a; Kothari and Steurer, 2017; Diakonikolas et al., 2017; Steinhardt, 2018; Klivans et al., 2018; Dong et al., 2019; Cheng et al., 2019b,a).

The large variety of computationally efficient estimators suggests that robust estimation is easier than we would have expected given its non-convexity. How can we explain this? Here we analyze the non-convex optimization landscape for several problems—mean estimation, covariance estimation, and linear regression—and show that, while the landscape is indeed non-convex, it is nevertheless nice enough to admit efficient optimization algorithms.

This claim is easiest to formalize for mean estimation under bounded covariance. In this case, we observe points $X_1, \ldots, X_n \in \boldsymbol{R}^d$, such that a subset S of $(1-\epsilon)n$ "good" points is guaranteed to have bounded covariance: $\|\Sigma_{p_S}\| \leq \sigma$, where p_S is the empirical distribution over S, Σ_p is the covariance matrix under p, and $\|\cdot\|$ is operator norm. As shown in Diakonikolas et al. (2016); Steinhardt et al. (2018), estimating the mean of p_S only requires finding *any* large subset of the data with small covariance, as in the (non-convex) optimization problem below:

Example 1.1 (Mean estimation with bounded operator norm of covariance matrix)**.** *Let $\Delta_{n,\epsilon}$ denote the set of ϵ-deleted distribution:*

$$\Delta_{n,\epsilon} \triangleq \{q \mid \sum_{i=1}^{n} q_i = 1, 0 \leq q_i \leq \frac{1}{(1-\epsilon)n}\}, \tag{1}$$

where q_i is the probability q assigns to point X_i. We solve the feasibility problem[1]

$$\begin{array}{ll} \textit{find} & q \\ \textit{subject to} & q \in \Delta_{n,\epsilon}, \|\Sigma_q\| \leq \sigma'^2, \end{array} \tag{2}$$

where the parameter $\sigma'^2 \geq \sigma^2$ depends on ϵ, and is close to σ^2 when ϵ is small. In this case the mean μ_q satisfies $\|\mu_q - \mu_{p_S}\| = O((\sigma + \sigma')\sqrt{\epsilon})$ for any feasible q.

Although $\Delta_{n,\epsilon}$ is a convex constraint on q, the function $\|\Sigma_q\|$ is non-convex in q. In dimension one, it reduces to the variance of q, which is concave, and not even quasiconvex. Nevertheless, we show that all stationary points (or approximate stationary points) of Example 1.1 are approximate global optima (formal version Theorem 2.1):

[1] We discuss other formulations of the mean estimation problem in Appendix B.

Theorem 1.1 (Informal). *All first-order stationary points of minimizing $\|\Sigma_q\|$ subject to $q \in \Delta_{n,\epsilon}$ are an approximate global minimum with worst case approximation ratio $(1-\epsilon)^2/(1-3\epsilon)^2$ (and infinite approximation ratio if $\epsilon \geq 1/3$). Approximate stationary points are also approximate global minimum, and gradient descent algorithms can approach them efficiently.*

The approximation ratio $(1-\epsilon)^2/(1-3\epsilon)^2$ is tight even in the constant. For $\epsilon \geq 1/3$, we exhibit examples where stationary points (indeed, local minima) can be arbitrarily far from the global minimum (Theorem C.1). However, we show that a carefully initialized gradient descent algorithm approximates the global optimum whenever $\epsilon < 1/2$ (Theorem 4.2), which is the highest breakdown point for any translation equivariant mean estimator (Rousseeuw and Leroy, 1987, Page 270). This gradient algorithm is an instance of the commonly-used filtering algorithm in the literature (Li, 2018, 2019; Diakonikolas et al., 2017; Steinhardt, 2018), but we provide a tighter analysis with optimal breakdown point. This algorithm achieves approximation ratio $2(1-\epsilon)/(1-2\epsilon)^2$ (Theorem 4.2) for all $\epsilon \in (0, 1/2)$.

We might hope that the optimization landscape is similarly well-behaved for other robust estimation problems beyond mean estimation. However, this is not true in general. For linear regression, we show that the analogous optimization problem can have arbitrarily bad stationary points even as $\epsilon \to 0$ (Section 3). We nevertheless show that the landscape is tractable, by identifying a property that we call *generalized quasi-gradients*. Such quasi-gradients allow many gradient descent algorithms to approximate the global optima.

Definition 1.1 (Generalized quasi-gradient). *In the optimization problem $\min_{q \in A} F(q)$, we say $g(q)$ is a generalized quasi-gradient with parameter $C \geq 1$ if the following holds for all $q, p \in A$:*

$$\langle g(q), q - p \rangle \leq 0 \implies F(q) \leq C \cdot F(p). \tag{3}$$

Moreover, we call $g(q)$ a strict generalized quasi-gradient with parameters $C_1(\alpha, \beta), C_2(\alpha, \beta)$ if the following holds for all $q, p \in A, \alpha, \beta \geq 0$,

$$\langle g(q), q - p \rangle \leq \alpha \langle |g(q)|, p \rangle + \beta \implies F(q) \leq C_1(\alpha, \beta) \cdot F(p) + C_2(\alpha, \beta), \tag{4}$$

where $|g(q)|$ is the point-wise absolute value of vector $g(q)$.

The conventional quasi-gradient is a generalized quasi-gradient with parameter $C = 1$ (Boyd and Mutapcic, 2007), which only exists for quasi-convex functions. Our next result shows that even though the target functions we consider are not quasi-convex ($q \mapsto \|\Sigma_q\|$ is not quasi-convex as a concrete example), we can still find generalized quasi-gradients:

Theorem 1.2 (Informal). *Generalized quasi-gradient exists for all the optimization problems investigated in the paper, including mean estimation (Example 1.1, 2.1), linear regression (Example 3.1), and joint mean and covariance estimation (Example 3.2). Here the set $A = \Delta_{n,\epsilon}$.*

Strict generalized quasi-gradients are important because every low-regret algorithm can approach points q such that the inequality $\langle g(q), q - p_S \rangle \leq 0$ approximately holds (Nesterov, 2009; Arora et al., 2012). This then immediately implies that $F(q) \leq C_1 \cdot F(p_S) + C_2$ (see e.g. Theorem 2.2). Thus once we identify (strict) generalized quasi-gradients that are efficiently computable, we immediately obtain a family of algorithms that approximately solve the feasibility problem. We elaborate on our concrete algorithm constructions in Section 1.2.

1.1 Constructing generalized quasi-gradients

We next describe generalized quasi-gradients for several tasks. Our starting point is the following optimization problem, which generalizes Example 1.1:

***Problem* 1.1** (Approximate Minimum Distance (AMD) functional with TV)**.** *We solve the feasibility problem*

$$\begin{array}{ll} \textit{find} & q \\ \textit{subject to} & q \in \Delta_{n,\epsilon}, F(q) \leq \xi, \end{array} \tag{5}$$

where $\Delta_{n,\epsilon}$ is defined in (1).

The only difference between optimization problem (5) and (2) is that we have replaced $\|\Sigma_q\|$ with a more general function $F(q)$. We often also consider the minimization form $\min_{q \in \Delta_{n,\epsilon}} F(q)$.

The appropriate F to use is problem-dependent and depends on what distributional assumptions we are willing to make. Zhu et al. (2019) provides a general treatment for how to choose F. For linear regression (Example 3.1, also in Zhu et al. (2019, Example 3.2)) and joint mean and covariance estimation under the Mahalanobis distances (Example 3.2, also in Kothari and Steurer (2017)), the appropriate F is closely related to the hypercontractivity coefficient of q, represented by the function

$$F_1(q) = \sup_{v \in \boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top X)^4]}{\mathbb{E}_q[(v^\top X)^2]^2} \leq \kappa^2. \tag{6}$$

As with the covariance $\|\Sigma_q\|$, the function $F_1(q)$ in (6) is generally not a convex function of q. Indeed, if $d = 1$, then its sublevel set is the complementary set of a convex set, which makes the function not even quasi-convex. But more problematically, as mentioned above, we can construct first-order stationary points of (5) where $F_1(q)$ is arbitrarily big while ϵ and $F_1(p_S)$ are both small (Theorem D.1).

Nevertheless, the following function (among others) is a generalized quasi-gradient for $F_1(q)$ with $C = 4$ when $9\kappa^2\epsilon \leq 1$:

$$g_1(X; q) = \mathbb{E}_q[(v^\top X)^4], \text{ where } v \in \arg\max_{v \in \boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top X)^4]}{\mathbb{E}_q[(v^\top X)^2]^2}, \tag{7}$$

which we analyze in Section 3.2. Since the supremum in (7) is not generally efficiently computable, in practice we make the stronger assumption that p_S has Sum-of-Squares (SoS) certifiable hypercontractivity, and construct an efficient relaxation of (7) using pseudoexpectations (see Appendix A for formal definitions).

Given a quasi-gradient for F_1, we are most of the way to designing algorithms for joint mean and covariance estimation, as well as linear regression. For joint mean and covariance (Example 3.2), we actually need to handle a centered version of F_1, where we consider $X - \mu_q$ instead of X. We show in Section 3.3 that the analogous quasi-gradient has constant $C = 7$ when $200\kappa^2\epsilon < 1$.

For linear regression (Example 3.1), we do not need to center X, but we do need to impose the following bounded noise condition in addition to the bound on F_1:

$$F_2(q) = \frac{\mathbb{E}_q[(Y - \theta(q)^\top X)^2 (v^\top X)^2]}{\mathbb{E}_q[(v^\top X)^2]} \leq \sigma^2, \tag{8}$$

where $\theta(p) = \arg\min_{\theta \in \boldsymbol{R}^d} \mathbb{E}_p[(Y - \theta^\top X)^2]$ is the optimal regression parameters for p. The corresponding quasi-gradient is

$$g_2(X; q) = (Y - X^\top \theta(q))^2 (v^\top X)^2, \text{ where } v \in \arg\max_{v \in \boldsymbol{R}^d} \frac{\mathbb{E}_q[(Y - X^\top \theta(q))^2 (v^\top X)^2]}{\mathbb{E}_q[(v^\top X)^2]}. \tag{9}$$

which we show in Section 3.2 has $C = 3$ when $64\kappa^3\epsilon < 1$.

Other robust estimation problems have been studied such as sparse mean estimation, sparse PCA and moment estimation (Li, 2017; Diakonikolas et al., 2019b; Li, 2018). For most cases we are aware of, we can similarly construct generalized quasi-gradients and obtain efficient algorithms. As a concrete example we exhibit quasig-radients for sparse mean estimation in Theorem D.2.

1.2 Efficient algorithms from generalized quasi-gradients

Having constructed (strict) generalized quasi-gradients for several robust estimation problems, we next show that such generalized quasi-gradients enable efficient optimization. Specifically, any algorithm with vanishing regret $\sum_t \langle g(q_t), q_t - p_S \rangle = o(t)$ as $t \to \infty$ converges to an approximate global minimum, assuming g is a strict generalized quasi-gradient. Typically, any online learning algorithm will yield vanishing regret, but the robust setting is complicated by the fact that online convergence rates typically depend on the maximum norm of the gradients, and an adversary can include outliers that make these gradients arbitrarily large. We provide two strategies to handle this: explicit low-regret with naïve pruning, and filtering. The first removes large points as a preprocessing step, after which we can employ standard regret bounds; the second picks the step size carefully to ensure convergence in $O(\epsilon n)$ steps even if the gradients can be arbitrarily large. Both algorithms are a form of gradient descent on q using the generalized quasi-gradients.

For the explicit low-regret algorithm, after the gradient step we project the distribution back to the set of deleted distribution $\Delta_{n,\epsilon} = \{q \mid \sum_{i=1}^n q_i = 1, 0 \le q_i \le \frac{1}{(1-\epsilon)n}\}$ after one-step update. This explicitly ensures that $q \in \Delta_{n,\epsilon}$. The performance of explicit low-regret algorithms are analyzed in Lemma 4.2.

For the filter algorithm, we only project the distribution back to the probability simplex $\Delta_n = \{q \mid \sum_{i=1}^n q_i = 1, \forall i \in [n], q_i \ge 0\}$. We show that if the strict generalized quasi-gradient is coordinate-wise non-negative with appropriate parameters (Lemma 4.4), then the algorithm will output some q with $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$. The set of q satisfying this property is a supserset of $\Delta_{n,\epsilon}$, but is exactly what we need for statistical inference. Our analysis closely follows previous analyses (see e.g. Li (2018, 2019); Steinhardt (2018); Diakonikolas et al. (2017)), but we provide tighter bounds at several points that lead to better breakdown point.

Both algorithms converge to approximate global optima, but need different assumptions to achieve fast convergence. The explicit low-regret algorithm requires us to identify and remove bad points that can blow up the gradient, which is only possible in some settings. The filtering algorithm works if the strict generalized quasi-gradients are non-negative with appropriate parameters, which again only holds in some settings. Together, however, these cover all the settings we need for our analysis.

Concrete algorithmic results. Our result for mean estimation with bounded covariance provides an efficient algorithm with breakdown point 1/2 and iteration complexity ϵn. Our analysis is the first that achieves both optimal breakdown point and optimal rate $\Theta(\sqrt{\epsilon})$ for $\epsilon \le 1/4$ in this task.

For mean estimation with near identity covariance, the projected gradient algorithm has breakdown point 1/3 and iteration complexity $\tilde{O}(d/\epsilon^2)$ (Theorem 4.5), which improves the iteration

complexity of $\tilde{O}(nd^3/\epsilon)$ in the concurrent work of Cheng et al. (2020), since $\epsilon \geq 1/n$ without loss of generality. The breakdown point is also consistent with the lower bound in Theorem 1.1 if we allow arbitrary initialization.

The generalized quasi-gradients for linear regression immediately yield a filtering algorithm that achieves estimation error $O(\epsilon)$ for $\epsilon < 1/(200\kappa^3)$ under certified hypercontractivity (Theorem 4.3), which is optimal and improves over the previous bound of $O(\sqrt{\epsilon})$ in Klivans et al. (2018). We similarly obtain a filtering algorithm for joint mean and covariance estimation, which matches the $O(\epsilon^{3/4})$ rate in mean estimation and $O(\sqrt{\epsilon})$ in covariance estimation for $\epsilon \leq 1/(4\kappa^2)$ in Kothari and Steurer (2017) but with a simpler algorithm (Theorem 4.4).

1.3 Notation and discussion on the corruption model

Notations: We use X for random variables, p for the population distribution, and p_n for the corresponding empirical distribution from n samples. Blackbold letter $\mathbb{E}$ is used for expectation. We write $A \lesssim B$ to denote that $A \leq CB$ for an absolute constant C. We let $\mu_p = \mathbb{E}_p[X]$ and $\Sigma_p = \mathbb{E}_p[(X-\mu_p)(X-\mu_p)^\top]$ denote the mean and covariance of a distribution p. We also use $\mathsf{Cov}(X)$ to denote the covariance of a random variable X.

We use $\mathsf{TV}(p,q) = \sup_A p(A) - q(A)$ to denote the total variation distance between p and q. We use $\mathsf{supp}(\cdot)$ to denote the support of a distribution, $\mathsf{conv}(\cdot)$ to denote convex hull. We say that a distribution q is an *ϵ-deletion* of p if for any set A, $q(A) \leq p(A)/(1-\epsilon)$. This implies that $\mathsf{TV}(p,q) \leq \epsilon$ since $\mathsf{TV}(p,q) = \sup_A q(A) - p(A) \leq \sup_A \epsilon q(A) \leq \epsilon$. We use $\Delta_n = \{p \mid \sum_{i=1}^n p_i = 1\}$ to denote the probability simplex.

We write $f(x) = O(g(x))$ for $x \in A$ if there exists some positive real number M such that $|f(x)| \leq Mg(x)$ for all $x \in A$. If A is not specified, we have $|f(x)| \leq Mg(x)$ for all $x \in [0, +\infty)$ (thus the notation is non-asymptotic). We use $\tilde{O}(\cdot)$ to be the big-O notation ignoring logarithmic factors.

We discuss our formal corruption model here. In the traditional finite-sample total-variation corruption model in (Donoho, 1982), it is assumed that there exists a set of n good samples, and the adversary is allowed to either add or delete an ϵ fraction of points. In contrast, throughout the paper we assume that there exists a set of n possibly corrupted samples, and a set S of $(1-\epsilon)n$ of them are good samples. The final goal is to estimate some property of the good samples, assuming that the good samples satisfy some nice property $F(p_S) \leq \xi$, where p_S is the empirical distribution.

Although our formulation only allows the adversary to add points, for all the tasks we consider, the property $F(p_S) \leq \xi$ is stable under deletions. For instance, an ϵ-deletion of a bounded-covariance distribution also has small covariance. Therefore, our results also apply to the total variation setting (additions and deletions) without loss of generality: if S^* is the original set of all good points, and S is the remaining set after deletions, then $F(p_S)$ is small whenever $F(p_{S^*})$ is small. This point is shown in more detail in Steinhardt et al. (2018); Zhu et al. (2019) in the form of a *generalized resilience* property.

Throughout the paper, we only impose assumptions on the true empirical distribution p_S instead of the population distribution. However, for all of the assumptions we consider, Zhu et al. (2019) show that if they hold for the population distribution then they also hold in finite samples for large enough n. The deterministic finite-sample setting frees us from probabilistic considerations and lets us focus on the deterministic optimization problem, and directly implies results in the statistical setting via the aforementioned generalization bounds.

2 Mean estimation: a landscape theory

In this section, we study the landscape of the optimization problem induced by mean estimation. We first show that any first-order stationary point is an approximate global minimum, and then show that it suffices to have an approximate first-order stationary point to guarantee approximate global minimum.

We start by analyzing mean estimation with bounded covariance (Example 1.1). We consider the optimization problem of minimizing $\|\Sigma_q\|$ subject to $q \in \Delta_{n,\epsilon}$. Since we have the representation $\|\Sigma_q\|_2 = \sup_{v\in \boldsymbol{R}^d, \|v\|_2=1} \sum_{i=1}^n q_i (v^\top X_i)^2 - (v^\top \mu_q)^2$, the optimization problem can be formulated as

$$\begin{aligned} &\min_q \sup_{v\in \boldsymbol{R}^d, \|v\|_2=1} \sum_{i=1}^n q_i (v^\top X_i)^2 - (v^\top \mu_q)^2 \\ &\text{s.t. } q \in \Delta_{n,\epsilon} \end{aligned} \tag{10}$$

While $\Delta_{n,\epsilon}$ is a convex set, the objective function is non-convex. In the following sections, we show that any stationary point is an approximate global minimum for this non-convex problem if and only if the corruption level $\epsilon < 1/3$. We further show that approximate stationary points are also approximate global minima, with slightly worse breakdown point.

2.1 Stationary points are approximate global minimum

We first recall the definition of first-order stationary points for locally Lipschitz functions (Clarke, 1990, Proposition 2.4.3 and Corollary) (Bian and Chen, 2017; Lacoste-Julien, 2016).

Definition 2.1 (First-order stationary points)**.** *Consider the constrained optimization problem* $\min_{x\in A} F(x)$*, where* A *is a closed convex set and* $F(\cdot) : B \mapsto \boldsymbol{R}$ *is a locally-Lipschitz function with domain* $B \supset A$*. We say that* $x \in A$ *is a first-order stationary point if there exists* $g \in \partial F(x)$ *such that*

$$\langle g, x - y\rangle \le 0, \forall y \in A, \tag{11}$$

where $\partial F(x)$ *is the Clarke subdifferential of the function* $F(x)$ *on* B*, which is defined in Definition A.1*

We interpret this definition for the minimization problem (10):

Lemma 2.1. *If* q *is a first-order stationary point of* (10)*, then for any* $p \in \Delta_{n,\epsilon}$*, there exists some* $v \in \mathbf{R}^d, \|v\|_2 = 1$ *such that*

$$\mathbb{E}_q[(v^\top (X - \mu_q))^2] \le \mathbb{E}_p[(v^\top (X - \mu_q))^2]. \tag{12}$$

Moreover, v *is a principal eigenvector of* Σ_q*:* $v \in \arg\max_{\|v\|_2} \mathbb{E}_q[(v^\top (X - \mu_q))^2]$.

Proof. Let $F(q) = \sup_{v\in \boldsymbol{R}^d, \|v\|_2=1} \sum_{i=1}^n q_i (v^\top X_i)^2 - (v^\top \mu_q)^2$ be defined on $\boldsymbol{R}^n$. From Danskins formula (Clarke, 2013, Theorem 10.22), we know that the subdifferential of $F(q)$ with respect to q_i is

$$\partial_{q_i} F(q) = (X_i - \mu_q)^\top V (X_i - \mu_q) - \mu_q^\top V \mu_q, \tag{13}$$

where $V = \sum_i \alpha_i v_i v_i^\top$ is a convex combination of supremum-achieving $v_i \in \arg\max_{\|v\|_2=1} \sum_{i=1}^n q_i (v^\top (X_i - \mu_q))^2$. By taking $x = q$, $y = p$ in (11), we have

$$\mathbb{E}_q[(X-\mu_q)^\top V (X-\mu_q)] \le \mathbb{E}_p[(X-\mu_q)^\top V (X-\mu_q)]. \tag{14}$$

Since the equality holds for a combination of v_i, it must hold for some single v_i that maximizes $\mathbb{E}_q[(v^\top (X-\mu_q))^2]$. Thus we derive the conclusion. □

Define p_* as the global minimum of the optimization problem (10). By taking $p = p_*$ in Equation (12), we know that $\|\Sigma_q\| = \mathbb{E}_q[(v^\top (X-\mu_q))^2] \le \mathbb{E}_{p_*}[(v^\top (X-\mu_q))^2]$. With this condition, we show in the following theorem that any first-order stationary point for the minimization problem (10) is an approximate global minimum.

Theorem 2.1 (Stationary points are approximate global minimum)**.** *Assume $\epsilon \in [0, 1/3)$. Then for any $q \in \Delta_{n,\epsilon}$ that satisfies (12), we have*

$$\|\Sigma_q\| \le \left(\frac{1-\epsilon}{1-3\epsilon}\right)^2 \|\Sigma_{p_*}\|, \tag{15}$$

which is tight in that there exists a first-order stationary point $q \in \Delta_{n,\epsilon}$ such that $\|\Sigma_q\| = \left(\frac{1-\epsilon}{1-3\epsilon}\right)^2 \|\Sigma_{p_*}\|$ for some set of observations $X_1, X_2, \ldots, X_n$.

Proof of Theorem 2.1. For the supremum achieving v chosen in Lemma 2.1, we have

$$\begin{aligned}
\|\Sigma_q\| &= \mathbb{E}_q[(v^\top (X-\mu_q)^2)] && (16)\\
&\overset{(i)}{\le} \mathbb{E}_{p_*}[(v^\top (X-\mu_q)^2)] && (17)\\
&= \mathbb{E}_{p_*}[(v^\top (X-\mu_{p_*})^2) + (v^\top (\mu_q - \mu_{p_*}))^2] && (18)\\
&\overset{(ii)}{\le} \sup_{v\in\mathbf{R}^d, \|v\|_2\le 1} \mathbb{E}_{p_*}[(v^\top (X-\mu_{p_*})^2)] + \sup_{v\in\mathbf{R}^d, \|v\|_2\le 1} (v^\top (\mu_q - \mu_{p_*}))^2 && (19)\\
&= \|\Sigma_{p_*}\| + \|\mu_q - \mu_{p_*}\|^2. && (20)
\end{aligned}$$

Here (i) comes from Lemma 2.1, (ii) comes from substituting v with the largest unit-norm vector. To bound $\|\mu_q - \mu_{p_*}\|$, we introduce the following two lemmas. The first lemma upper bounds $\|\mu_q - \mu_{p^*}\|_2$ in terms of $\mathsf{TV}(q, p^*)$, while the second establishes that $\mathsf{TV}(q, p^*)$ is small. These types of results are standard in literature (Li, 2019, Lemma 2.1, Lecture 4), (Zhu et al., 2019, Lemma C.2), (Diakonikolas et al., 2017; Steinhardt, 2018; Dong et al., 2019). Here we provide the tight results that improve over the existing results:

Lemma 2.2. *For any distributions p, q with $\mathsf{TV}(p, q) \le \epsilon$, we have*

$$\|\mu_p - \mu_q\| \le \sqrt{\frac{\|\Sigma_p\|\epsilon}{1-\epsilon}} + \sqrt{\frac{\|\Sigma_q\|\epsilon}{1-\epsilon}}. \tag{21}$$

Lemma 2.3. *For a distribution p_n, suppose that $q \in \Delta_{n,\epsilon_1}$ and $q' \in \Delta_{n,\epsilon_2}$. Then,*

$$\mathsf{TV}(q, q') \le \frac{\max\{\epsilon_2, \epsilon_1\}}{1 - \min\{\epsilon_2, \epsilon_1\}}. \tag{22}$$

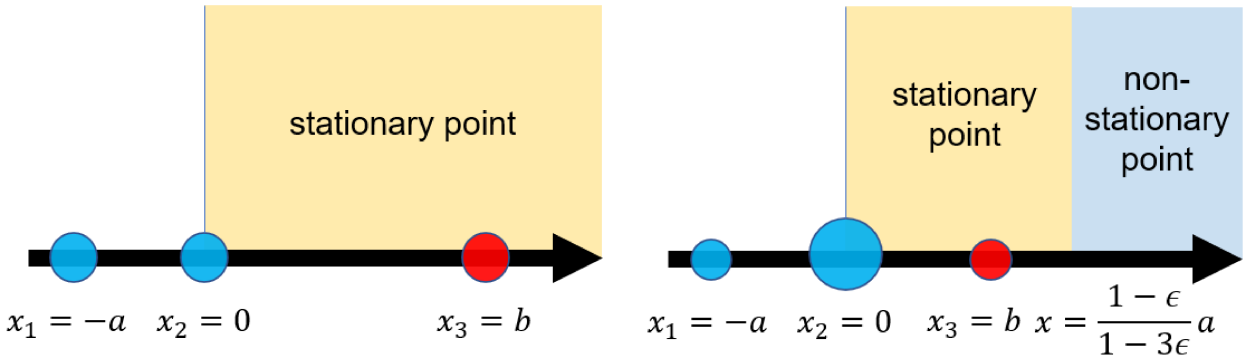

Figure 1: Illustration of the stationary points under different corruption level. The blue points are probability mass on p_* (or p_S) and the red point is added by adversary. Both x_1 and x_3 have mass ϵ and x_2 has mass $1-2\epsilon$ in corrupted distribution p. Left: When $\epsilon = 1/3$, the three points share equal probability. As long as $a > 0$, deleting x_1 and keeping the rest two points will yield a valid stationary point (in fact, a local minimum). Thus the adversary can drive the mean to infinity. Right: When $\epsilon < 1/3$, deleting x_1 only yields a valid stationary point when $b < \frac{1-\epsilon}{1-3\epsilon} \cdot a$. Thus the adversary cannot create stationary points far from mean.

The proofs are deferred to Appendices C.3 and C.2. Since both q and p_* are ϵ-deletions of corrupted distribution p_n, from Lemma 2.3 we have $\mathsf{TV}(q, p_*) \le \frac{\epsilon}{1-\epsilon}$. Combining it with Lemma 2.2, we have

$$\|\Sigma_q\| \le \|\Sigma_{p_*}\| + \left(\sqrt{\frac{\|\Sigma_q\|\epsilon}{1-2\epsilon}} + \sqrt{\frac{\|\Sigma_{p_*}\|\epsilon}{1-2\epsilon}}\right)^2. \tag{23}$$

Solving the above inequality on $\|\Sigma_q\|$, we know that when $\epsilon \in [0, 1/3)$,

$$\|\Sigma_q\| \le \left(\frac{1-\epsilon}{1-3\epsilon}\right)^2 \|\Sigma_{p_*}\| \tag{24}$$

We defer the result of tightness of (24) to Appendix C.4, and illustrate the example in Figure 1. □

Since we always have $\|\Sigma_{p_*}\| \le \|\Sigma_{p_S}\|$, Equation (15) also implies[2] $\|\Sigma_q\| \le \left(\frac{1-\epsilon}{1-3\epsilon}\right)^2 \|\Sigma_{p_S}\|$. The approximation ratio on covariance matrix is exactly tight, which indicates that a $1/3$ corruption level is tight for stationary points to be approximate global minimum: when $\epsilon \ge 1/3$, there exist cases when a local minimum is arbitrarily far from the true mean. We illustrate this phenomenon in Figure 1, and provide a formal analysis in Appendix C.4.

As a direct corollary of Theorem 2.1 and Lemma 2.2, since $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$ and both q and p_S have bounded covariance, we can bound the distance between the two means.

Corollary 2.1. *For any $q \in \Delta_{n,\epsilon}$ that satisfies (12), for the true distribution p_S we have*

$$\|\mu_q - \mu_{p_S}\|_2 \le \sigma\sqrt{\frac{\epsilon}{1-2\epsilon}} + \sigma\sqrt{\frac{\epsilon(1-\epsilon)^2}{(1-3\epsilon)^2(1-2\epsilon)}} = O\left(\sigma\sqrt{\frac{\epsilon}{1-2\epsilon}} \cdot \frac{1-\epsilon}{1-3\epsilon}\right). \tag{25}$$

[2] In fact, p_* and p_S are interchangeable throughout the section, i.e. for all the results that relate q and p_*, it is also true for q and p_S.

This corollary shows that any first-order stationary point of the optimization problem is a near-optimal estimator for mean under bounded covariance assumption up to the ratio $(1-\epsilon)/(1-3\epsilon)$, since $\sigma\sqrt{\frac{\epsilon}{1-2\epsilon}}$ is the information theoretic limit in mean estimation under the bounded covariance assumption up to universal constants for all $\epsilon \in (0, 1/2)$ (see Donoho and Liu (1988); Zhu et al. (2019) and Lemma 2.2).

2.2 Approximate stationary points are approximate global minimum

In the previous section, we show that any stationary point is an approximate global minimum. However, in practice we cannot find an exact stationary point, but only an approximate stationary point. In this section, we show that even when (12) only holds approximately, q is still an approximate global minimum. This proves to be important for generalization to other tasks in Section 3 and algorithm design in Section 4.

Concretely, we relax the condition (12) to the following for $\alpha, \beta \geq 0$:

$$\mathbb{E}_q[(v^\top(X-\mu_q))^2] \leq (1+\alpha)\mathbb{E}_{p_*}[(v^\top(X-\mu_q))^2] + \beta, \tag{26}$$

where v is a principal eigenvector of Σ_q: $v \in \arg\max_{\|v\|_2} \mathbb{E}_q[(v^\top(X-\mu_q))^2]$. Compared with (12), we only require the relationship holds for the global minimum p_* instead of all p, and allow both multiplicative error $\alpha\mathbb{E}_{p_*}[(v^\top(X-\mu_q))^2]$ and additive error β.

In fact, equation (26) is more general than the traditional definition of approximate stationary point where one requires that $\|\partial_q F(q)\|$ small (Dutta et al., 2013; Cheng et al., 2020). We show that as long as $\|\partial_q F(q)\|$ is upper bounded by some $\gamma \geq 0$, the condition (26) holds with $\alpha = 0$, $\beta = O(\epsilon\gamma)$.

Assume for simplicity that v is the only principal eigenvector of Σ_q. Then we have $\partial_{q_i} F(q) = (v^\top(X_i-\mu_q))^2 - (v^\top\mu_q)^2$, and

$$\begin{aligned}
&\mathbb{E}_q[(v^\top(X-\mu_q))^2] - \mathbb{E}_{p_*}[(v^\top(X-\mu_q))^2] \\
=&\mathbb{E}_q[(v^\top(X-\mu_q))^2 - (v^\top\mu_q)^2] - \mathbb{E}_{p_*}[(v^\top(X-\mu_q))^2 - (v^\top\mu_q)^2] \\
=&\sum_{i\in[n]} (q_i - p_{*,i}) \cdot \partial_{q_i} F(q) \\
\overset{(i)}{\leq}&\sqrt{\sum_{i\in[n]} (q_i - p_{*,i})^2} \cdot \|\partial_q F(q)\| \\
\overset{(ii)}{\leq}&\sqrt{\frac{2\epsilon}{(1-\epsilon)^2 n}} \cdot \|\partial_q F(q)\| \\
\overset{(iii)}{\leq}&O(\epsilon \cdot \|\partial_q F(q)\|).
\end{aligned} \tag{27}$$

Here (i) comes from Cauchy Schwarz, (ii) comes from optimizing over all $q, p_{*,i} \in \Delta_{n,\epsilon}$, (iii) comes from that $1/n \leq \epsilon \leq 1/2$. Thus any approximate stationary point with small $\|\partial_q F(q)\|$ will also satisfy (26) with $\alpha = 0, \beta = O(\epsilon \cdot \|\partial_q F(q)\|)$.

Now we show that any point that satisfies (26) is an approximate global minimum.

Theorem 2.2. *Assume $\epsilon \in [0, 1/3)$. Define p_* as the global minimum of the optimization problem (10). Assume $q \in \Delta_{n,\epsilon}$ and there exists some $v \in \arg\max_{\|v\|\leq 1} \mathbb{E}_q[(v^\top(X-\mu_q))^2]$ such that*

$$\mathbb{E}_q[(v^\top(X-\mu_q))^2] - \mathbb{E}_{p_*}[(v^\top(X-\mu_q))^2] \leq \alpha\mathbb{E}_{p_*}[(v^\top(X-\mu_q))^2] + \beta. \tag{28}$$

Then for some universal constants C_1, C_2, we have

$$\|\Sigma_q\| \le \left(1 + \frac{C_1(\alpha+\epsilon)}{(1-(3+\alpha)\epsilon)^2}\right)\|\Sigma_{p_*}\| + \frac{C_2\beta}{(1-(3+\alpha)\epsilon)^2}. \tag{29}$$

We defer the proof to Appendix C.5. For the task of mean estimation with bounded covariance, we want that $\|\Sigma_q\| \le C \cdot \|\Sigma_{p_S}\|$. This is satisfied when α is constant and $\beta \lesssim \|\Sigma_{p_S}\|$, in sacrifice of a smaller breakdown point $1/(3+\alpha)$.

2.3 Application to the case of mean estimation with near identity covariance

Beyond bounded covariance, we can make the stronger assumption that the covariance is close to the identity on all large subsets of the good data and the distribution has stronger tail bound (e.g. bounded higher moments or sub-Gaussianity). This stronger assumption yields tighter bounds (Diakonikolas et al., 2017; Zhu et al., 2019). We can adapt our previous landscape analysis to this setting as well.

Example 2.1 (Mean estimation with near identity covariance)**.** *Let $\Delta_{S,\epsilon} = \{r \mid \forall i \in [n], r_i \le \frac{p_{S,i}}{1-\epsilon}\}$ denote the set of ϵ-deletions on p_S. We assume that the true distribution p_S has near identity covariance, and its mean is stable under deletions, i.e. the following holds for any $r \in \Delta_{S,\epsilon}$:*

$$\|\mu_r - \mu_{p_S}\| \le \rho, \|\Sigma_{p_S} - I\| \le \tau.$$

Our goal is to solve the following feasibility problem for some $\tau' \ge \tau$ that may depend on ϵ,

$$\begin{aligned} &\textit{find} && q \\ &\textit{subject to} && q \in \Delta_{n,\epsilon}, \|\Sigma_q\| \le 1 + \tau'. \end{aligned} \tag{30}$$

Once we find such a q, we have the following lemma to guarantee mean recovery:

Lemma 2.4. *Under the same assumption on p_S as Example 2.1, any solution q to the feasibility problem (30) satisfies $\|\mu_p - \mu_q\| = O(\rho + \sqrt{\epsilon(\tau+\tau')} + \epsilon)$.*

We defer statement with detailed constants and the proof of the above lemma to Lemma D.1, where we provide a tighter analysis than Zhu et al. (2019, Lemma E.3).

The optimization is the same as in (10); only the assumptions on p_S are different. Applying Theorem 2.1, we therefore know that the stationary point q satisfies

$$\|\Sigma_q\|_2 \le \left(\frac{1-\epsilon}{1-3\epsilon}\right)^2 \|\Sigma_{p_*}\|_2 \le \left(\frac{1-\epsilon}{1-3\epsilon}\right)^2 \cdot (1+\tau) \le 1 + \frac{C(\tau+\epsilon)}{(1-3\epsilon)^2} \tag{31}$$

for some universal constant C. Thus we can guarantee that τ' is close to τ up to some constant when $\tau \gtrsim \epsilon$. The breakdown point $1/3$ is still tight in this case. Indeed, the counterexample and argument in Figure 1 and Appendix C.4 still applies.

The result for approximate stationary points in Theorem 2.2 also applies here. We know that any approximate stationary point q that satisfies (26) will satisfy

$$\|\Sigma_q\| \le \left(1 + \frac{C_1(\alpha+\epsilon)}{(1-(3+\alpha)\epsilon)^2}\right)\|\Sigma_{p_*}\| + \frac{C_2\beta}{(1-(3+\alpha)\epsilon)^2} \le 1 + \frac{C_3(\tau+\epsilon+\alpha+\beta)}{(1-(3+\alpha)\epsilon)^2} \tag{32}$$

for some universal constants C_1, C_2, C_3.

We interpret the assumptions and the results in Example 2.1 under concrete cases as below. Assume the true population distribution is sub-Gaussian, we have $\rho = C_1 \cdot \epsilon\sqrt{\log(1/\epsilon)}, \tau = C_2 \cdot \epsilon\log(1/\epsilon)$ (Diakonikolas et al., 2016; Zhu et al., 2019; Cheng et al., 2020). Thus when $\alpha \lesssim \tau$, $\beta \lesssim \tau$, we know that τ' is close to τ up to some constant. From Lemma 2.4 we know that $\|\mu_q - \mu_{p_S}\| = O(\epsilon\sqrt{\log(1/\epsilon)})$. This improves over the bound $O(\sqrt{\epsilon})$ in Corollary 2.1 since we have imposed stronger tail bound assumption.

Furthermore, if we know that q is an approximate stationary point with $\|\partial_q F(q)\| \lesssim \log(1/\epsilon)$, from (27) we know that $\alpha = 0, \beta \lesssim \epsilon\log(1/\epsilon)$, thus we have $\|\mu_q - \mu_{p_S}\| = O(\epsilon\sqrt{\log(1/\epsilon)})$. This implies the result in the independent and concurrent work (Cheng et al., 2020, Theorem 3.2).

3 From gradient to generalized quasi-gradient

Given the success of the landscape analysis (Theorem 2.1 and 2.2) for mean estimation, a natural question is whether the stationary point story holds for other tasks, such as linear regression or joint mean and covariance estimation. We might hope that, as for mean estimation, all first-order stationary points of minimizing $F(q)$ subject to $q \in \Delta_{n,\epsilon}$ are approximate global minimum.

We show that this is in general not true. The counterexample is minimizing the hypercontractivity coefficient, which appears as part of linear regression (Zhu et al., 2019) and joint mean and covariance estimation (Kothari and Steurer, 2017). The target function $F(q)$ to minimize takes the form of $\sup_{v\in\boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top X)^4]}{\mathbb{E}_q[(v^\top X)^2]^2}$, as is defined in (6). We show in Appendix D.1 there exist first-order stationary points of minimizing $F(q)$ subject to $q \in \Delta_{n,\epsilon}$ such that its hypercontractivity coefficient is arbitrarily big.

Instead, we identify a more general property, the existence of *generalized quasi-gradients*, that allows us to handle the new tasks. This also motivates the algorithm design in Section 4, where we use the generalized quasi-gradients as the gradient passed to algorithms.

3.1 Generalized quasi-gradients and mean estimation

In this section, we interpret the result of mean estimation in Section 2 in the lens of generalized quasi-gradients.

Recall that in the minimization problem $\min_{q\in\Delta_{n,\epsilon}} F(q)$, the generalized quasi-gradients, as defined in Definition 1.1, refers to any $g(X;q)$ such that for all $p, q \in \Delta_{n,\epsilon}$[3]:

$$\mathbb{E}_q[g(X;q)] - \mathbb{E}_p[g(X;q)] \le 0 \implies F(q) \le C \cdot F(p), \tag{33}$$

here C is some constant that may depend on ϵ. We call it 'generalized' quasi-gradient since in the literature, quasi-gradient usually refers to the case when $C = 1$ in the implication (Boyd and Mutapcic, 2007; Hazan et al., 2015), which requires the function F to be quasi-convex. In the case of mean estimation, we set the generalized quasi-gradient as

$$g(X;q) = (v^\top(X-\mu_q))^2, v \in \arg\max_{\|v\|_2\le 1} \mathbb{E}_q[(v^\top(X-\mu_q))^2]. \tag{34}$$

Theorem 2.1 shows that $g(X;q)$ is a valid generalized quasi-gradients for $F(q) = \|\Sigma_q\|$ with approximation ratio $C = (1-\epsilon)^2/(1-3\epsilon)^2$. For strict generalized quasi-gradient, the condition in the left of the implication in (4) reduces to (26). Theorem 2.2 shows that $g(X;q)$ is also a valid strict

[3]Indeed, it suffices to show the implication holds for any $q \in \Delta_{n,\epsilon}$ and p_S instead of any $q, p \in \Delta_{n,\epsilon}$, since we only want to relate $F(q)$ with $F(p_S)$.

generalized quasi-gradient with parameter $C_1(\alpha,\beta) = (1 + \frac{C_3(\alpha+\epsilon)}{(1-(3+\alpha)\epsilon)^2}), C_2(\alpha,\beta) = \frac{C_4\beta}{(1-(3+\alpha)\epsilon)^2}$ for some universal constant C_3, C_4.

Once we identify a generalized quasi-gradient g for $F(q)$, it suffices to find some q such that $\mathbb{E}_q[g] \le \mathbb{E}_{p_S}[g]$ to guarantee that $F(q)$ is bounded by $C \cdot F(p_S)$. However, the condition $\mathbb{E}_q[g] \le \mathbb{E}_{p_S}[g]$ is impossible to check since we do not know the true distribution p_S, why is this (strict) generalized quasi-gradient still important?

First, the conditions in strict generalized quasi-gradient can be approached via low-regret algorithms (Arora et al., 2012; Nesterov, 2009). By viewing $g(X;q)$ as loss, the low-regret algorithms usually provide the guarantee in the form of $\langle g, q - p_S\rangle \le \alpha\langle |g|, p_S\rangle + \beta$. Thus if we pick g as a strict generalized quasi-gradient, we know that the output of the low-regret algorithm will guarantee an upper bound on $F(q)$. This is the key idea for algorithm design in Section 4.

Second, as we have shown in (27), the condition in strict generalized quasi-gradient is more general than traditional approximate stationary points. Furthermore, as we show throughout this section, stationary point fails to provide a good solution to the optimization problem in general. In the meantime the viewpoint of generalized quasi-gradients enables the flexibility of selecting g, and thus succeeds in all the tasks we considered.

Third, the termination condition can be based on the function value $F(q)$. As a concrete example, Theorem 2.2 indicates that $\|\Sigma_q\|$ will be small when we reach the condition $\langle g, q - p_S\rangle \le \alpha\langle |g|, p_S\rangle + \beta$. Thus in practice it suffices for us to check $\|\Sigma_q\|$ for termination.

Now we show that for the tasks of linear regression and joint mean and covariance estimation, we can identify the generalized quasi-gradients, which enable algorithm design for these tasks in Section 4.

3.2 Linear regression

To demonstrate the power of generalized quasi-gradients, we first consider robust linear regression:

Example 3.1 (Linear regression)**.** *We assume that the true distribution p_S satisfies $F_1(p_S) \le \kappa^2, F_2(p_S) \le \sigma^2$ for $F_1(q) = \sup_{v\in\boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top X)^4]}{\mathbb{E}_q[(v^\top X)^2]^2}, F_2(q) = \frac{\mathbb{E}_q[(Y-\theta(q)^\top X)^2(v^\top X)^2]}{\mathbb{E}_q[(v^\top X)^2]}$ in (6) and (8). Our goal is to solve the following feasibility problem for some $\kappa' \ge \kappa, \sigma' \ge \sigma$ that may depend on ϵ,*

$$\begin{aligned} &\textit{find} && q \\ &\textit{subject to} && q \in \Delta_{n,\epsilon}, F_1(q) \le \kappa'^2, F_2(q) \le \sigma'^2. \end{aligned} \tag{35}$$

As is shown in Zhu et al. (2019, Example 3.2), any q that satisfies the above condition would guarantee a small worst-case excess predictive loss (regression error) of $\Theta((\kappa+\kappa')(\sigma+\sigma')\epsilon)$.

The linear regression problem is special in that we need to guarantee both F_1, F_2 small simultaneously. In fact, we can do it sequentially: we first solve (35) without the constraint that $F_2(q) \le \sigma'^2$, then we treat the output distribution as a new 'corrupted' distribution and further delete it such that $F_2(q)$ is small. The hypercontractivity is approximately closed under deletion (Zhu et al., 2020b, Lemma C.6). Thus the distribution will be hypercontractive throughout the second step.

Now we will design two generalized quasi-gradients g_1, g_2 separately. We would like to find g_1, g_2 such that

- g_1 is a generalized quasi-gradient for F_1, i.e. $\mathbb{E}_q[g_1] \le \mathbb{E}_{p_S}[g_1]$ implies $F_1(q) \le CF_1(p_S)$ for some C that may depend on ϵ.

- under the assumption of hypercontractive on the corrupted distribution p_n (hence on any deletion of it), g_2 is a generalized quasi-gradient for F_2, i.e. $\mathbb{E}_q[g_2] \le \mathbb{E}_{p_S}[g_2]$ implies $F_2(q) \le CF_2(p_S)$ for some C that may depend on ϵ.

For hypercontractivity, we take

$$g_1(X;q) = (v^\top X)^4, \text{ where } v \in \arg\max_{v \in \boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top X)^4]}{\mathbb{E}_q[(v^\top X)^2]^2}. \tag{36}$$

For bounded noise, we use the generalized quasi-gradient

$$g_2(X;q) = (Y - X^\top \theta(q))^2 (v^\top X)^2, \text{ where } v \in \arg\max_{v \in \boldsymbol{R}^d} \frac{\mathbb{E}_q[(Y - X^\top \theta(q))^2 (v^\top X)^2]}{\mathbb{E}_q[(v^\top X)^2]}. \tag{37}$$

It is in general computationally hard to solve the maximization problem in (36) related to higher moments. To guarantee the efficiency of computing g, we make a stronger assumption of hypercontracitivity under Sum-Of-Squares (SOS) proof. We refer to Appendix A for the definition of SOS proofs. To be precise, we make the hypercontractivity condition stronger in the sense that there exists a sum-of-squares proof for the inequality $\mathbb{E}_q[(v^\top X)^4] \le \kappa^2 \mathbb{E}_q[(v^\top X)^2]^2$, in particular, we let

$$\tilde{F}_1(q) = \sup_{E_v \in \mathcal{E}_4} \frac{E_v[\mathbb{E}_q[(v^\top X)^4]]}{E_v[\mathbb{E}_q[(v^\top X)^2]^2]}, \tag{38}$$

and assume that $\tilde{F}_1(p_S) \le \kappa^2$. Here $\mathcal{E}_4$ is the set of all degree-4 pseudo-expectations on the sphere, and E_v denotes one of the pseudo-expectation with respect to the polynomials of v. We provide its concrete definition in Definition A.4. We call the inequality $\tilde{F}_1(q) \le \kappa^2$ *certifiable hypercontractivity*. In this case, we would like to find some distribution q that is also certifiably hypercontractive. To guarantee certifiable hypercontractivity, We take $\tilde{g}_1$ as the pseudo-expectation of $(v^\top X)^4$, i.e.

$$\tilde{g}_1(X;q) = E_v[(v^\top X)^4], \text{ where } E_v \in \arg\max_{E_v \in \mathcal{E}_4} \frac{E_v[\mathbb{E}_q[(v^\top X)^4]]}{E_v[\mathbb{E}_q[(v^\top X)^2]^2]}. \tag{39}$$

Now we are ready to show the main result: $\tilde{g}_1, g_2$ as selected above are generalized quasi-gradients for $\tilde{F}_1, F_2$.

Theorem 3.1 (Generalized quasi-gradients for linear regression)**.** *Assume that $\tilde{F}_1(p_S) \le \kappa^2, F_2(p_S) \le \sigma^2$. For any $q \in \Delta_{n,\epsilon}$, when $\kappa^3\epsilon < 1/64$, the following implications are true:*

$$\mathbb{E}_q[\tilde{g}_1(X;q)] \le \mathbb{E}_{p_S}[\tilde{g}_1(X;q)] \Rightarrow \tilde{F}_1(q) \le 4\tilde{F}_1(p_S),$$
$$\tilde{F}_1(q) \le 4\kappa^2, \mathbb{E}_q[g_2(X;q)] \le \mathbb{E}_{p_S}[g_2(X;q)] \Rightarrow F_2(q) \le 3F_2(p_S).$$

Thus $\tilde{g}_1$ is a generalized quasi-gradients for $\tilde{F}_1$. Given that q is certifiably hypercontractive, g_2 is a generalized quasi-gradients for F_2.

We defer the proof to Appendix D.3. Theorem 3.1 shows that as long as we can find some q_1 that are approximate global minimum for F_1, and further delete it to get q_2 as a global minimum for F_2, we are guaranteed to have a q_2 such that both $\tilde{F}_1(q_2)$ and $F_2(q_2)$ are small enough, which leads to a near-optimal worst-case regression error[4].

[4]Although in the second part of further deleting q_1, $\tilde{F}_1(q_2)$ is guaranteed to be upper bounded by not $4\kappa^2$ but $C \cdot \kappa^2$ for some C that is larger than 4 and may depend on κ, ϵ, we can still achieve the guarantee $F_2(q_2) \le 3F_2(p_S)$ by substituting the assumption $\kappa^3\epsilon < 1/64$ with a tighter assumption.

Similarly to mean estimation, $\tilde{g}_1, g_2$ are also strict generalized quasi-gradients. We leave the detailed analysis to Section 4.4.

We sketch the proof of g_1 in (36) being the generalized quasi-gradients of F_1 in (6) as below for the intuition of choosing generalized quasi-gradients.

Proof sketch of quasi-graident property for g_1. We would like to show that for g_1 in (36) and $F_1(q) = \sup_{v\in\boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top X)^4]}{\mathbb{E}_q[(v^\top X)^2]^2}$, we have the following holds for any $p, q \in \Delta_{n,\epsilon}$

$$\forall q, p \in \Delta_{n,\epsilon}, \mathbb{E}_q[g_1(X;q)] \le \mathbb{E}_p[g_1(X;q)] \Rightarrow F_1(q) \le C \cdot F_1(p). \tag{40}$$

Define $\kappa^2 = F_1(p_S)$ and $\kappa'^2 = F_1(q)$. Assume that $\kappa' \ge \kappa$, since otherwise we already have the desired result. We know from Lemma 2.3 that $\mathsf{TV}(p,q) \le \frac{\epsilon}{1-\epsilon}$. From Zhu et al. (2019, Lemma C.2) (or Lemma D.3), we know that the second moments of the two hypercontractive distributions are multiplicatively close, with the ratio depending on κ', i.e.

$$\mathbb{E}_q[(v^\top X)^2]^2 \le \gamma^2 \mathbb{E}_{p_S}[(v^\top X)^2]^2, \tag{41}$$

where $\gamma^2 = \frac{(1+\sqrt{\epsilon\kappa^2/(1-2\epsilon)^2})^2}{(1-\sqrt{\epsilon\kappa^2/(1-2\epsilon)^2})^2}$. Thus we know that for the supremum achieving v picked in g_1,

$$\begin{aligned}
\mathbb{E}_q[(v^\top X)^4] &\le \mathbb{E}_{p_S}[(v^\top X)^4] \\
&\overset{(i)}{\le} \kappa^2 \mathbb{E}_{p_S}[(v^\top X)^2)]^2 \\
&\overset{(ii)}{\le} \gamma^2\kappa^2 \mathbb{E}_q[(v^\top X)^2)]^2 \qquad (42)\\
&\overset{(iii)}{=} \frac{\gamma^2\kappa^2}{\kappa'^2} \mathbb{E}_q[(v^\top X)^4)]. \qquad (43)
\end{aligned}$$

Here (i) comes from the assumption that $F_1(p_S) \le \kappa^2$, (ii) comes from (41), (iii) comes from that $\kappa'^2 = F_1(q)$. By solving the above inequality on κ', we have when $9\epsilon\kappa^2/(1-2\epsilon)^2 \le 1$,

$$\kappa' \le 2\kappa. \tag{44}$$

□

3.3 Joint mean and covariance estimation

We summarize the setting and target for joint mean and covariance estimation as below.

Example 3.2 (Joint mean and covariance estimation)**.** *We assume that the true distribution p_S satisfies $F(p_S) \le \kappa^2$ for $F(q) = \sup_{v\in\boldsymbol{R}^d} \frac{\mathbb{E}_q[(v^\top (X-\mu_q))^4]}{\mathbb{E}_q[(v^\top (X-\mu_q))^2]^2}$. Our goal is to solve the following feasibility problem for some $\kappa' \ge \kappa$ that may depend on ϵ:*

$$\begin{aligned}
&\textit{find} && q \\
&\textit{subject to} && q \in \Delta_{n,\epsilon}, F(q) \le \kappa'^2.
\end{aligned} \tag{45}$$

As is shown in Zhu et al. (2019, Example 3.3), any q that satisfies the above condition will guarantee good recovery of mean and covariance: $\|\Sigma_{p_S}^{-1/2}(\mu_q - \mu_{p_S})\| \le \Theta(\sqrt{(\kappa+\kappa')}\epsilon^{3/4})$, and $\|I_d - \Sigma_{p_S}^{-1/2}\Sigma_q\Sigma_{p_S}^{-1/2}\|_2 \le \Theta((\kappa+\kappa')\sqrt{\epsilon})$.

Different from the recovery metric in traditional mean estimation and covariance estimation, here we use the metric in transformed space as in Kothari and Steurer (2017). For mean the metric is also known as the Mahalanobis distance.

For the efficiency of computing generalized quasi-gradient, we further restrict the assumption by making it SOS certifiable, as the case for linear regression in Section 3.2. We take $\tilde{F}(q)$ as the coefficient for *certifiable* hypercontractivity:

$$\tilde{F}(q) = \sup_{E_v \in \mathcal{E}_4} \frac{E_v[\mathbb{E}_q[(v^\top (X - \mu_q))^4]]}{E_v[\mathbb{E}_q[(v^\top (X - \mu_q))^2]^2]}. \tag{46}$$

Making certifiably hypercontractive assumption on empirical distribution instead of population distribution is still reasonable. Kothari and Steurer (2017, Lemma 5.5) shows that when the population distribution is certifiably hypercontractive with parameter κ and the sample size $n \gtrsim \frac{(d \log(d/\delta))^2}{\epsilon^2}$, the empirical distribution is certifiably hypercontractive with parameter $\kappa + 4\epsilon$ with probability at least $1 - \delta$.

The assumptions and target are very similar to the case of hypercontractivity in linear regression, with the only difference that joint mean and covariance estimation has all moments centered while hypercontractivity has all moments non-centered. Indeed, the choice of quasi-gradient is also following a similar route as the case of linear regression. We identify the generalized quasi-gradient g as

$$g(X; q) = E_v[(v^\top (X - \mu_q))^4], \text{ where } E_v \in \arg\max_{E_v \in \mathcal{E}_4} \frac{E_v[\mathbb{E}_q[(v^\top (X - \mu_q))^4]]}{E_v[\mathbb{E}_q[(v^\top (X - \mu_q))^2]^2]}. \tag{47}$$

Here $\mathcal{E}_4$ is the set of all degree-4 pseudo-expectations on the sphere, and E_v denotes one of the pseudo-expectation with respect to the polynomials of v. We show that g is a generalized quasi-gradient for F in the following theorem.

Theorem 3.2 (Generalized quasi-gradients for joint mean and covariance estimation)**.** *Assume that $\tilde{F}(p_S) \le \kappa^2$. For any $q \in \Delta_{n,\epsilon}$, when $\kappa^2 \epsilon < 1/200$, the following implication holds:*

$$\mathbb{E}_q[g(X; q)] \le \mathbb{E}_{p_S}[g(X; q)] \Rightarrow F(q) \le 7F(p_S). \tag{48}$$

Thus g is a generalized quasi-gradient for F.

We defer the proof to Appendix D.4. Similar to the case of hypercontractivity in linear regression, we can extend it to the case of approximate quasi-gradient. We defer the detailed analysis to Section 4.5.

4 Designing gradient descent algorithms

From Section 2 and 3, we know that we can approximately solve the minimization problem $\min_{q \in \Delta_{n,\epsilon}} F(q)$ as long as we identify (strict) generalized quasi-gradients g and some $q \in \Delta_{n,\epsilon}$ that satisfies

$$\mathbb{E}_{X \sim q}[g(X; q)] \le \mathbb{E}_{p_S}[g(X; q)] + \alpha \mathbb{E}_{p_S}[|g(X; q)|] + \beta \tag{49}$$

for small enough α, β. In this section, we design algorithms to find such q points given strict generalized quasi-gradients. This yields efficient algorithms for all the examples discussed above.

Any low-regret online learning algorithm will eventually provide us with a q that satisfies (49). To see this, for a sequence of iterates $q_1, \ldots, q_T$, define the loss function $\ell_t(p) = \mathbb{E}_{X\sim p}[g(X; q_t)]$. Then the average regret relative to the distribution p_S is

$$\text{Regret}(p_S) = \frac{1}{T}\sum_{t=1}^{T} \ell_t(q_t) - \ell_t(p_S) \tag{50}$$

$$= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{q_t}[g(X; q_t)] - \mathbb{E}_{p_S}[g(X; q_t)] \tag{51}$$

As long as $\text{Regret}(p_S) \to 0$, the average value of $\mathbb{E}_{q_t}[g(X; q_t)] - \mathbb{E}_{p_S}[g(X; q_t)]$ also approaches zero, and so eventually (49) will hold for at least one of the q_t.

We make this argument more concrete by providing two low-regret algorithms based on gradient descent. The first, which we call the *explicit low-regret* algorithm, takes gradient steps with respect to the quasi-gradients g, and then projects back onto the constraint set $\Delta_{n,\epsilon}$. The second algorithm, which we call the *filter* algorithm, instead projects only onto the probability simplex Δ_n. Although this is larger than the original constraint set, we can guarantee that $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$ under certain conditions including that each coordinate of the generalized quasi-gradient $g_i, i \in [n]$ is always non-negative. This bound on TV distance is all we needed to guarantee small worst-case error. The filter algorithm is commonly used in the literature for mean estimation and beyond (see e.g. Diakonikolas et al. (2017); Li (2018, 2019); Steinhardt (2018)). Our result in this paper improves over the breakdown point in the literature, and also provide new results under different settings.

We provide a general analysis of these two algorithms in Section 4.1 and 4.2. Then we apply the analysis and establish the performance guarantee of the two algorithms for mean estimation with bounded covariance. For simplicity we only show the guarantee for one of the two algorithms in each of the other tasks including linear regression, joint mean and covariance estimation, and mean estimation with near-identity covariance.

Our results provide the first near-optimal polynomial-time algorithm for linear regression under the hypercontractivity and bounded noise assumption, which improves the rate in Klivans et al. (2018) from $O(\sqrt{\epsilon})$ to $O(\epsilon)$. We also give new efficient algorithms for joint mean and covariance estimation under the same setting as Kothari and Steurer (2017). We present an explicit low-regret algorithm for the case of mean estimation with near identity covariance, which improves over the independent and concurrent work (Cheng et al., 2020) in both iteration complexity (from $O(nd^3/\epsilon)$ to $O(d/\epsilon^2)$) and breakdown point (up to 1/3).

4.1 Designing general explicit low-regret algorithm

We now describe the general framework for the explicit low-regret algorithm, which runs gradient descent using the quasigradients and projects onto the set $\Delta_{n,\epsilon}$. Pseudocode is provided in Algorithm 1.

Approaching the approximate global minimum. The algorithm has the following regret bound from Arora et al. (2012, Theorem 2.4):

Lemma 4.1 (Regret bound for explicit low-regret algorithm)**.** *Denote $g^{(k)}(X) = g(X; q^{(k)})$. In Algorithm 1, assume that $|g_i^{(k)}| \le B$ for all k and i, and $\eta^{(k)} = \eta/(2B)$, $\eta \in [0, 1]$. If the algorithm*

Algorithm 1 Explicit low-regret algorithm(p_n, ξ)

Input: corrupted empirical distribution $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, ξ.
Initialize $q_i^{(0)} = 1/n$, $i \in [n]$.
for $k = 0, 1, \ldots$ **do**
  if $F(q^{(k)}) \le \xi$ **then**
    return $q^{(k)}$
  else
    Compute $g_i^{(k)} = g(X_i; q^{(k)})$, $\tilde{q}_i^{(k+1)} = q_i^{(k)} \cdot (1 - \eta^{(k)} \cdot g_i^{(k)})$.
    Update $q^{(k+1)} = \mathsf{Proj}_{\Delta_{n,\epsilon}}^{KL}(\tilde{q}^{(k+1)})$.
  end if
end for

does not terminate before step T, we have[5]

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{q^{(k)}}[g^{(k)}(X)] - \mathbb{E}_{p_S}[g^{(k)}(X)]\right) \le \frac{\eta}{T}\sum_{k=1}^{T}\mathbb{E}_{p_S}[|g^{(k)}(X)|] + \frac{2B\epsilon}{T\eta}. \tag{52}$$

The lemma directly implies that there exists some $q^{(t_0)}$ with $t_0 \in [T]$ such that

$$\mathbb{E}_{q^{(t_0)}}[g^{(t_0)}(X)] - \mathbb{E}_{p_S}[g^{(t_0)}(X)] \le \eta\mathbb{E}_{p_S}[|g^{(t_0)}(X)|] + \frac{2B\epsilon}{T\eta}. \tag{53}$$

If $g(X; q)$ is a strict generalized quasi-gradient, then it implies that q is an approximate global minimum.

Bounding the iteration complexity To bound the iteration complexity, one may adjust the iteration number T and step size $\eta^{(k)}$ in (53) to get the desired precision. Assume that we want $F(q^{(k)}) \le \xi$ and the strict generalized quasi-gradient states that

$$\mathbb{E}_q[g(X; q)] - \mathbb{E}_{p_S}[g(X; q)] \le \eta\mathbb{E}_{p_S}[|g(X; q)|] + \beta \Rightarrow F(q) \le \xi. \tag{54}$$

(See e.g. Theorem 2.2.) Then by taking $T = 2B\epsilon/(\eta\beta)$, Equation (53) shows that we are able to find q with $F(q) \le \xi$ within $O(B\epsilon/\eta\beta)$ iterations.

Furthermore, by taking $\eta^{(k)} = \min(1/(2B), 1/\sqrt{T})$ and letting $T \to \infty$, we will eventually obtain some $q^{(t_0)}$ that approaches the approximate global minimum. We summarize this analysis in the lemma below.

Lemma 4.2 (Sufficient condition for the success of explicit low-regret algorithm)**.** *Assume that $|g(X_i; q)| \le B$ for all $q \in \Delta_{n,\epsilon}$ and $i \in [n]$, and that (54) holds for the strict generalized quasi-gradient g. Take $\eta^{(k)} = \eta/(2B), \eta \in [0, 1]$. Then within $O(B\epsilon/\eta\beta)$ iterations, Algorithm 1 will terminate and output a q such that $F(q) \le \xi, q \in \Delta_{n,\epsilon}$.*

4.2 Designing general filter algorithm

Now we introduce the filter algorithm. As noted, the only difference from the explicit low-regret algorithm is that we project onto the probability simplex Δ_n instead of the set of deleted distributions $\Delta_{n,\epsilon}$. We recall that projecting to the simplex under KL divergence is equivalent to renormalization[6]. Pseudocode is provided in Algorithm 2.

[5] In Arora et al. (2012), the last term is $\mathsf{KL}(p_S||p_n)/T\eta$. One can upper bound $\mathsf{KL}(p_S||p_n) = \log(1/(1-\epsilon))$ by 2ϵ

Algorithm 2 filter algorithm (p_n, ξ)

Input: corrupted empirical distribution $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, ξ.
Initialize $q_i^{(0)} = 1/n$, $i \in [n]$.
for $k = 0, 1, \ldots$ **do**
  if $F(q^{(k)}) \le \xi$ **then**
    return $q^{(k)}$
  else
    Compute $g_i^{(k)} = g(X_i; q^{(k)}) \ge 0$, $\tilde{q}_i^{(k+1)} = q_i^{(k)} \cdot (1 - \eta^{(k)} \cdot g_i^{(k)})$.
    Update $q^{(k+1)} = \mathsf{Proj}_{\Delta_n}^{KL}(\tilde{q}^{(k+1)}) = \tilde{q}^{(k+1)} / \sum_{i\in[n]} \tilde{q}_i^{(k+1)}$.
  end if
end for

Approaching the approximate global minimum. The filter algorithm may at first seem strange, since we project onto the larger set Δ_n and so do not satisfy the constraint $q \in \Delta_{n,\epsilon}$. The key to analyzing this algorithm is that we can guarantee that $\mathsf{TV}(q, p_S)$ is small implicitly, and this property is all we need to guarantee small worst-case estimation error (see e.g. Lemma 2.2).

To bound $\mathsf{TV}(q, p_S)$, we keep track of the unnormalized weights, denoted as $c^{(k)}$. Concretely, for any $i \in [n]$, we let $c_i^{(0)} = 1/n, c_i^{(k+1)} = c_i^{(k)}(1 - \eta^{(k)} \cdot g_i^{(k)})$. Then we always have the relationship $q_i^{(k)} = c_i^{(k)} / (\sum_i c_i^{(k)})$. We show that $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$ by establishing the following invariant:

$$\sum_{i\in S}(\frac{1}{n} - c_i^{(k)}) \le \sum_{i\in[n]/S}(\frac{1}{n} - c_i^{(k)}), \tag{55}$$

which can be interpreted as saying that we delete more probability mass from bad points $i \in [n]/S$ than from good points $i \in S$. This type of analysis based on (55) are well known in the literature (Diakonikolas et al., 2017; Li, 2018, 2019; Steinhardt, 2018). We show in the following lemma that as long as (55) holds, it is guaranteed that $\mathsf{TV}(q, p_S)$ is small:

Lemma 4.3. *If c_i satisfies (55) and $c_i \le \frac{1}{n}$, for all $i \in [n]$, then the normalized distribution $q_i = c_i / \sum_i c_i$ satisfies $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$.*

We defer the detailed statement and proof to Lemma C.1. To guarantee $c_i \le \frac{1}{n}$ always holds, we need to impose another assumption that $g(X_i; q^{(k)})$ is always non-negative.

We need to carefully design and balance the threshold ξ in the algorithm such that

1. it is small enough such that the condition $F(q^{(k)}) \le \xi$ guarantees a good bound when we terminate;

2. it is large enough such that if we do not terminate at step k (which means $F(q^{(k)}) > \xi$) and the invariance (55) holds at $c^{(k)}$, then in step k we still delete more probability mass from bad points than good points, i.e. the invariance (55) still holds at step $k+1$. Since the deleted probability mass is proportional to $q_i^{(k)} g_i^{(k)}$ for each point X_i, it suffices to check that when $F(q^{(k)}) > \xi$, we have

$$\sum_{i\in S} q_i^{(k)} g_i^{(k)} \le \frac{1}{2}\sum_{i=1}^n q_i^{(k)} g_i^{(k)}. \tag{56}$$

when $\epsilon < 1/2$.

[6] Here $\mathsf{Proj}_{\Delta_n}^{KL}(\tilde{q}^{(k+1)}) = \arg\min_{p\in\Delta_n} \sum_{i=1}^n p_i \ln(\frac{p_i}{\tilde{q}_i^{(k+1)}})$ is projecting the updated distribution $\tilde{q}^{(k+1)}$ under Kullback-Leibler divergence onto the probability simplex set Δ_n, which is equivalent to renormalizing $\tilde{q}^{(k+1)}$ by dividing $\sum_{i=1}^n \tilde{q}_i^{(k+1)}$.

Although the above argument does not provide a regret bound explicitly, the low-regret argument in Lemma 4.1 still applies here. Thus there must exists some $q^{(t_0)}$ that satisfies (53). Combined with the fact that $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$ we know that $q^{(t_0)}$ is an approximate global minimum.

Bounding the iteration complexity. To bound the iteration complexity, we choose the largest possible step size

$$\eta^{(k)} = 1/g_{\max}^{(k)}, \text{where } g_{\max}^{(k)} = \max_{i\in[n]} g_i^{(k)}, \tag{57}$$

We can easily bound the iteration complexity from the choice of $\eta^{(k)}$: since we set the mass of at least one point to zero each time, the invariance (55) implies we will terminate in $O(\epsilon n)$ steps. By the time the algorithm terminates, we must have $F(q) \le \xi$, which is the desired result.

By combining the two arguments together, we derive the sufficient condition for the success of Algorithm 2.

Lemma 4.4 (Sufficient condition guaranteeing the success of filter algorithm)**.** *In Algorithm 2, take $\eta^{(k)}$ as in (57). Assume that for all $i \in [n]$ and $k \ge 0$, we have $g_i^{(k)} \ge 0$, and that when $F(q^{(k)}) > \xi$ and the invariance (55) holds, (56) always holds. Then Algorithm 2 would output some q within $O(\epsilon n)$ iterations such that $F(q) \le \xi$.*

In the above lemma, we require that (56) holds when $F(q^{(k)}) \ge \xi$ and the invariance (55) holds. We now show that if g is a strict generalized quasi-gradient with appropriate parameters, the implication is satisfied. It suffices to check that under the invariance (55), $\sum_{i\in S} q_i^{(k)} g_i^{(k)} > \frac{1}{2}\sum_{i=1}^n q_i^{(k)} g_i^{(k)}$ implies $F(q) \le \xi$. From the invariance (55) we know that $q_i \le \frac{1-\epsilon}{1-2\epsilon} \cdot p_{S,i}$ for all $i \in S$, we have

$$\mathbb{E}_{q^{(k)}}[g(X; q^{(k)})] = \sum_{i=1}^n q_i^{(k)} g_i^{(k)} < 2\sum_{i\in S} q_i^{(k)} g_i^{(k)} \le \frac{2(1-\epsilon)}{1-2\epsilon} \cdot \mathbb{E}_{p_S}[g(X; q^{(k)})]. \tag{58}$$

If g is a strict generalized quasi-gradient, from Definition 1.1 we know that the above formula implies $F(q) \le C_1(1/(1-2\epsilon), 0)\cdot F(p_S) + C_2(1/(1-2\epsilon), 0)$. Thus as long as g is coordinate-wise non-negative and a strict generalized quasi-gradient with $\xi \ge C_1(1/(1-2\epsilon), 0) \cdot F(p_S) + C_2(1/(1-2\epsilon), 0)$, the filter algorithm would work.

In the rest of the section, we will apply Lemma 4.2 and Lemma 4.4 to all the examples in the paper.

4.3 Application to mean estimation with bounded covariance

For mean estimation with bounded covariance, we want to minimize $F(q) = \|\Sigma_q\|$. The corresponding generalized quasi-gradient is $g(X; \mu_q) = (v^\top(X - \mu_q))^2$, where $v \in \arg\max_{\|v\|\le 1} \mathbb{E}_q[(v^\top(X - \mu_q))^2]$ is any of the supremum-achieving direction (Theorem 2.2). We show that under this choice of F, g, both the explicit and filter algorithm output a near-optimal estimator for the mean.

4.3.1 Explicit low-regret algorithm

Assume that $\|\Sigma_{p_S}\| \le \sigma^2$. To apply Lemma 4.2, we need $|g(X_i; q)| \le B$. Although X_i may come from the adversary and thus $g(X_i; q) = (v^\top(X_i - \mu_q^{(k)}))^2$ can be unbounded, a standard "naive

pruning" procedure (see e.g. Li (2019, Lecture 5) and Diakonikolas et al. (2019a); Dong et al. (2019)) yields data such that all points X_i satisfy $\|X_i - \mu\| \le \sigma\sqrt{d/\alpha\epsilon}$ for some $\mu \in \boldsymbol{R}^d$ while only deleting $O(\alpha\epsilon)$ fraction of points. Thus we know that $\|X_i - X_j\| \le 2\sigma\sqrt{d/\alpha\epsilon}$ for any i, j. By re-centering the points we can get $\|X_i\| \lesssim \sigma\sqrt{d/\epsilon}$. Thus in the later argument, we assume that $(v^\top(X_i - \mu_{q^{(k)}}))^2 \le \sigma^2 d/\epsilon$ for any $X_i, q, \|v\| = 1$ after naive pruning.

After applying this pruning procedure, our low-regret algorithm yields the following guarantee for mean estimation:

Theorem 4.1. *Assume that $\|\Sigma_{p_S}\| \le \sigma^2$ and $\|X_i\| \le \sigma\sqrt{d/\epsilon}/2, \forall i \in [n]$. Take (describe F and g). For any $\gamma \in (0,1)$, instantiate Algorithm 1 with fixed step size $\gamma \cdot \frac{\epsilon}{2\sigma^2 d}$, and set*

$$\xi = \Big(\frac{2\eta + 7}{3(1 - (3+\eta)\epsilon)}\Big)^2 \cdot \sigma^2. \tag{59}$$

$\eta^{(k)} = \eta \cdot \epsilon/(2\sigma^2 d), \eta \in [0,1]$, $F(q) = \|\Sigma_q\|$, $g(X; \mu_q)$ as in (34)[7]. Then for $\epsilon \in (0, 1/(3+\eta))$, Algorithm 1 will output some $q \in \Delta_{n,\epsilon}$ within $d/(\gamma\sigma^2)$ iterations such that

$$\|\Sigma_q\| \le \xi. \tag{60}$$

The conclusion also holds if we have arbitrary initialization $q^{(0)}$ instead of uniform. The result follows directly from Lemma 4.2; see Appendix E.2 for proof. We also show there that the computational complexity within each iteration is $O(nd\log(d))$.

Combining the result with Lemma 2.2, we know the output q satisfies

$$\|\mu_q - \mu_{p_S}\| \le \sigma \cdot \sqrt{\frac{\epsilon}{1-\epsilon}} + \frac{(2\eta+7)\sigma}{3(1-(3+\eta)\epsilon)} \cdot \sqrt{\frac{\epsilon}{1-\epsilon}}. \tag{61}$$

When $\eta \to 0$, the breakdown point approaches $1/3$, which is consistent with the result in Theorem 2.1.

4.3.2 Filter algorithm

Different from explicit low-regret algorithm, pre-pruning is not required for filter algorithm. We present the guarantee for filter algorithm in the following theorem.

Theorem 4.2 (Filter algorithm achieves optimal breakdown point and near-optimal error)**.** *In Algorithm 2, take $\xi = \frac{2(1-\epsilon)}{(1-2\epsilon)^2} \cdot \sigma^2$, $\eta^{(k)}$ as in (57), $F(q) = \|\Sigma_q\|$, $g(X; \mu_q) = (v^\top(X - \mu_q))^2$, where $v \in \arg\max_{\|v\|\le 1} \mathbb{E}_q[(v^\top(X-\mu_q))^2]$ is any of the supremum-achieving direction. Then for $\epsilon \in [0, 1/2)$, Algorithm 2 will output some q within $O(\epsilon n)$ iterations such that*

$$\mathsf{TV}(q, p_S) \le \frac{\epsilon}{1-\epsilon}, \|\Sigma_q\| \le \xi. \tag{62}$$

This result follows directly from Lemma 4.4 and we defer the proof to Appendix E.3. The conclusion only holds if we have uniform initialization.

Combining the result with Lemma 2.2, we know the output q satisfies

$$\|\mu_q - \mu_{p_S}\| \le \sigma \cdot \sqrt{\frac{\epsilon}{1-\epsilon}} + \sigma \cdot \frac{\sqrt{2\epsilon}}{1-2\epsilon}. \tag{63}$$

[7]For the sake of computation efficiency, it suffices to find any v such that $\mathbb{E}_q(v^\top(X_i - \mu_{q^{(k)}}))^2 \ge 0.9\|\Sigma_q\|_2$, which can be achieved by power method within $O(\log(d))$ time. One can see from the later proof that this will only affect the final bound up to some constant factor. In the rest of the paper, all the arg max in the algorithm can be substituted with this feasibility problem on v for computational efficiency.

Breakdown points and optimality. The filter algorithm achieves near-optimal error and optimal breakdown point 1/2 (Rousseeuw and Leroy, 1987, Page 270) under good initialization, while the explicit low-regret algorithm has breakdown point 1/3 with arbitrary initialization. It may sound confusing given the negative results that the stationary point may be a bad estimate when $\epsilon \geq 1/3$ in the Section 2. Indeed, if one does not initialize properly, both approaches would definitely fail for $\epsilon \geq 1/3$, and the filter algorithm may fail even for much smaller ϵ. In some sense, the success of the filter algorithm is due to appropriate initialization at uniform distribution and the landscape. We believe that one can show the breakdown point of the explicit low-regret algorithm initialized at uniform is also 1/2 with a better analysis. Theorem 4.2 also improves over previous analyses (Steinhardt, 2018; Li, 2019) of filtering algorithms in breakdown point and obtains the sharper rate.

However, the exact rate obtained in the theorem is still not optimal. If we solve the minimum distance functional in Donoho and Liu (1988) by minimizing TV distance between q and the set of distributions with covariance bounded by σ^2, the error would be $\sqrt{\frac{8\sigma^2\epsilon}{1-2\epsilon}}$ (Zhu et al., 2019, Example 3.1), which is significantly smaller than what Theorem 2.1 and Theorem 4.2 achieve when ϵ is close to 1/2. It remains an open problem whether there exists an efficient algorithm that achieves error $O(\sqrt{\frac{\sigma^2\epsilon}{1-2\epsilon}})$ for all $\epsilon \in (0, 1/2)$.

4.4 Application to linear regression

Under the same setting as in Section 3.2, we would like to find $q \in \Delta_{n,\epsilon}$ such that $\tilde{F}_1(q) \leq \kappa'$ and $F_2(q) \leq \sigma'^2$, where κ' and σ' are parameters to be specified later. Due to computational consideration, we relax the choice of maximizer in generalized quasi-gradient in Section 3.2, and take g_1, g_2 as

$$\begin{aligned} g_1(X;q) &= E_v[(v^\top X)^4], \text{ where } E_v \in \mathcal{E}_4 \text{ satisfies } E_v[\mathbb{E}_q[(v^\top X)^4]] \geq \kappa'^2 E_v[\mathbb{E}_q[(v^\top X)^2]^2], \\ g_2(X;q) &= (Y - X^\top\theta(q))^2(v^\top X)^2, \text{ where } v \in \boldsymbol{R}^d \text{ satisfies } \mathbb{E}_q[(Y - X^\top\theta(q))^2(v^\top X)^2] \geq \sigma'^2\mathbb{E}_q[(v^\top X)^2]. \end{aligned} \tag{64}$$

To guarantee that both $\tilde{F}_1$ and F_2 small simultaneously, we will check both and update q sequentially within the filter algorithm. We summarize the algorithm in Algorithm 3, and the guarantee in Theorem 4.3[8].

Theorem 4.3 (Filter algorithm for linear regression)**.** *Under the same setting as Theorem 3.1, in Algorithm 3, take $\eta^{(k)} = 1/g_{\max}^{(k)}$, where $g_{\max}^{(k)} = \max_i g_i^{(k)}$. Take $\tilde{F}_1, F_2$ as in (38) and (8), g_1, g_2 as in (64), $\kappa'^2 = \frac{2\kappa^2}{1-2\kappa^2\epsilon}$, $\sigma'^2 = \frac{4\sigma^2(1+2\kappa'\sqrt{\epsilon(1-\epsilon)})}{(1-2\epsilon)^3-20\kappa'^3\epsilon(1-\epsilon)}$. Then Algorithm 3 will output q within $O(\epsilon n)$ iterations such that*

$$\tilde{F}_1(q) \leq \kappa'^2, F_2(q) \leq \sigma'^2, \mathsf{TV}(q, p_S) \leq \frac{\epsilon}{1-\epsilon}. \tag{65}$$

We defer the proof to Appendix E.4. Combining the theorem with the regression error bound in Example 3.1, we know that the output satisfies

$$\mathbb{E}_{p_S}[(Y - X^\top\theta(q))^2] - \mathbb{E}_{p_S}[(Y - X^\top\theta(p_S))^2] \leq O((\kappa + \kappa')(\sigma + \sigma')\epsilon). \tag{66}$$

Our result provides the first near-optimal and polynomial time algorithm for linear regression under the assumption of hypercontractivity and bounded noise, which improves the rate in Klivans et al. (2018) from $O(\sqrt{\epsilon})$ to $O(\epsilon)$.

[8] An initial version of the regression result appeared in unpublished lecture notes (Steinhardt, 2019).

Algorithm 3 filter algorithm for linear regression (p_n, κ', σ')

Input: corrupted empirical distribution $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, ξ'.
Initialize $q_i^{(0)} = 1/n$, $i \in [n]$.
for $k = 0, 1, \ldots$ **do**
  if $F_1(q^{(k)}) \geq \kappa'^2$ **then**
    Compute $g_i^{(k)} = g_1(X_i; q^{(k)})$.
  else
    if $F_2(q^{(k)}) \geq \sigma'^2$ **then**
      Compute $g_i^{(k)} = g_2(X_i; q^{(k)})$.
    else
      return $q^{(k)}$.
    end if
  end if
  Compute $\tilde{q}_i^{(k+1)} = q_i^{(k)} \cdot (1 - \eta^{(k)} \cdot g_i^{(k)})$.
  Update $q^{(k+1)} = \mathsf{Proj}_{\Delta_n}^{KL}(\tilde{q}^{(k+1)}) = \tilde{q}^{(k+1)} / \sum_{i\in[n]} \tilde{q}_i^{(k+1)}$.
end for

4.5 Application to joint mean and covariance estimation

Under the setting of joint mean and covariance estimation with sum-of-squares condition in Section 3.3, we take the generalized quasi-gradient as

$$g(X; q) = E_v[(v^\top (X - \mu_q))^4], \text{ where } E_v \in \mathcal{E}_4 \text{ satisfies } E_v[\mathbb{E}_q[(v^\top (X - \mu_q))^4]] \geq \kappa'^2 E_v[\mathbb{E}_q[(v^\top (X - \mu_q))^2]^2]. \tag{67}$$

Then the filter algorithm in Algorithm 2 will guarantee a good solution q.

Theorem 4.4 (Filter algorithm for joint mean and covariance estimation)**.** *Under the same setting as Theorem 3.2, in Algorithm 2, take* $\eta^{(k)} = 1/g_{\max}^{(k)}$, *where* $g_{\max}^{(k)} = \max_i g_i^{(k)}$. *Take* F, g *as in* (46) *and* (67), $\kappa' = 7\kappa$. *Assume* $\kappa^2\epsilon \leq 1/4$. *Then Algorithm 2 will output some* q *within* $O(\epsilon n)$ *iterations such that*

$$F(q) \leq \kappa'^2, \mathsf{TV}(q, p_S) \leq \frac{\epsilon}{1-\epsilon}. \tag{68}$$

We defer the proof to Appendix E.5. Combining the theorem with the error bound in Example 3.2, we know that the output satisfies

$$\|\Sigma_{p_S}^{-1/2}(\mu_q - \mu_{p_S})\| \leq \Theta(\sqrt{(\kappa + \kappa')}\epsilon^{3/4})$$
$$\|I_d - \Sigma_{p_S}^{-1/2}\Sigma_q\Sigma_{p_S}^{-1/2}\|_2 \leq \Theta((\kappa + \kappa')\sqrt{\epsilon}).$$

Our result provides a new efficient algorithm for joint mean and covariance estimation under the same setting as Kothari and Steurer (2017).

4.6 Application to mean estimation with near identity covariance

Under the setting of mean estimation with near identity covariance (Example 2.1), we assume the following holds for any $r \in \Delta_{S,\epsilon}$:

$$\|\mu_r - \mu_{p_S}\| \leq \rho, \|\Sigma_{p_S} - I\| \leq \tau, \tag{69}$$

and would like to find some q such that $\mathsf{TV}(q, p_S) \le \frac{\epsilon}{1-\epsilon}, F(q) = \|\Sigma_q\| \le 1 + C \cdot \tau$. It is shown in Section 2.3 that we can take the quasi-gradient g the same as the case of bounded covariance.

We present an explicit low-regret algorithm for the case of mean estimation with near identity covariance. For better bound of iteration complexity, we choose a slightly different generalized quasi-gradient g as

$$g(X; q) = (v^\top (X - \mu_q))^2 - 1, \text{ where } v \in \boldsymbol{R}^d \text{ satisfies } \mathbb{E}_q[(v^\top (X - \mu_q))^2] \ge (1 - \gamma)\|\Sigma_q\|, \quad (70)$$

where $\gamma \in (0, 1)$ is the desired precision, and v can be found via power method within $O(\log(d)/\gamma)$ time. Here we lose the property that $g(X_i; q) \ge 0$. Thus Lemma 4.4 for filter algorithm does not apply directly. However, we can still run explicit low-regret algorithm.

The explicit low-regret algorithm will still work even if we take $g(X; q) = (v^\top (X - \mu_q))^2$, since as $T \to \infty$, we can find some q that satisfies the condition $\mathbb{E}_q[(v^\top (X - \mu_q))^2] \le \mathbb{E}_{p_S}[(v^\top (X - \mu_q))^2] + \gamma$ for arbitrarily small γ. The choice of adding -1 in the generalized quasi-gradient is only due to the consideration of iteration complexity, which will be elaborated in the proof.

Since we know that p_S has bounded covariance, we assume that $\|X_i\|_2 \le \sqrt{d/\epsilon}$ after the same naive filtering method in Section 4.3.1. Then we have the following result for mean estimation.

Theorem 4.5 (Explicit low-regret algorithm for mean estimation with near identity covariance)**.** *Assume that p_S satisfies (69), and $\forall i \in [n], \|X_i\| \le \sqrt{d/\epsilon}$. In Algorithm 1, take $\eta^{(k)} = \frac{\beta\tau}{1+\tau/2} \cdot \frac{\epsilon}{8d}, \beta \in (0, 1)$, $\xi = 1 + C_1 \cdot \frac{\tau + \epsilon\rho^2 + \epsilon}{(1 - 3(1 + \beta\tau/(1-\gamma\epsilon))\epsilon)^2}$ for some universal constant C_1, $F(q) = \|\Sigma_q\|$, g as in (70). Then Algorithm 1 will output some $q \in \Delta_{n,\epsilon}$ within $O(d/\tau^2)$ iterations such that*

$$\|\Sigma_q\|_2 \le 1 + C \cdot \frac{(1 + 1/\beta)\tau + \rho^2 + \epsilon}{(1 - 3(1 + \beta\tau/(1 - \gamma\epsilon))\epsilon)^2}. \quad (71)$$

We defer the proof to Appendix E.6. Combining the result with Lemma 2.4, we know that the output satisfies

$$\|\mu_{p_S} - \mu_q\| = O(\rho + \frac{\sqrt{\epsilon((1 + 1/\beta)\tau + \rho^2 + \epsilon)}}{(1 - 3(1 + \beta\tau/(1 - \gamma\epsilon))\epsilon)}). \quad (72)$$

As $\beta, \gamma \to 0$, we can see the breakdown point of the algorithm approaches $1/3$, which is tight and consistent with the result in Section 2.3.

Application of the guarantee. The result applies to two different cases: true distribution as either sub-Gaussian distribution with identity covariance or bounded k-th moment with identity covariance.

1. If p_S is the empirical distribution from a sub-Gaussian distribution, and the sample size satisfies $n \gtrsim d/(\epsilon \log(1/\epsilon))$, then $\rho = C_1 \cdot \epsilon\sqrt{\log(1/\epsilon)}$, $\tau = C_2 \cdot \epsilon\log(1/\epsilon)$ for some univeresal constants C_1, C_2 (see e.g. Zhu et al. (2019, Lemma E.11), Diakonikolas et al. (2019a, Lemma 4.4)). Thus the algorithm will output q such that $\|\mu_q - \mu_{p_S}\|_2 \le C_3 \cdot \epsilon\sqrt{\log(1/\epsilon)}$ for some universal constant C_3.

   In this case, our result improves over the concurrent and independent work in Cheng et al. (2020) in both iteration complexity and breakdown point. The best iteration complexity in Cheng et al. (2020) is $O(nd^3/\epsilon)$, while our algorithms achieves $O(d/\tau^2)$. In the case of sub-Gaussian distributions, $\tau = C \cdot \epsilon \log(1/\epsilon)$, and $\epsilon \ge 1/n$, thus we always have $d/\tau^2 \gtrsim nd^3/\epsilon$.

2. If p_S is the empirical distribution from a distribution with bounded k-th moment, and the sample size satisfies $n \gtrsim d\log(d)/\epsilon^{2-2/k}$, then $\rho = C_1\epsilon^{1-1/k}$, $\tau = C_2\epsilon^{1-2/k}$ for some constant C_1, C_2 (see Zhu et al. (2019, Theorem 5.6)). Thus the algorithm will output q such that $\|\mu_q - \mu_{p_S}\| \le C_3\epsilon^{1-1/k}$ for some universal constants C_3.

Designing filter algorithm. For mean estimation with near identity covariance, it is well known that the filtering algorithm can work under a different choice of objective function $F(q)$, which only considers the second moment on the 2ϵ tail of the points (see e.g. (Diakonikolas et al., 2017; Li, 2018, 2019)). However, running filter algorithm using the generalized quasi-gradient in Section 2.3 would fail since it is only able to guarantee that $\|\Sigma_q\| \le C(\epsilon)\cdot\|\Sigma_{p_S}\|$ for some $C(\epsilon)$ that cannot approach 1 when ϵ is small (Theorem 4.2). Indeed, $C(\epsilon) \to 2$ as $\epsilon \to 0$. However, it is fine to have constant $C(\epsilon)$ in $\|\Sigma_q\|_2 - 1 \le C(\epsilon)\cdot|\|\Sigma_{p_S}\| - 1|$ instead. The failure of naively running filtering algorithm is also discussed in (Li, 2019, Lecture 7). From another point of view, Lemma 4.4 shows that filtering in general cannot guarantee that the constant $C(\epsilon)$ approaches 1 as $\epsilon \to 0$, so the success of the upper tail filtering algorithms in (Diakonikolas et al., 2017; Li, 2018, 2019) may be explained by constructing a new $F(q)$ such that a constant approximation ratio gives good estimation error.

5 Conclusion

In this paper, we investigate why the feasibility problem (Problem (1.1)) can be efficiently solved, which was the target of essentially all computationally efficient robust estimators in high dimensions. Based on the insights, we are able to develop new algorithms for different robust inference tasks.

We start from exploring the landscape for mean estimation. We show that any approximate stationary point is an approximate global minimum for the associated minimization problem under either bounded covariance assumption or near identity covariance assumption with stronger tail bounds.

We then generalize the insights from mean estimation to other tasks. We identify generalized quasi-gradients for different tasks. Based on the generalized quasi-gradients, we design algorithms to approach approximate global minimum for a variety of tasks, which produces efficient algorithms for mean estimation with bounded covariance which is near optimal in both rate and breakdown point, first polynomial time and near-optimal algorithm for linear regression under hypercontractivity assumption, and new efficient algorithm for joint mean and covariance estimation. Our algorithm also improves both the breakdown point and computational complexity for the task of mean estimation with near identity covariance.

Beyond the questions we investigated, the framework applies to a large family of robust inference questions, including sparse mean estimation and sparse linear regression. The following steps may be followed to deal with a new robust statistics problem: first identify the bound on the worst case error using modulus of continuity (Donoho and Liu, 1988; Zhu et al., 2019), then formulate an approximate MD problem in the form of Problem 1.1, at last, identify the efficiently computable generalized quasi-gradients for the approximate MD problem, and approach the approximate global minimum using either explicit low-regret or implicit low-regret algorithm.

References

Jorge Adrover and Víctor Yohai. Projection estimates of multivariate location. *The Annals of Statistics*, 30(6):1760–1781, 2002.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

Thorsten Bernholt. Robust estimators are hard to compute. Technical report, Technical Report, 2006.

Wei Bian and Xiaojun Chen. Optimality and complexity for constrained optimization problems with nonconvex regularization. *Mathematics of Operations Research*, 42(4):1063–1084, 2017.

Stephen Boyd and Almir Mutapcic. Subgradient methods. lecture notes of ee364b. *Standford Univ., Stanford, CA, USA, Tech. Rep*, 2007.

RW Butler, PL Davies, and M Jhun. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, pages 1385–1400, 1993.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under hubers contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Zhiqiang Chen and David E Tyler. The influence function and maximum bias of Tukey's median. *The Annals of Statistics*, 30(6):1737–1759, 2002.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019a.

Yu Cheng, Ilias Diakonikolas, Rong Ge, and David Woodruff. Faster algorithms for high-dimensional robust covariance estimation. *arXiv preprint arXiv:1906.04661*, 2019b.

Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent, 2020.

Francis Clarke. *Functional analysis, calculus of variations and optimal control*, volume 264. Springer Science & Business Media, 2013.

Frank H Clarke. A new approach to lagrange multipliers. *Mathematics of Operations Research*, 1 (2):165–174, 1976.

Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. Siam, 1990.

Laurie Davies et al. The asymptotics of rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, 20(4):1828–1843, 1992.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664. IEEE, 2016.

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, 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, Daniel Kane, Sushrut Karmalkar, Eric Price, and Alistair Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Advances in Neural Information Processing Systems*, pages 10688–10699, 2019b.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019c.

Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems*, pages 6065–6075, 2019.

David L Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston, 1982.

David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

David L Donoho and Richard C Liu. The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

Joydeep Dutta, Kalyanmoy Deb, Rupesh Tulshyan, and Ramnik Arora. Approximate kkt points and a proximity measure for termination. *Journal of Global Optimization*, 56(4):1463–1499, 2013.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

Mia Hubert and Michiel Debruyne. Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, 2(1):36–43, 2010.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Jerry Li. Robust sparse estimation tasks in high dimensions. *arXiv preprint arXiv:1702.05860*, 2017.

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

Jerry Zheng Li. Lecture notes on robustness in machine learning, 2019.

Jingbo Liu. A note on affine invariant cost functions, 2020.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 1. Wiley Online Library, 1987.

Werner A Stahel. *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch., 1981.

Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

Jacob Steinhardt. Lecture notes for stat260 (robust statistics), 2019.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, page 45. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. When does the Tukey median work? *IEEE International Symposium on Information Theory (ISIT)*, 2020a.

Banghua Zhu, Jiantao Jiao, and David Tse. Deconstructing generative adversarial networks. *IEEE Transactions on Information Theory*, 2020b.

A Omitted definitions and notations

Definition A.1 (Clarke subdifferential (Clarke, 1990, Chapter 2))**.** *We work in a Banach space X. Let Y be a subset of X. For a given function $f : Y \mapsto \boldsymbol{R}$ that is locally Lipschitz near a given point $x \in X$, let $f^\circ(x; v)$ denote its generalized directional derivative at x in the direction $v \in X$:*

$$f^\circ(x; v) = \limsup_{y \to x, t \to 0+} \frac{f(y + tv) - f(y)}{t}. \tag{73}$$

Consider the dual space X^*. The Clarke subdifferential of f at x, denoted $\partial f(x)$, is the subset of X^* given by

$$\{\xi \in X^* \mid f^\circ(x; v) \geq \langle \xi, v \rangle \text{ for all } v \in X\}. \tag{74}$$

Definition A.2 (Sum-of-squares (SOS) proof)**.** *For any two polynomial functions $p(v), q(v)$ with degree at most d, we say $p(v) \geq q(v)$ has a degree-d sum-of-squares proof if there exists some degree-$d/2$ polynomials $r_i(v)$ such that*

$$p(v) - q(v) = \sum_i r_i^2(v), \tag{75}$$

we denote it as

$$p(v) \succeq_{sos} q(v). \tag{76}$$

Definition A.3 (Certifiable k-hypercontractivity)**.** *We say a d-dimensional random variable X is certifiably k-hypercontractive with parameter κ if there exists a degree-$2k$ sum-of-squares proof for the k-hypercontractivity condition, i.e.*

$$\mathbb{E}_p[(v^\top X)^{2k}] \preceq_{\text{sos}} (\kappa \mathbb{E}_p[(v^\top X)^2])^k, \tag{77}$$

We will also need to introduce one additional piece of sum-of-squares machinery, called *pseudoexpectations on the sphere*:

Definition A.4 (pseudoexpectation on the sphere)**.** *A degree-$2k$ pseudoexpectation on the sphere is a linear map E from the space of degree-$2k$ polynomials to $\boldsymbol{R}$ satisfying the following three properties:*

- *$E[1] = 1$ (where 1 on the LHS is the constant polynomial).*
- *$E[p^2] \geq 0$ for all polynomials p of degree at most k.*
- *$E[(\|v\|^2 - 1)p] = 0$ for all polynomials p of degree at most k.*

We let $\mathcal{E}_{2k}$ denote the set of degree-$2k$ pseudoexpectations on the sphere.

The space $\mathcal{E}_{2k}$ can be optimized over efficiently, because it has a separation oracle expressible as a sum-of-squares program. Indeed, checking that $E \in \mathcal{E}_{2k}$ amounts to solving the problem $\min\{E[p] \mid p \succeq_{\text{sos}} 0\}$, which is a sum-of-squares program because $E[p]$ is a linear function of p. Throughout the paper, we use E_v to denote the pseudoexpectation with respect to v.

B Connections with classical literature

In this section, we discuss the progress of robust statistics and the connections to our paper. For problems such as mean and covariance estimation, where the loss function $L(p_S, \hat{\theta}(p_n))$ takes a special form $\|\theta(p_S) - \hat{\theta}(p_n)\|$ for some norm $\|\cdot\|$, the task of robust estimation is usually decomposed to two separate goals: bounding the *maximum bias* and being *Fisher-consistent*. The maximum bias measures $\|\hat{\theta}(p_n) - \hat{\theta}(p_S)\|$ over the worst case corruption p_n, while Fisher-consistency means that the estimator's output given the real distribution, $\hat{\theta}(p_S)$, is exactly the same as the real parameter one wants to compute given p_S: $\theta(p_S)$.

Checking Fisher-consistency may be doable, but bounding the maximum bias proves to be challenging for various estimators. (Huber, 1973; Donoho, 1982; Donoho and Gasko, 1992; Chen and Tyler, 2002; Chen et al., 2018; Zhu et al., 2020a) analyzed the maximum bias for the Tukey median, while Davies et al. (1992) analyzed that for the *minimum volume ellipsoid (MVE)*, but the maximum bias for the *minimum covariance determinant (MCD)* is still largely open (Adrover and Yohai, 2002; Hubert and Debruyne, 2010). Given the difficulty of analyzing the maximum bias, statisticians turned to surrogates of these concepts. Two popular notions are the *breakdown point* and *equivariance* property. The breakdown point is defined as the smallest corruption level ϵ such that the maximum bias is infinity. In other words, it measures the smallest level of corruption p_n does to p_S to drive the estimate $\hat{\theta}(p_n)$ to infinity. Ideally we want a large breakdown point, but this single criterion is not enough. Indeed, the constant zero estimator has breakdown point one but is completely useless. The second criteria, *equivariance* mandates that the estimator $\hat{\theta}$ has to follow similar group transformations if we transform the data. For example, in the case of mean estimation, *translation equivariance* means that $\hat{\theta}(\{X_1 + b, X_2 + b, \ldots, X_n + b\}) = \hat{\theta}(X_1, X_2, \ldots, X_n) + b$ for any vector $b \in \mathbf{R}^d$, and *affine equivariance* means that $\hat{\theta}(\{AX_1 + b, AX_2 + b, \ldots, AX_n + b\}) = A\hat{\theta}(X_1, X_2, \ldots, X_n) + b$ for any vector $b \in \mathbf{R}^d$ and nonsingular matrix A. It was shown that the the maximal breakdown point for any translation equivariant mean estimator is at most $\lfloor (n+1)/2 \rfloor / n$ (Rousseeuw and Leroy, 1987, Page 270), which as $n \to \infty$ approaches $1/2$. If we enforce affine equivariance, then the maximum breakdown point decreases to $\lfloor (n-d+1)/2 \rfloor / n$ as shown in (Rousseeuw and Leroy, 1987, Page 271), which is way below $1/2$ when n and d are comparable.

Translation equivariance for mean estimation looks natural since it is implied by Fisher-consistency for mean estimation, but why do we additionally consider affine equivariance? One observation might be that there exist translation equivariant estimators with $1/2$ breakdown point, but it fails to achieve good maximum bias. Indeed, if p_S comes from d-dimensional isotropic Gaussian and the estimator is coordinatewise median, then its maximum bias is of order $\epsilon\sqrt{d}$ while the information theoretic optimal error $O(\epsilon)$. Moreover, it was shown in (Rousseeuw and Leroy, 1987, Page 250) that it does not necessarily lie in the convex hull of the samples when $d \geq 3$. Requiring affine equivariance rules out the coordinatewise median estimator, and may be a desirable property since the Tukey median is affine equivariant. However, it is quite challenging to find estimators that are both affine equivariant and have the largest possible breakdown point. A few notable examples are the Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982), the minimum volume ellipsoid (MVE) (Davies et al., 1992), and the minimum covariance determinant (MCD) (Butler et al., 1993; Rousseeuw and Driessen, 1999; Hubert and Debruyne, 2010), which are all shown to be NP-hard to compute in the worst case in (Bernholt, 2006). It was also shown in (Davies et al., 1992) that even if we can compute MVE, its maximum bias is suboptimal.

Till today, researchers have not found any efficiently computable estimator that is both affine equivariant and has the largest possible breakdown point among affine equivariant estimators. However, the computationally efficient schemes we are discussing in this paper are solving the *original problem* of analyzing maximum bias and Fisher-consistency. It appears that once we remove

the affine-equivariance requirement, the problem becomes computationally efficiently solvable.

The most interesting connection between the classical literature and recent computationally efficient robust estimators is the MCD (Butler et al., 1993; Rousseeuw and Driessen, 1999; Hubert and Debruyne, 2010), which is defined as

$$\begin{aligned} &\text{minimize} \quad \det(\Sigma_q) && (78)\\ &\text{subject to} \quad q \in \Delta_{n,\epsilon}. && (79) \end{aligned}$$

It looks strikely similar to our example of robust mean estimation under bounded covariance:

$$\begin{aligned} &\text{find} \quad q && (80)\\ &\text{subject to} \quad q \in \Delta_{n,\epsilon}, \|\Sigma_q\| \le \sigma'^2. && (81) \end{aligned}$$

We can see that the major difference is that our problem is a feasibility problem while MCD is a minimization problem, and MCD considers the determinant but we use the operator norm. Interestingly, it was shown that among all minimization problems using the covariance matrix Σ_q, the determinant is the *only* function that guarantees affine equivariance (Liu, 2020).

C Proof for Section 2

C.1 Proof of Auxillary Lemmas

Lemma C.1. *Denote q_S as the distribution of q conditioned on the good set S, i.e.*

$$q_{S,i} = \begin{cases} \frac{q_i}{\sum_{i\in S} q_i}, & i \in S, \\ 0, & \text{otherwise}. \end{cases} \tag{82}$$

Assume that (55) holds and $\forall i \in [n], c_i \le \frac{1}{n}$, then q_S is an $\epsilon/(1-\epsilon)$-deletion of p_S, and an ϵ-deletion of q. We also have $\mathsf{TV}(q, p_S) \le \frac{\epsilon}{1-\epsilon}$.

Proof. From the update rule we have $\forall i, c_i \le \frac{1}{n}$. Furthermore, we know that

$$\sum_{i\in S}(\frac{1}{n} - c_i) \le \sum_{i\in[n]/S}(\frac{1}{n} - c_i) \le \epsilon. \tag{83}$$

Thus we have $\sum_{i\in S} c_i \ge 1 - 2\epsilon$, and

$$\forall i \in S, q_{S,i} = \frac{c_i}{\sum_{i\in S} c_i} \le \frac{1}{(1-2\epsilon)n} = \frac{1}{1-\epsilon/(1-\epsilon)} \cdot \frac{1}{(1-\epsilon)n}. \tag{84}$$

Thus we can conclude that q_S is an $\epsilon/(1-\epsilon)$ deletion of p_S. On the other hand,

$$\forall i \in S, q_{S,i} = \frac{c_i}{\sum_{i\in S} c_i} \le \frac{c_i}{\sum_{i=1}^n c_i} \cdot \frac{\sum_{i=1}^n c_i}{\sum_{i\in S} c_i} \le q_i \cdot \frac{1}{1-\epsilon}. \tag{85}$$

Now we show that $\mathsf{TV}(p_S, q) \le \frac{\epsilon}{1-\epsilon}$. We use the following formula for TV: $\mathsf{TV}(p, q) = \int \max(q(x) - p(x), 0)dx$. Let β be such that $\sum_{i=1}^n c_i = (1-\beta)n$. Then we have

$$\mathsf{TV}(p_S, q) = \sum_{i\in S} \max\big(\frac{c_i}{(1-\beta)n} - \frac{1}{(1-\epsilon)n}, 0\big) + \sum_{i\notin S} \frac{c_i}{(1-\beta)n}. \tag{86}$$

If $\beta \le \epsilon$, then the first sum is zero while the second sum is at most $\frac{\epsilon}{1-\beta} \le \frac{\epsilon}{1-\epsilon}$. If on the other hand $\beta > \epsilon$, we will instead use the equality obtained by swapping p and q, which yields

$$\mathsf{TV}(p_S, q) = \sum_{i \in S} \max\left(\frac{1}{(1-\epsilon)n} - \frac{c_i}{(1-\beta)n}, 0\right) \tag{87}$$

$$= \frac{1}{(1-\epsilon)(1-\beta)n} \sum_{i \in S} \max((1-\beta)(1-c_i) + (\epsilon - \beta)c_i, 0). \tag{88}$$

Since $(\epsilon - \beta)c_i \le 0$ and $\sum_{i \in S}(1 - c_i) \le \epsilon n$, this yields a bound of $\frac{(1-\beta)\epsilon}{(1-\epsilon)(1-\beta)} = \frac{\epsilon}{1-\epsilon}$. We thus obtain the desired bound no matter the value of β, so $\mathsf{TV}(p_S, q) \le \frac{\epsilon}{1-\epsilon}$. □

C.2 Proof of Lemma 2.3

Proof. Note that for any set A, we have

$$p(A) \le \frac{r(A)}{1-\epsilon_1} \tag{89}$$

$$q(A) \le \frac{r(A)}{1-\epsilon_2}. \tag{90}$$

Apply it to the complement of A, we have

$$p(A) \ge \frac{r(A) - \epsilon_1}{1-\epsilon_1} \tag{91}$$

$$q(A) \ge \frac{r(A) - \epsilon_2}{1-\epsilon_2}. \tag{92}$$

It then implies that

$$p(A) \le \frac{\epsilon_2 + (1-\epsilon_2)q(A)}{1-\epsilon_1} \tag{93}$$

$$q(A) \le \frac{\epsilon_1 + (1-\epsilon_1)p(A)}{1-\epsilon_2} \tag{94}$$

If $\epsilon_2 \ge \epsilon_1$, we have

$$\mathsf{TV}(p, q) = \sup_A p(A) - q(A) \tag{95}$$

$$\le \sup_A \frac{\epsilon_2 + (1-\epsilon_2)q(A)}{1-\epsilon_1} - q(A) \tag{96}$$

$$= \frac{\epsilon_2}{1-\epsilon_1} + \sup_A \frac{\epsilon_1 - \epsilon_2}{1-\epsilon_1} q(A) \tag{97}$$

$$\le \frac{\epsilon_2}{1-\epsilon_1} \tag{98}$$

$$= \frac{\max\{\epsilon_2, \epsilon_1\}}{1 - \min\{\epsilon_2, \epsilon_1\}}. \tag{99}$$

The case of $\epsilon_1 > \epsilon_2$ follows similarly by writing $\mathsf{TV}(p, q) = \sup_A q(A) - p(A)$. □

C.3 Proof of Lemma 2.2

We first prove the following lemma for the resilience property of mean estimation.

Lemma C.2 (Resilience for mean estimation)**.** *For any q and event E such that $q(E) \geq 1-\epsilon$, we have*

$$\|\mathbb{E}_q[X] - \mathbb{E}_q[X|E]\| \leq \sqrt{\|\Sigma_q\| \frac{\epsilon}{1-\epsilon}}. \tag{100}$$

Proof. For any $a \in \boldsymbol{R}^d$ and event E such that $q(E) \geq 1-\epsilon$, we have

$$\mathbb{E}_q[X] - a = \mathbb{E}_q[(X-a)\mathbb{1}(E)] + \mathbb{E}_q[(X-a)\mathbb{1}(E^c)] \tag{101}$$

For any direction $v \in \boldsymbol{R}^d, \|v\|_2 \leq 1$, by applying Hölder's inequality to $\mathbb{E}_q[v^\top (X-a)\mathbb{1}(E^c)]$ we obtain

$$\begin{aligned} v^\top(\mathbb{E}_q[X] - a) &= v^\top(\mathbb{E}_q[(X-a)\mathbb{1}(E)] + \mathbb{E}_q[(X-a)\mathbb{1}(E^c)]) && (102)\\ &\geq v^\top \mathbb{E}_q[(X-a)\mathbb{1}(E)] - \sqrt{q(E^c)}\sqrt{\mathbb{E}_q[(v^\top(X-a))^2]} \\ &= q(E)(\mathbb{E}_q[v^\top X|E] - v^\top a) - \sqrt{q(E^c)}\sqrt{\mathbb{E}_q[(v^\top(X-a))^2]} && (103) \end{aligned}$$

By taking $a = \mathbb{E}_q[X] + \sqrt{\|\Sigma_q\| \frac{1-q(E)}{q(E)}}$, we have

$$\mathbb{E}_q[v^\top X|E] - \mathbb{E}_q[v^\top X] \leq \sqrt{\|\Sigma_q\| \frac{1-q(E)}{q(E)}} \leq \sqrt{\|\Sigma_q\| \frac{\epsilon}{1-\epsilon}} \tag{104}$$

Thus we can conclude that $\|\mathbb{E}_q[X|E] - \mathbb{E}_q[X]\| \leq \sqrt{\|\Sigma_q\| \frac{\epsilon}{1-\epsilon}}$ by taking the supremum over v : $\|v\|_2 = 1$. □

From $\mathsf{TV}(p,q) \leq \epsilon$, we know that there exists some r such that $r \leq \frac{p}{1-\epsilon}$, $r \leq \frac{q}{1-\epsilon}$ (Zhu et al., 2019, Lemma C.1). Thus we have

$$\begin{aligned} \|\mathbb{E}_q[X] - \mathbb{E}_p[X]\| &\leq \|\mathbb{E}_q[X] - \mathbb{E}_r[X]\| + \|\mathbb{E}_r[X] - \mathbb{E}_p[X]\| \\ &\leq \sqrt{\|\Sigma_q\| \frac{\epsilon}{1-\epsilon}} + \sqrt{\|\Sigma_p\| \frac{\epsilon}{1-\epsilon}} && (105) \end{aligned}$$

Remark C.1. *Lemma C.2 is tight since it achieves equality for the distribution q where $q(0) = 1-\epsilon, q(a) = \epsilon$, and the set $E = \{0\}$. It improves over existing results in the literature such as (Zhu et al., 2019, Example 3.1), (Cheng et al., 2019a, Lemma 5.3), (Li, 2019, Lecture 5, Lemma 1.1).*

C.4 Discussions related to the lower bound for breakdown point

Now we show that not all the stationary points are global minimum, and provide some sufficient conditions when the distribution q is stationary point the via the following example.

Example C.1. *Let $a > 0$, and the corruption level $\epsilon = 1/n$. Let one dimensional corrupted distribution p be*

$$\mathbb{P}_p[X = x] = \begin{cases} \frac{1}{n}, & x = -1 \\ \frac{n-2}{n}, & x = 0 \\ \frac{1}{n}, & x = a. \end{cases} \tag{106}$$

Let distribution q be

$$\mathbb{P}_q[X = x] = \begin{cases} \frac{n-2}{n-1}, & x = 0 \\ \frac{1}{n-1}, & x = a. \end{cases} \tag{107}$$

Then q, μ_q is a stationary point in optimization problem (10) when $a \le \frac{n-1}{n-3}$, and is not a stationary point otherwise.

Proof. The Lagrangian for the optimization problem is

$$L(q, w, u, y, \lambda) = F(q, w) + \sum_{i=1}^{n} u_i\Big(-q_i\Big) + \sum_{i=1}^{n} y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) + \lambda\Big(\sum_{i}^{n} q_i - 1\Big).$$

From the KKT conditions for locally Lipischitz functions (Clarke, 1976), we know that the stationary points must satisfy

$$\begin{aligned} &\text{(stationarity)} \quad 0 \in \partial_{q,w}\Big(F(q, w) + \sum_{i=1}^{n} u_i q_i + \sum_{i=1}^{n} y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) + \lambda\Big(\sum_{i}^{n} q_i - 1\Big)\Big), \\ &\text{(complementary slackness)} \quad u_i(-q_i) = 0,\ y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) = 0,\ i \in [n], \\ &\text{(primal feasibility)} \quad -q_i \le 0,\ q_i - \frac{1}{(1-\epsilon)n} \le 0,\ \sum_{i}^{n} q_i = 1, \\ &\text{(dual feasibility)} \quad u_i \ge 0,\ y_i \ge 0,\ i \in [n]. \end{aligned} \tag{108, 109}$$

It suffices to check the KKT condition in (108). Denote q_1, q_2, q_3 as the probability mass on $-1, 0, a$. Then the KKT conditions are equivalent to

$$\begin{aligned} 0 &= (-1 - \mu_q)^2 - u_1 + \lambda, \\ 0 &= \mu_q^2 + y_2 + \lambda, \\ 0 &= (a - \mu_q)^2 + y_3 + \lambda, \\ u_1, &y_2, y_3 \ge 0. \end{aligned}$$

Since $a > 0$, the necessary and sufficient condition for the KKT conditions to hold is

$$(a - \mu_q)^2 \le (-1 - \mu_q)^2, \tag{110}$$

Solving this inequality, we get $a \le \frac{n-1}{n-3}$. □

By substituting $1/n$ with ϵ and scaling all the points, we derive the example in Figure 1, which also shows the tightness of Theorem 2.1. This example shows that when the adversary puts the corrupted point ($X = a$) far away from the other points, the distribution that puts mass on the corrupted point will not be a local minimum if $n > 3$. On the other hand, when the corrupted point is near the other points, the distribution can be a local minimum, but not a global minimum in general. What happens when $n = 3$? The next result shows when $n = 3$, one may have a arbitrarily big and *break down* when $\epsilon = 1/3$.

Theorem C.1. *For $\epsilon = 1/3$ and any $a > 0$, there exists some distribution p_S such that $\|\Sigma_{p_S}\| \le \sigma^2$, while the mean of some local minimum of (10) μ_q satisfies*

$$\|\mu_q - \mu_{p_S}\| \ge a. \tag{111}$$

Proof. Here we consider the simple case when $n = 3, d = 1, \epsilon = 1/3$, and the number of 'good' points is 2, and the number of 'bad' points is 1, i.e.,

$$x_1 = 0, x_2 = 1, x_3 = a,$$

where x_1 and x_2 are good points and x_3 is the outlier. If we set q as follows,

$$q_1 = 0, q_2 = \frac{1}{2}, q_3 = \frac{1}{2},$$

then we set w as

$$w = \sum_{i=1}^{3} q_i x_i = \frac{1}{2} \cdot a = \frac{1+a}{2},$$

then we have

$$\begin{aligned}
(x_1 - w)^2 &= (0 - \frac{1+a}{2})^2 = (\frac{a+1}{2})^2 \\
(x_2 - w)^2 &= (1 - \frac{1+a}{2})^2 = (\frac{a-1}{2})^2 \\
(x_3 - w)^2 &= (a - \frac{1+a}{2})^2 = (\frac{a-1}{2})^2,
\end{aligned}$$

thus, if we set

$$\lambda = -(\frac{a-1}{2})^2, u_1 = (\frac{a+1}{2})^2 - (\frac{a-1}{2})^2, v_1 = 0, u_2 = u_3 = 0, v_2 = v_3 = 0,$$

then the current $\{x_i\}_{i=1}^3$, $\{q_i\}_{i=1}^3$, $\{u_i\}_{i=1}^3$, $\{v_i\}_{i=1}^3$, λ, w, satisfy the KKT condition, but it is not a good solution.

Now we verify that it is also a local minimum. For any fixed q, the optimal w is always $w = \sum_{i=1}^3 q_i x_i$. Thus it suffices to consider any perturbation on q. Denote $q_1' = s+t, q_2' = \frac{1}{2} - t, q_3' = \frac{1}{2} - s$ for small $s, t > 0$. Then we have

$$\begin{aligned}
&\sum_{i=1}^{3} q_i (x_i - \sum_{i=1}^{3} q_i x_i)^2 - \sum_{i=1}^{3} q_i' (x_i - \sum_{i=1}^{3} q_i' x_i)^2 \\
=&(\frac{a-1}{2})^2 - \frac{(s+t)(a+1)^2}{2} - (\frac{1}{2} - t)(\frac{a-1}{2} - t - sa)^2 - (\frac{1}{2} - s)(\frac{-a+1}{2} - t - sa)^2 \\
=&(\frac{a-1}{2})^2 - \frac{(s+t)(a+1)^2}{2} - (1 - s - t)((\frac{a-1}{2})^2 + (t + sa)^2) \\
&- (s - t)(1 - a)(t + sa) \\
=&- O(a(s+t)) < 0.
\end{aligned} \tag{112}$$

Thus we can see q is a local minimum. □

C.5 Proof of Theorem 2.2

For the supremum achieving v, we have

$$\begin{aligned}
\|\Sigma_q\| &= \mathbb{E}_q[(v^\top(X-\mu_q)^2)] && (113)\\
&\overset{(i)}{\le} (1+\alpha)\mathbb{E}_{p_*}[(v^\top(X-\mu_q)^2)]+\beta && (114)\\
&= (1+\alpha)\mathbb{E}_{p_*}[(v^\top(X-\mu_{p_*})^2)+(v^\top(\mu_q-\mu_{p_*}))^2]+\beta && (115)\\
&\le (1+\alpha)(\|\Sigma_{p_*}\|+\|\mu_q-\mu_{p_*}\|^2)+\beta && (116)\\
&\le (1+\alpha)\left(\|\Sigma_{p_*}\|+\left(\sqrt{\frac{\|\Sigma_q\|\epsilon}{1-2\epsilon}}+\sqrt{\frac{\|\Sigma_{p_*}\|\epsilon}{1-2\epsilon}}\right)^2\right)+\beta. && (117)
\end{aligned}$$

Here (i) comes from the assumption in the theorem, (ii) comes from Lemma 2.2 and Lemma 2.3.

Solving the above inequality on $\|\Sigma_q\|$, we know that when $\epsilon \in [0, 1/(3+\alpha))$,

$$\|\Sigma_q\| \le \left(1+\frac{C_1(\alpha+\epsilon)}{(1-(3+\alpha)\epsilon)^2}\right)\|\Sigma_{p_*}\|+\frac{C_2\beta}{(1-(3+\alpha)\epsilon)^2}, \tag{118}$$

for some constant C_1, C_2.

D Proof for Section 3

D.1 Stationary point for hypercontractivity is not an approximately good solution

Consider the task of finding some distribution q that is hypercontractive given corrupted distribution from a hypercontractive distribution, which is a sub-question from linear regression. To be concrete, we assume that the true distribution p_S satisfies $F(p_S) \le \kappa^2$, where

$$F(p) = \frac{\mathbb{E}_p[(v^\top X)^4]}{\mathbb{E}_p[(v^\top X)^2]^2}. \tag{119}$$

The feasibility problem in Problem 1.1 reduces to finding some distribution q such that $\mathsf{TV}(q, p_S) \le \epsilon/(1-\epsilon)$, $F(q) \le \kappa'^2$.

As the case of mean estimation, a natural approach to solve the feasibility problem is transfer that to the optimization problem of $\min_{q\in\Delta_{n,\epsilon}} F(q)$. However, different from the case of mean estimation, the stationary point for this function is not a good solution for the feasibility problem. We show it in the below theorem.

Theorem D.1. *Given any $\kappa' > \kappa > 0$, there exists some n, ϵ such that one can design some one-dimensional distribution p_n satisfying: (a) there exists some set $S \subset [n], |S| \ge (1-\epsilon)n$, $F(p_S) \le \kappa^2$; (b) there exists a distribution q with $F(q) \ge \kappa'^2$, while q is a stationary point for the optimization problem $\min_{q\in\Delta_{n,\epsilon}} F(q)$.*

Remark D.1. *Since any stationary point guarantees $\mathbb{E}_q[g] \le \mathbb{E}_{p_S}[g]$ with g taken as the partial derivative of $F(q)$ with respect to q. The counter example also shows that we cannot take this g as generalized quasi-gradient in general.*

Proof. The Lagrangian for the optimization problem is

$$L(q,u,y,\lambda) = F(q) + \sum_{i=1}^{n} u_i\Big(-q_i\Big) + \sum_{i=1}^{n} y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) + \lambda\Big(\sum_{i}^{n} q_i - 1\Big).$$

From the KKT conditions, we know that the stationary points must satisfy

$$\begin{aligned}
&\text{(stationarity)} \quad 0 \in \partial_q\Big(F(q) + \sum_{i=1}^{n} u_i q_i + \sum_{i=1}^{n} y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) + \lambda\Big(\sum_{i}^{n} q_i - 1\Big)\Big),\\
&\text{(complementary slackness)} \quad u_i(-q_i) = 0,\ y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) = 0,\ i \in [n],\\
&\text{(primal feasibility)} \quad -q_i \le 0,\ q_i - \frac{1}{(1-\epsilon)n} \le 0,\ \sum_{i}^{n} q_i = 1,\\
&\text{(dual feasibility)} \quad u_i \ge 0,\ y_i \ge 0,\ i \in [n].
\end{aligned}$$

Denote τ_i as

$$\tau_i = \partial_{q_i} F(q) = \frac{X_i^4}{(\sum_{i\in[n]} q_i X_i^2)^2} - \frac{2X_i^2(\sum_{i\in[n]} q_i X_i^4)}{(\sum_{i\in[n]} q_i X_i^2)^3}. \tag{120}$$

Then we will have

$$0 = \tau_i - u_i + y_i + \lambda,\ i \in [n], \tag{121}$$

Next, we define two sets,

$$S_+ = \Big\{i\Big|q_i > 0\Big\}, \quad S_- = \Big\{i\Big|q_i = 0\Big\}$$

and $(1-\epsilon)n \le |S_+| \le n$, $|S_-| \le \epsilon n$. For $i \in [n]$ such that $q_i = \frac{1}{(1-\epsilon)n}$, which implies that $u_i = 0$, we have

$$\tau_i = \underbrace{-y_i}_{\le 0} - \lambda \le -\lambda.$$

For $i \in S_-$, $q_i = 0$, which implies that $y_i = 0$, we know that

$$\tau_i = \underbrace{u_i}_{\ge 0} - \lambda \ge -\lambda.$$

For i such that $0 < q_i < 1/(1-\epsilon)n$, which implies that $u_i = y_i = 0$, we know that

$$\tau_i = -\lambda.$$

Now we are ready to construct p_n, p_S, q such that q is a stationary point of the optimization problem.

Consider distribution p_n with $1-\delta-\gamma$ fraction of points to be 0, δ fraction of points at $a > 0$, and γ fraction of points at b. We assume that $\delta > \epsilon > \gamma$ and $a < b$. Assume that n is picked such that all the fractions listed below multiplying n will be integer. Let p_S be the distribution of completely deleting point b from p_n, i.e. p_S has $\frac{1-\delta-\gamma}{1-\gamma}$ fraction of points on 0, and $\frac{\delta}{1-\gamma}$ fraction of points on a. We take δ, γ such that it satisfies

$$\frac{1-\gamma}{\delta} \le \kappa^2, \tag{122}$$

which implies that $F(p_S) \le \kappa^2$.

Let q be the distribution of deleting ϵ mass from point a, i.e. q has $\frac{1-\delta-\gamma}{1-\epsilon}$ mass on 0, $\frac{\delta-\epsilon}{1-\epsilon}$ mass on a, $\frac{\gamma}{1-\epsilon}$ mass on b. We first verify that q is a stationary point for this problem. Denote A, B, C as the set of indexes of points that are supported on $0, a$ and b, separately. For any point $i \in A$, from (120), we have $\tau_i = 0$. Since all the points have $q_i = \frac{1}{(1-\epsilon)n}$, we know that

$$\forall i \in A, \tau_i = 0 = -y_i - \lambda, y_i \ge 0. \tag{123}$$

For any point $i \in C$, from (120), we have $\tau_i = 0$. Since all the points have $q_i = \frac{1}{(1-\epsilon)n}$, we know that

$$\begin{aligned}\forall i \in C, \tau_i &= \frac{X_i^4}{(\sum_{i\in[n]} q_i X_i^2)^2} - \frac{2X_i^2(\sum_{i\in[n]} q_i X_i^4)}{(\sum_{i\in[n]} q_i X_i^2)^3} \\ &= \frac{b^4}{((\delta-\epsilon)a^2/(1-\epsilon) + \gamma b^2/(1-\epsilon))^2} - \frac{2b^2((\delta-\epsilon)a^4/(1-\epsilon) + \gamma b^4/(1-\epsilon))}{((\delta-\epsilon)a^2/(1-\epsilon) + \gamma b^2/(1-\epsilon))^3}.\end{aligned} \tag{124}$$

For some fixed δ, a, we set γ such that $(\delta-\epsilon)a^2/(1-\epsilon) = \gamma b^2/(1-\epsilon)$. Thus we have

$$\begin{aligned}\forall i \in C, \tau_i = -y_i - \lambda &= \frac{(1-\epsilon)^2}{4\gamma^2} - \frac{2b^2 \cdot ((\delta-\epsilon)a^4/(1-\epsilon) + \gamma b^4/(1-\epsilon))}{(2\gamma b^2/(1-\epsilon))^3} \\ &< \frac{(1-\epsilon)^2}{4\gamma^2} - \frac{2b^2 \cdot (\gamma b^4/(1-\epsilon))}{(2\gamma b^2/(1-\epsilon))^3} \\ &= 0, y_i \ge 0.\end{aligned} \tag{125}$$

Similarly, for any point $i \in B$, there are some points i with $q_i = \frac{1}{(1-\epsilon)n}$. Denote the set as D, and the rest as B/D. Then we can similarly compute that

$$\begin{aligned}\forall i \in D, \tau_i = c = -y_i - \lambda > 0, y_i \le 0 \\ \forall i \in B/D, \tau_i = c = u_i - \lambda > 0, u_i \ge 0\end{aligned} \tag{126}$$

where c is some positive value. We can see that there exists $y_i \le 0, u_i \ge 0, \lambda$ such that Equation (123), (125) and (126) hold simultaneously by taking $\lambda = -c$, $\forall i \in B, u_i = y_i = 0$,, $\forall i \in A, y_i = c > 0$, $\forall i \in B, y_i = -\tau_i + c > 0$. Thus q is a stationary point. Now we let $b \to \infty$, since we have set γ such that $(\delta-\epsilon)a^2/(1-\epsilon) = \gamma b^2/(1-\epsilon)$, we have $\gamma \to 0$, and

$$\frac{\mathbb{E}_q[X^4]}{\mathbb{E}_q[X^2]^2} = \frac{\gamma b^4(1-\epsilon)}{\gamma^2 b^4} = \frac{1-\epsilon}{\gamma} \to \infty. \tag{127}$$

□

D.2 Proof of Auxillary Lemmas

The following Lemma gives a tighter bound on the modulus with respect to the coefficient in front of τ than (Zhu et al., 2019, Lemma E.3).

Lemma D.1 (Modulus of continuity for mean estimation with near identity covariance)**.** *For some fixed $\epsilon \in [0, 1)$ and non-negative constants ρ, τ and τ', define*

$$\mathcal{G}_1 = \{p \mid \forall r \le \frac{p}{1-\epsilon}, \|\mu_r - \mu_p\|_2 \le \rho, \lambda_{\mathsf{min}}(\mathbb{E}_r[(X-\mu_p)(X-\mu_p)^\top]) \ge 1-\tau\} \tag{128}$$

$$\mathcal{G}_2 = \{p \mid \|\mathbb{E}_p[(X-\mu_p)(X-\mu_p)^\top]\|_2 \le 1+\tau'\}. \tag{129}$$

Here $\lambda_{\mathsf{min}}(A)$ is the smallest eigenvalue of symmetric matrix A. Assume $p_S \in \mathcal{G}_1$, $q \in \mathcal{G}_2$, $\mathsf{TV}(q, p_S) \le \epsilon$. Then we have

$$\sup_{p\in\mathcal{G}_1, q\in\mathcal{G}_2, \mathsf{TV}(p,q)\le\epsilon} \|\mu_p - \mu_q\|_2 \le \frac{\rho}{1-\epsilon} + \sqrt{\frac{\epsilon(\tau+\tau'+\epsilon)}{1-\epsilon} + \frac{\epsilon\rho^2}{(1-\epsilon)^2}}. \tag{130}$$

Here C is some universal constant.

Proof. Assume $p \in \mathcal{G}_1, q \in \mathcal{G}_2$, $p \neq q$. Without loss of generality, we assume $\mu_p = 0$. From $\mathsf{TV}(p, q) = \epsilon_0 \le \epsilon$, we construct distribution $r = \frac{\min(p,q)}{1-\epsilon_0}$. Then we know that $r \le \frac{p}{1-\epsilon_0}$, $r \le \frac{q}{1-\epsilon_0}$. Denote $\tilde{r} = (1-\epsilon_0)r$. Consider measure $p-\tilde{r}, q-\tilde{r}$. We have $\mu_q = \mu_p - \mu_{p-\tilde{r}} + \mu_{q-\tilde{r}} = -\mu_{p-\tilde{r}} + \mu_{q-\tilde{r}}$. Note that $\|\mu_{p-\tilde{r}}\|_2 = \|\mu_p - \mu_{\tilde{r}}\|_2 = \|\mu_{\tilde{r}}\|_2 \le (1-\epsilon_0)\rho \le \rho$. For any $v \in \boldsymbol{R}^d$, $\|v\|_2 = 1$, we have

$$\begin{aligned}
v^\top \Sigma_q v^\top &= v^\top(\mathbb{E}_q[XX^\top] - \mu_q\mu_q^\top)v \\
&= v^\top(\mathbb{E}_{\tilde{r}}[XX^\top] + \mathbb{E}_{q-\tilde{r}}[XX^\top] - (\mu_{q-\tilde{r}} - \mu_{p-\tilde{r}})(\mu_{q-\tilde{r}} - \mu_{p-\tilde{r}})^\top)v \\
&\ge (1-\tau)(1-\epsilon_0) + \mathbb{E}_{q-\tilde{r}}[(v^\top X)^2] - (v^\top\mu_{q-\tilde{r}})^2 + 2v^\top\mu_{q-\tilde{r}}v^\top\mu_{p-\tilde{r}} - (v^\top\mu_{p-\tilde{r}})^2 \\
&\ge 1 - \tau - \epsilon_0 + \mathbb{E}_{q-\tilde{r}}[(v^\top X)^2] - (v^\top\mu_{q-\tilde{r}})^2 - 2\|\mu_{q-\tilde{r}}\|_2\|\mu_{p-\tilde{r}}\|_2 - \|\mu_{p-\tilde{r}}\|^2 \\
&\ge 1 - \tau - \epsilon_0 + \mathbb{E}_{q-\tilde{r}}[(v^\top X)^2] - (v^\top\mu_{q-\tilde{r}})^2 - 2\rho\|\mu_{q-\tilde{r}}\|_2 - \rho^2.
\end{aligned} \tag{131}$$

Denote $b_q = \frac{q-\tilde{r}}{\epsilon_0}$. Then b_q is a distribution. If $\mu_{b_q} = 0$, then we already know that $\|\mu_q - \mu_r\| \le \epsilon_0\|\mu_{b_q}\|_2 = 0$. Otherwise we take $v = \frac{\mu_{b_q}}{\|\mu_{b_q}\|_2}$. Then we can see $\mathbb{E}_{q-\tilde{r}}[(v^\top X)^2] = \epsilon_0\mathbb{E}_{b_q}[(v^\top X)^2] \ge \epsilon_0\|\mu_{b_q}\|_2^2$. From $q \in \mathcal{G}_2$, we know that $v^\top\Sigma_q v \le 1 + \tau'$. Thus

$$(\epsilon_0 - \epsilon_0^2)\|\mu_{b_q}\|_2^2 - 2\epsilon_0\rho\|\mu_{b_q}\|_2 \le \rho^2 + \tau + \tau' + \epsilon_0. \tag{132}$$

Solving the inequality, we derive that

$$\|\mu_q - \mu_r\|_2 \le \epsilon_0\|\mu_{b_q}\|_2 \le \frac{\epsilon\rho}{1-\epsilon} + \sqrt{\frac{\epsilon(\tau+\tau'+\epsilon)}{1-\epsilon} + \frac{\epsilon\rho^2}{(1-\epsilon)^2}}. \tag{133}$$

where C is some universal constant. Thus we can conclude

$$\|\mu_p - \mu_q\|_2 \le \|\mu_p - \mu_r\|_2 + \|\mu_r - \mu_q\|_2 \le \frac{\rho}{1-\epsilon} + \sqrt{\frac{\epsilon(\tau+\tau'+\epsilon)}{1-\epsilon} + \frac{\epsilon\rho^2}{(1-\epsilon)^2}}. \tag{134}$$

□

Lemma D.2 (Sum-of-squares modulus of continuity for bounded covariance distributions)**.** *Assume that $\mathsf{TV}(p, q) \le \epsilon$ and both p, q has finite second moment. Then we have*

$$\begin{aligned}
(\mathbb{E}_q[v^\top X] - \mathbb{E}_p[v^\top X])^2 &\preceq_{\text{sos}} \frac{2\epsilon}{(1-\epsilon)^2} \cdot (\mathbb{E}_p[(v^\top(X-\mu_p))^2] + \mathbb{E}_q[(v^\top(X-\mu_q))^2]) \\
&\preceq_{\text{sos}} \frac{2\epsilon}{(1-\epsilon)^2} \cdot (\mathbb{E}_p[(v^\top X)^2] + \mathbb{E}_q[(v^\top X)^2]).
\end{aligned}$$

Proof. For any distribution p, q with $\mathsf{TV}(p, q) \le \epsilon$, there exists some distribution r such that r is an ϵ-deletion of both distributions, i.e. $r \le \frac{p}{1-\epsilon}$, $r \le \frac{q}{1-\epsilon}$.

The proof uses the property that for any $r \le \frac{p}{1-\eta}$, there exists some event E such that $\mathbb{P}_p(E) \ge 1-\epsilon$ and $\mathbb{E}_r[f(X)] = \mathbb{E}_p[f(X)|E]$ for any measurable f (see e.g. (Zhu et al., 2019, Lemma C.1)). For any event E with $\mathbb{P}_p(E) \ge 1-\epsilon$, denote its compliment as E^c. We have

$$\begin{aligned}
\mathbb{E}_q[(v^\top X)^2] &\succeq_{\text{sos}} \mathbb{E}_q[(v^\top (X-\mu_q))^2] \\
&\succeq_{\text{sos}} \mathbb{E}_q[(v^\top (X-\mu_q))^2 \mathbb{1}(E^c)] \\
&\overset{(i)}{\succeq}_{\text{sos}} \mathbb{E}_q[(v^\top (X-\mu_q))\mathbb{1}(E^c)]^2/\epsilon \\
&\overset{(ii)}{=} \mathbb{E}_q[(v^\top (X-\mu_q))\mathbb{1}(E)]^2/\epsilon \\
&\overset{(iii)}{\succeq}_{\text{sos}} (\mathbb{E}_r[v^\top X] - \mathbb{E}_q[v^\top X])^2 (1-\epsilon)^2/\epsilon.
\end{aligned}$$

Here (i) comes from SOS-Hölder's inequality, (ii) comes from the fact that $\mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])^2 \mathbb{1}(E^c)] + \mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])^2 \mathbb{1}(E)] = 0$, (iii) comes from that $r = p|E$. Thus we have

$$(\mathbb{E}_r[v^\top X] - \mathbb{E}_q[v^\top X])^2 \preceq_{\text{sos}} \frac{\epsilon}{(1-\epsilon)^2} \mathbb{E}_q[(v^\top (X-\mu_q))^2].$$

Using the same argument for p, we have

$$(\mathbb{E}_r[v^\top X] - \mathbb{E}_p[v^\top X])^2 \preceq_{\text{sos}} \frac{\epsilon}{(1-\epsilon)^2} \mathbb{E}_p[(v^\top (X-\mu_p))^2].$$

By summing the two SOS-inequalities, we have

$$\begin{aligned}
(\mathbb{E}_q[v^\top X] - \mathbb{E}_p[v^\top X])^2 &\preceq_{\text{sos}} 2((\mathbb{E}_r[v^\top X] - \mathbb{E}_q[v^\top X])^2 + (\mathbb{E}_r[v^\top X] - \mathbb{E}_p[v^\top X])^2) \\
&\preceq_{\text{sos}} \frac{2\epsilon}{(1-\epsilon)^2} (\mathbb{E}_p[(v^\top (X-\mu_p))^2] + \mathbb{E}_q[(v^\top (X-\mu_q))^2]) \\
&\preceq_{\text{sos}} \frac{2\epsilon}{(1-\epsilon)^2} (\mathbb{E}_p[(v^\top X)^2] + \mathbb{E}_q[(v^\top X)^2]).
\end{aligned}$$

□

Lemma D.3 (Modulus of continuity for certifiable hypercontractivity)**.** *Denote* $\mathcal{G}(\kappa^2) = \{p \mid \mathbb{E}_p[(v^\top X)^4] \preceq_{\text{sos}} \kappa^2 \mathbb{E}_p[(v^\top X)^2]^2\}$. *Assume that* $p \in \mathcal{G}(\kappa^2), q \in \mathcal{G}(\kappa'^2)$, $\mathsf{TV}(p,q) \le \epsilon$. *Then we have*

$$\frac{(1-\epsilon-\kappa\sqrt{\epsilon})^2}{(1-\epsilon+\kappa\sqrt{\epsilon})^2} \mathbb{E}_q[(v^\top X)^2]^2 \preceq_{\text{sos}} \mathbb{E}_p[(v^\top X)^2]^2 \preceq_{\text{sos}} \frac{(1-\epsilon+\kappa\sqrt{\epsilon})^2}{(1-\epsilon-\kappa\sqrt{\epsilon})^2} \mathbb{E}_q[(v^\top X)^2]^2.$$

Proof. For any distribution p, q with $\mathsf{TV}(p,q) \le \epsilon$, there exists some distribution r such that r is an ϵ-deletion of both distributions, i.e. $r \le \frac{p}{1-\epsilon}$, $r \le \frac{q}{1-\epsilon}$ from (Zhu et al., 2019, Lemma C.1). The proof uses the property that for any $r \le \frac{p}{1-\eta}$, there exists some event E such that $\mathbb{P}_p(E) \ge 1-\epsilon$ and $\mathbb{E}_r[f(X)] = \mathbb{E}_p[f(X)|E]$ for any measurable f (see e.g. (Zhu et al., 2019, Lemma C.1)). For any event E with $\mathbb{P}_p(E) \ge 1-\epsilon$, denote its compliment as E^c. We have

$$\begin{aligned}
\kappa^2 \mathbb{E}_q[(v^\top X)^2]^2 &\succeq_{\text{sos}} \mathbb{E}_q[(v^\top X)^4] \\
&\succeq_{\text{sos}} \mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])^2] \\
&\succeq_{\text{sos}} \mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])^2 \mathbb{1}(E^c)] \\
&\overset{(i)}{\succeq}_{\text{sos}} \mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])\mathbb{1}(E^c)]^2/\epsilon \\
&\overset{(ii)}{=} \mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])\mathbb{1}(E)]^2/\epsilon \\
&\overset{(iii)}{\succeq}_{\text{sos}} (\mathbb{E}_r[(v^\top X)^2] - \mathbb{E}_q[(v^\top X)^2])^2 (1-\epsilon)^2/\epsilon.
\end{aligned}$$

Here (i) comes from SOS-Hölder's inequality, (ii) comes from the fact that $\mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])^2 \mathbb{1}(E^c)] + \mathbb{E}_q[((v^\top X)^2 - \mathbb{E}_q[(v^\top X)^2])^2 \mathbb{1}(E)] = 0$, (iii) comes from that $r = p|E$. Thus we have

$$\begin{aligned}
&(\mathbb{E}_r[(v^\top X)^2] - \mathbb{E}_q[(v^\top X)^2])^2 \preceq_{\text{sos}} \frac{\epsilon\kappa^2}{(1-\epsilon)^2}\mathbb{E}_q[(v^\top X)^2]^2 \\
\Rightarrow\ &\mathbb{E}_r[(v^\top X)^2]^2 + \mathbb{E}_q[(v^\top X)^2] \preceq_{\text{sos}} \frac{\epsilon\kappa^2}{(1-\epsilon)^2}\mathbb{E}_q[(v^\top X)^2]^2 + 2\mathbb{E}_r[(v^\top X)^2]^2\mathbb{E}_q[(v^\top X)^2]^2 \\
&\preceq_{\text{sos}} \frac{\epsilon\kappa^2}{(1-\epsilon)^2}\mathbb{E}_q[(v^\top X)^2]^2 + \frac{1}{\alpha}\mathbb{E}_r[(v^\top X)^2]^4 + \alpha\mathbb{E}_q[(v^\top X)^2]^2.
\end{aligned}$$

By optimizing over α in the regime $\alpha > 1$ and $\alpha \in (0,1)$ separately, we can get

$$\frac{1}{(1+\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})}\mathbb{E}_q[(v^\top X)^2]^2 \preceq_{\text{sos}} \mathbb{E}_r[(v^\top X)^2]^2 \preceq_{\text{sos}} \frac{1}{(1-\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})}\mathbb{E}_q[(v^\top X)^2]^2. \quad (135)$$

Using the same argument we have

$$\frac{1}{(1+\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})^2}\mathbb{E}_p[(v^\top X)^2]^2 \preceq_{\text{sos}} \mathbb{E}_r[(v^\top X)^2]^2 \preceq_{\text{sos}} \frac{1}{(1-\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})^2}\mathbb{E}_p[(v^\top X)^2]^2. \quad (136)$$

Thus we have

$$\frac{(1-\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})^2}{(1+\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})^2}\mathbb{E}_q[(v^\top X)^2]^2 \preceq_{\text{sos}} \mathbb{E}_p[(v^\top X)^2]^2 \preceq_{\text{sos}} \frac{(1+\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})^2}{(1-\sqrt{\epsilon\kappa^2/(1-\epsilon)^2})^2}\mathbb{E}_q[(v^\top X)^2]^2. \quad (137)$$

□

D.3 Proof of Theorem 3.1

We prove the theorem via two separate arguments. First, we show that $\tilde{g}_1$ is a valid generalized quasi-gradient for certifiable hypercontractivity. Then, given the knowledge that q is certifiably hypercontractive, we show that g_2 is a valid generalized quasi-gradient for bounded noise condition.

Lemma D.4 (Generalized quasi-gradient for hypercontractivity)**.** *Under the same assumption as Theorem 3.1, for any $q \in \Delta_{n,\epsilon}$ that satisfies $\mathbb{E}_q[g_1(X;q)] \le \mathbb{E}_{p_S}[g_1(X;q)]$, when $9\epsilon\kappa^2/(1-2\epsilon)^2 \le 1$ we have*

$$\mathbb{E}_q[(v^\top X)^4] \preceq_{\text{sos}} 4\kappa^2\mathbb{E}_q[(v^\top X)^2]^2. \quad (138)$$

Proof. Denote $\tilde{\kappa}^2 = \sup_{E_v \in \mathcal{E}_4} \frac{E_v[\mathbb{E}_q[(v^\top X)^4]]}{E_v[\mathbb{E}_q[(v^\top X)^2]^2]}$. Assume that $\tilde{\kappa} \ge \kappa$, since otherwise we already have $F(q) \le \kappa^2$. From the modulus of continuity for second moment in Lemma D.2, since $\mathsf{TV}(p_S, q) \le \frac{\epsilon}{1-\epsilon}$, we have

$$(\mathbb{E}_q[(v^\top X)^2] - \mathbb{E}_{p_S}[(v^\top X)^2])^2 \preceq_{\text{sos}} \frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}(\mathbb{E}_{p_S}[(v^\top X)^4] + \mathbb{E}_q[(v^\top X)^4]). \quad (139)$$

Thus we know that for the specific choice of pseudoexpectation E_v, we have

$$\begin{aligned} E_v[(\mathbb{E}_q[(v^\top X)^2] - \mathbb{E}_{p_S}[(v^\top X)^2])^2] &\le \frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2} E_v[\mathbb{E}_{p_S}[(v^\top X)^4] + \mathbb{E}_q[(v^\top X)^4]] \\ &\overset{(i)}{\le} \frac{4\epsilon}{(1-2\epsilon)^2} \cdot E_v[\mathbb{E}_{p_S}[(v^\top X)^4]] \\ &\overset{(ii)}{\le} \frac{4\kappa^2\epsilon}{(1-2\epsilon)^2} \cdot E_v[\mathbb{E}_{p_S}[(v^\top X)^2]^2]. \end{aligned} \tag{140}$$

Here (i) comes from the assumption, and (ii) is from the assumption that $\tilde{F}(p_S) \le \kappa^2$. Rearranging the inequality gives us

$$\begin{aligned} E_v[(\mathbb{E}_q[(v^\top X)^2]^2] + E_v[\mathbb{E}_{p_S}[(v^\top X)^2]^2] &\le \frac{4\kappa^2\epsilon}{(1-2\epsilon)^2} \cdot E_v[\mathbb{E}_{p_S}[(v^\top X)^2]^2] + 2E_v[\mathbb{E}_q[(v^\top X)^2] \cdot \mathbb{E}_{p_S}[(v^\top X)^2]] \\ &\le \frac{4\kappa^2\epsilon}{(1-2\epsilon)^2} \cdot E_v[\mathbb{E}_{p_S}[(v^\top X)^2]^2] + \frac{1}{\alpha} E_v[\mathbb{E}_q[(v^\top X)^2]^2] + \alpha E_v[\mathbb{E}_{p_S}[(v^\top X)^2]^2], \end{aligned} \tag{141}$$

for any $\alpha > 0$. By optimizing over α, we have

$$E_v[\mathbb{E}_q[(v^\top X)^2]^2] \le \gamma^2 E_v[\mathbb{E}_{p_S}[(v^\top X)^2]^2], \tag{142}$$

where $\gamma^2 = \frac{(1+\sqrt{\epsilon\kappa^2/(1-2\epsilon)^2})^2}{(1-\sqrt{\epsilon\kappa^2/(1-2\epsilon)^2})^2}$. Thus we know that

$$\begin{aligned} E_v[\mathbb{E}_q[(v^\top X)^4)]] &\le E_v[\mathbb{E}_{p_S}[(v^\top X)^4)]] \\ &\le \kappa^2 E_v[\mathbb{E}_{p_S}[(v^\top X)^2)]^2] \\ &\le \gamma^2\kappa^2 E_v[\mathbb{E}_q[(v^\top X)^2)]^2] \qquad (143) \\ &= \frac{\gamma^2\kappa^2}{\tilde{\kappa}^2} E_v[\mathbb{E}_q[(v^\top X)^4)]]. \qquad (144) \end{aligned}$$

By solving the above inequality, we have when $9\epsilon\kappa^2/(1-2\epsilon)^2 \le 1$,

$$\tilde{\kappa} \le 2\kappa. \tag{145}$$

□

Now assume that we already know that q is hypercontractive with parameter 2κ. We show that q also satisfies bounded noise condition via the following lemma.

Lemma D.5. *Under the same assumption as Theorem 3.1, assume $q \in \Delta_{n,\epsilon}$ satisfies*

$$\mathbb{E}_q[g_2(X;q)] \le \mathbb{E}_{p_S}[g_2(X;q)], \forall v \in \boldsymbol{R}^d, \mathbb{E}_q[(v^\top X)^4] \le 4\kappa^2 \mathbb{E}_q[(v^\top X)^2]^2. \tag{146}$$

Then when $\kappa^3\epsilon < 1/64$, we have

$$\forall v \in \boldsymbol{R}^d, \mathbb{E}_q[(Y - X^\top\theta^*(q))^2(v^\top X)^2] \le 3\sigma^2 \mathbb{E}_q[(v^\top X)^2]. \tag{147}$$

Proof. Denote $\tilde{\sigma}^2 = \sup_{v\in\boldsymbol{R}^d} \mathbb{E}_q[(Y - X^\top\theta^*(q)^2(v^\top X)^2]/\mathbb{E}_q[(v^\top X)^2]$. Then we have

$$\begin{aligned}\mathbb{E}_q[(Y - X^\top\theta^*(q))^2(v^\top X)^2] &\le \mathbb{E}_{p_S}[(Y - X^\top\theta^*(q))^2(v^\top X)^2] \\ &\le \mathbb{E}_{p_S}[(Y - X^\top\theta^*(p_S))^2(v^\top X)^2] + \mathbb{E}_{p_S}[X^\top(\theta^*(q) - \theta^*(p_S))^2(v^\top X)^2] \\ &\overset{(i)}{\le} \mathbb{E}_{p_S}[(Y - X^\top\theta^*(p_S))^2(v^\top X)^2] + \mathbb{E}_{p_S}[(X^\top(\theta^*(p_S) - \theta^*(q)))^4]^{1/2}\cdot\mathbb{E}_{p_S}[(v^\top X)^4]^{1/2} \\ &\overset{(ii)}{\le} \sigma^2\mathbb{E}_{p_S}[(v^\top X)^2] + \kappa^2\mathbb{E}_{p_S}[(X^\top(\theta^*(p_S) - \theta^*(q)))^2]\cdot\mathbb{E}_{p_S}[(v^\top X)^2].\end{aligned} \tag{148}$$

Here (i) is a result of Cauchy-Schwarz inequality, (ii) comes from the hypercontractivity of p_S. From Lemma D.3, we know that

$$\mathbb{E}_{p_S}[(v^\top X)^2] \le \frac{1-\epsilon+2\kappa\sqrt{\epsilon}}{1-\epsilon-2\kappa\sqrt{\epsilon}}\mathbb{E}_q[(v^\top X)^2]. \tag{149}$$

From (Zhu et al., 2019, Theorem 3.4), we know that when $\kappa^2\epsilon < 1/32$

$$\mathbb{E}_q[((\theta^*(p_S) - \theta^*(q))^\top X)^2] \le \frac{2\kappa\tilde{\sigma}^2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}. \tag{150}$$

Thus

$$\begin{aligned}\mathbb{E}_q[(Y - X^\top\theta^*(q))^2(v^\top X)^2] &\le \frac{1-\epsilon+2\kappa\sqrt{\epsilon}}{1-\epsilon-2\kappa\sqrt{\epsilon}}\cdot(\sigma^2 + \frac{2\kappa^3\tilde{\sigma}^2\epsilon(1-\epsilon)}{(1-2\epsilon)^2})\mathbb{E}_q[(v^\top X)^2] \\ &\le \frac{1-\epsilon+2\kappa\sqrt{\epsilon}}{1-\epsilon-2\kappa\sqrt{\epsilon}}\cdot(\frac{\sigma^2}{\tilde{\sigma}^2} + \frac{2\kappa^3\epsilon(1-\epsilon)}{(1-2\epsilon)^2})\mathbb{E}_q[(Y - X^\top\theta^*(q))^2(v^\top X)^2].\end{aligned} \tag{151}$$

By solving the inequality, we know that when $\kappa^3\epsilon < 1/64$, $\tilde{\sigma}^2 \le 3\sigma^2$ (here we use the fact that $\kappa \ge 1$ always holds). □

D.4 Proof of Theorem 3.2

Proof. Denote $\tilde{\kappa}^2 = \sup_{E_v\in\mathcal{E}_4}\frac{E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^4]]}{E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^2]^2]}$. Assume that $\tilde{\kappa} \ge \kappa$, since otherwise we already have $F(q) \le \kappa^2$. From the SOS modulus of continuity for second moment in Lemma D.2, since $\mathsf{TV}(p_S, q) \le \frac{\epsilon}{1-\epsilon}$ from Lemma 2.3, we have,

$$(\mathbb{E}_q[(v^\top(X-\mu_q))^2] - \mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2])^2 \preceq_{\text{sos}} \frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}(\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^4] + \mathbb{E}_q[(v^\top(X-\mu_q))^4]). \tag{152}$$

Thus we know that for E_v, we have

$$
\begin{aligned}
&E_v[(\mathbb{E}_q[(v^\top(X-\mu_q))^2]-\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2])^2]\\
\leq&\frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\cdot E_v[\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^4]+\mathbb{E}_q[(v^\top(X-\mu_q))^4]]\\
\overset{(i)}{\leq}&\frac{4\epsilon}{(1-2\epsilon)^2}\cdot E_v[\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^4]]\\
\leq&\frac{32\epsilon}{(1-2\epsilon)^2}\cdot E_v[\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^4]+(v^\top(\mu_q-\mu_{p_S}))^4]\\
\overset{(ii)}{\leq}&\frac{32\epsilon}{(1-2\epsilon)^2}\cdot E_v[\kappa^2\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2]^2+(v^\top(\mu_q-\mu_{p_S}))^4]\\
\overset{(iii)}{\leq}&\frac{32\epsilon}{(1-2\epsilon)^2}\cdot E_v\left[\kappa^2\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2]^2+\left(\frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\left(\mathbb{E}_q[(v^\top(X-\mu_q))^2]+\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2]\right)\right)^2\right]\\
\leq&\frac{32\epsilon}{(1-2\epsilon)^2}\cdot E_v\left[\kappa^2\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2]^2+\left(\frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\left(\mathbb{E}_q[(v^\top(X-\mu_q))^2]+\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2]\right)\right)^2\right].
\end{aligned}
$$

Here (i) comes from the assumption that $\mathbb{E}_q[g]\leq\mathbb{E}_{p_S}[g]$, (ii) is by the certifiable hypercontractivity of p_S, (iii) is from Lemma D.2. By solving the above inequality, we can derive that when $\epsilon<1/(200\kappa^2)$,

$$
E_v[\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2]]\leq\frac{3}{2}E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^2]]. \tag{153}
$$

Thus following a similar line of argument as above, we know that

$$
\begin{aligned}
&E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^4]]\\
&\leq E_v[\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^4]]\\
&\leq 8E_v[\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^4]+(v^\top(\mu_q-\mu_{p_S}))^4]\\
&\leq 8E_v\left[\kappa^2\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2]^2+\left(\frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\left(\mathbb{E}_q[(v^\top(X-\mu_q))^2]+\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2]\right)\right)^2\right]\\
&\leq 8E_v\left[\kappa^2\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2]^2+\left(\frac{2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\left(\mathbb{E}_q[(v^\top(X-\mu_q))^2]+\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2]\right)\right)^2\right]\\
&\leq(6\kappa^2+0.03)E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^2)]^2]\\
&=\frac{7\kappa^2}{\tilde{\kappa}^2}\cdot E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^4)]].
\end{aligned} \tag{154}
$$

By solving the above inequality, we have

$$
\tilde{\kappa}\leq\sqrt{7}\kappa. \tag{155}
$$

□

D.5 Generalized quasi-gradient for sparse mean estimation

We discuss the generalized quasi-gradient for robust sparse mean estimation here. Let $\mathcal{A}_k$ denote the set

$$
\mathcal{A}_k=\{A\in\boldsymbol{R}^{d\times d}:\mathsf{Tr}(A)=1,\|A\|_1\leq k,A\succeq 0\}. \tag{156}
$$

The dual norm induced by $\mathcal{A}_k$, is defined by $\|B\|^*_{\mathcal{A}_k} = \sup_{A\in\mathcal{A}_k} \mathsf{Tr}(AB)$. In the task of robust sparse mean estimation, we set $F(q) = \|\Sigma_q - I\|^*_{\mathcal{X}_k}$ in Problem 1.1 (Li, 2018, Chapter 3), (Diakonikolas et al., 2019b; Li, 2017). Let $g(X;q)$ be

$$g(X;q) = \mathsf{Tr}(A((X-\mu_q)(X-\mu_q)^\top - I)), \text{ where } A \in \arg\max_{A\in\mathcal{A}_k} \mathsf{Tr}(A(\Sigma_q - I)). \tag{157}$$

We show in the following theorem that $g(X;q)$ is a valid generalized quasi-gradient for F.

Theorem D.2 (Generalized quasi-gradients for sparse mean estimation)**.** *Let $\Delta_{S,\epsilon} = \{r \mid \forall i \in [n], r_i \le \frac{p_{S,i}}{1-\epsilon}\}$ denote the set of ϵ-deletions on p_S. We assume that the true distribution p_S has near identity covariance, and its mean is stable under deletions under sparse norm, i.e. the following holds for any $r \in \Delta_{S,\epsilon}$:*

$$\sup_{\|v\|_2\le 1, \|v\|_0 \le k} v^\top(\mu_r - \mu_{p_S}) \le \rho.$$

Assume that $F(p_S) \ge \rho \ge \epsilon$. The following implication holds for $q \in \Delta_{n,\epsilon}$:

$$\mathbb{E}_q[g(X;q)] \le \mathbb{E}_{p_S}[g(X;q)] \Rightarrow F(q) \le C_2(\epsilon)\cdot F(p_S). \tag{158}$$

Here $C_2(\epsilon)$ is some constant that depends on ϵ. Thus g is a generalized quasi-gradient for F with parameter $C_2(\epsilon)$.

Proof. We have

$$\begin{aligned}
F(q) = \mathbb{E}_q[g(X;q)] &\le \mathbb{E}_{p_S}[g(X;q)] \\
&= \mathbb{E}_{p_S}[\mathsf{Tr}(A((X-\mu_q)(X-\mu_q)^\top - I))] \\
&= \mathbb{E}_{p_S}[\mathsf{Tr}(A((X-\mu_{p_S})(X-\mu_{p_S})^\top - I))] + \mathsf{Tr}(A(\mu_q-\mu_{p_S})(\mu_q-\mu_{p_S})^\top) \\
&\le F(p_S) + \mathsf{Tr}(A(\mu_q-\mu_{p_S})(\mu_q-\mu_{p_S})^\top) \\
&\overset{(i)}{\le} F(p_S) + 4 \sup_{\|v\|_2\le 1,\|v\|_0\le k} (v^\top(\mu_q-\mu_{p_S}))^2 \\
&\overset{(ii)}{\le} F(p_S) + C_1\cdot\sqrt{\epsilon(F(p_S)+F(q))}.
\end{aligned} \tag{159}$$

Here (i) comes from Li (2017, Lemma 5.5) , (ii) comes from Li (2017, Proposition 5.6), and C_1 is some universal constant. By solving the above self-normalzing inequality for $F(q)$, we have

$$F(q) \le C_2(\epsilon)\cdot F(p_S) \tag{160}$$

for some C_2 that is a function of ϵ. □

E Proof for Section 4

E.1 Proof of auxillary lemmas

We also introduce the following lemma, which shows that hypercontractivity is approximately closed under deletion.

Lemma E.1. *Suppose that the set S of good points is hypercontractive in the sense that $\frac{1}{|S|}\sum_{i\in S}(v^\top X_i)^4 \preceq_{\text{sos}} (\frac{\kappa}{|S|}\sum_{i\in S}(v^\top X_i)^2)^2$. Then, for any $c_i \le \frac{1}{|S|}$ such that $1-\sum_{i\in S} c_i \le \epsilon$, we have*

$$\frac{1}{n}\sum_{i\in S} c_i (v^\top X_i)^4 \preceq_{\text{sos}} \frac{\kappa^2}{1-\kappa^2\epsilon}(\frac{1}{|S|}\sum_{i\in S} c_i (v^\top X_i)^2)^2. \tag{161}$$

Proof. We expand directly; let

$$A = \frac{1}{|S|}\sum_{i\in S}(v^\top X_i)^4, \quad B = \frac{1}{|S|}\sum_{i\in S}(v^\top X_i)^2, \tag{162}$$

$$C = \sum_{i\in S}(\frac{1}{|S|} - c_i)(v^\top X_i)^4, \quad D = \sum_{i\in S}(\frac{1}{|S|} - c_i)(v^\top X_i)^2. \tag{163}$$

Then our goal is to show that $\frac{\kappa^2}{1-\kappa^2\epsilon}(B-D)^2 - (A-C) \succeq_{\text{sos}} 0$. We are also given that (i) $\kappa^2 B^2 \succeq_{\text{sos}} A$ and we observe that (ii) $C \succeq_{\text{sos}} D^2/(1-\sum_{i=1}^n c_i) \succeq_{\text{sos}} D^2/\epsilon$ by sum-of-squares Hölder's inequality. We thus have

$$\frac{\kappa^2}{1-\kappa^2\epsilon}(B-D)^2 - (A-C) = \frac{\kappa^2}{1-\kappa^2\epsilon}B^2 - \frac{2\kappa^2}{1-\kappa^2\epsilon}BD + \frac{\kappa^2}{1-\kappa^2\epsilon}D^2 - A + C \tag{164}$$

$$\overset{(i)}{\succeq_{\text{sos}}} (\frac{\kappa^2}{1-\kappa^2\epsilon} - \kappa^2)B^2 - \frac{2\kappa^2}{1-\kappa^2\epsilon}BD + (\frac{\kappa^2}{1-\kappa^2\epsilon}D^2 + C) \tag{165}$$

$$\overset{(ii)}{\succeq_{\text{sos}}} (\frac{\kappa^2}{1-\kappa^2\epsilon} - \kappa^2)B^2 - \frac{2\kappa^2}{1-\kappa^2\epsilon}BD + (\frac{\kappa^2}{1-\kappa^2\epsilon} + \frac{1}{\epsilon})D^2 \tag{166}$$

$$= \frac{\kappa^4\epsilon}{1-\kappa^2\epsilon}B^2 - \frac{2\kappa^2}{1-\kappa^2\epsilon}BD + \frac{1/\epsilon}{1-\kappa^2\epsilon}D^2 \tag{167}$$

$$= \frac{\epsilon}{1-\kappa^2\epsilon}(\kappa^2 B - D/\epsilon)^2 \succeq_{\text{sos}} 0, \tag{168}$$

as was to be shown. □

E.2 Proof of Theorem 4.1

With the help of Lemma 4.2, we are ready to prove Theorem 4.1.

Take $\beta = \sigma^2$ in Lemma 4.2. From Lemma 4.2, it suffices to verify Equation 54. Assume for contradiction that $F(q) = \|\Sigma_q\| \ge \sigma'^2$. Denote $\tilde{\sigma}^2 = \mathbb{E}_q[(v^\top(X_i - \mu_q))^2]$, where v is chosen such that $\tilde{\sigma}^2 = \|\Sigma_q\|_2$. Then we have

$$\begin{aligned}
\tilde{\sigma}^2 &\le (1+\eta)\mathbb{E}_{p_S}[(v^\top(X-\mu_q))^2] + \sigma^2 \\
&\le (1+\eta)(\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2] + (\mu_{p_S}-\mu_q)^2) + \sigma^2 \\
&\le (2+\eta)\sigma^2 + (1+\eta)\left(\sqrt{\frac{\sigma^2\epsilon}{1-2\epsilon}} + \sqrt{\frac{\tilde{\sigma}^2\epsilon}{1-2\epsilon}}\right)^2.
\end{aligned} \tag{169}$$

By solving the self-normalizing inequality, we have for $\epsilon < 1/(3+\eta)$,

$$\tilde{\sigma}^2 \le \frac{((1+\eta)\epsilon\sqrt{\sigma^2} + \sqrt{(1+\eta)^2\epsilon^2\sigma^2 + (2-3\epsilon-3\epsilon\eta+\eta)(1-(3+\eta)\epsilon)\sigma^2})^2}{(1-(3+\eta)\epsilon)^2} < \sigma'^2. \tag{170}$$

which contradicts the assumptions. Thus the algorithm must terminate. The conclusion can be seen from Lemma 2.2.

If we take v such that $\tilde{\sigma} \geq 0.9\|\Sigma_q\|_2$, then it suffices to replace $\tilde{\sigma}^2$ in (169) with $1.2\tilde{\sigma}^2$, which still gives a near-optimal bound up to constant.

In Theorem 4.1, we have shown that it suffices to run Algorithm 1 with $d/(\gamma\sigma^2)$ iterations to guarantee small operator norm of covariance of q. Here we bound the computational complexity within each iteration.

- **Finding v such that $\mathbb{E}_q[(v^\top(X-\mu_q))^2] \geq 0.9\|\Sigma_q\|$.** This can be done within $O(nd\log(d))$ time from power method (Hardt and Price, 2014).
- **Solving the projection step.** The projection step is $q_i^{(t+1)} = \mathsf{Proj}^{KL}_{\Delta_{n,\epsilon}}(c_i^{(t)}/(\sum_{i=1}^n c_i^{(t)})) = \arg\min_{q\in\Delta_{n,\epsilon}} \sum_{i=1}^n q_i \log((q_i \sum_{j=1}^n c_j^{(t)})/c_i^{(t)})$. This can be done within $O(n)$ time. We discuss it in detail as below.

Denote $p_i = c_i^{(t)}/\sum_{i=1}^n c_i^{(t)}$, and

$$F(q) = \sum_{i=1}^n q_i \log(q_i/p_i). \tag{171}$$

The Lagrangian for the optimization problem is

$$L(q,u,y,\lambda) = F(q) + \sum_{i=1}^n u_i\Big(-q_i\Big) + \sum_{i=1}^n y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) + \lambda\Big(\sum_i^n q_i - 1\Big).$$

From the KKT conditions, we have

$$\begin{aligned}
&\text{(stationarity)} \quad 0 = \partial_q\Big(F(q) + \sum_{i=1}^n u_i q_i + \sum_{i=1}^n y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) + \lambda\Big(\sum_i^n q_i - 1\Big)\Big),\\
&\text{(complementary slackness)} \quad u_i(-q_i) = 0,\ y_i\Big(q_i - \frac{1}{(1-\epsilon)n}\Big) = 0,\ i\in[n],\\
&\text{(primal feasibility)} \quad -q_i \leq 0,\ q_i - \frac{1}{(1-\epsilon)n} \leq 0,\ \sum_i^n q_i = 1,\\
&\text{(dual feasibility)} \quad u_i \geq 0,\ y_i \geq 0,\ i\in[n].
\end{aligned}$$

Let $g_i = \partial_{q_i} F(q) = 1 + \log(q_i/p_i)$, from the stationary conditions we have

$$0 = g_i - u_i + y_i + \lambda,\ i\in[n], \tag{172}$$

$$\mu_q = w. \tag{173}$$

For any $i\in[n]$, if $q_i = \frac{1}{(1-\epsilon)n}$, we have

$$0 = g_i + y_i + \lambda, y_i \geq 0. \tag{174}$$

If $q_i \in (0, \frac{1}{(1-\epsilon)n})$, we have

$$0 = g_i + \lambda. \tag{175}$$

If $q_i = 0$, we have

$$0 = g_i - u_i + \lambda, u_i \geq 0. \tag{176}$$

Combine the above three equalities, we know that

$$\frac{q_i}{p_i} \le \frac{q_j}{p_j} \le \frac{q_k}{p_k}, \quad \forall i,j,k \text{ with } q_i = \frac{1}{(1-\epsilon)n}, q_j \in (0, \frac{1}{(1-\epsilon)n}), q_k = 0. \tag{177}$$

From this we know that there does not exist k such that $q_k = 0$, otherwise all q_i, q_j shall be 0. We also know that there must be $p_i \ge p_j$. And for all j, we have $q_j/p_j = -\lambda$ for some constant λ.

Thus the algorithm to compute the projection is straightforward: order p_i in a descent way and compare the corresponding KL divergence. In the m-th itereation, we let $q_i = 1/((1-\epsilon)n)$ for all i such that p_i is among the m-th largest masses. For the left q_i, we just renormalize p_i such that $\sum_{i=1}^n q_i = 1$. We compare the n cases and pick one that has the smallest KL divergence. This can be done within $O(n)$ time since it suffices to compare the difference.

Overall, within each iteration, the computational complexity is $O(nd\log(d))$. Overall the computational complexity is $O(nd^2\log(d)/\eta\sigma^2)$. The same complexity within single iteration also applies to filter algorithm for bounded covariance case.

E.3 Proof of Theorem 4.2

Proof. In iteration k, if the algorithm terminates because of $\|\Sigma_{q^{(k)}}\| \le \sigma'^2$, then we know from $\mathsf{TV}(q^{(k)}, p_S) \le \frac{\epsilon}{1-\epsilon}$ and Lemma 2.2 that

$$\|\mu_q - \mu_{p_S}\| \le (\sigma + \sqrt{\xi'})\sqrt{\frac{\epsilon}{1-2\epsilon}} < \frac{4\sigma\sqrt{\epsilon}}{(1-2\epsilon)^{3/2}} \tag{178}$$

If $F(q^{(k)}, \xi') > 0$, we use induction to show that the invariance (55) holds at $c^{(k)}$. Obviously, it holds at $k = 0$. Assume it holds in step k, we will show that the invariance (55) still holds at step $k+1$. Since the deleted probability mass is proportional to $q_i^{(k)}\tau_i^{(k)}$ for each point X_i, it suffices to check

$$\sum_{i\in S} q_i^{(k)}\tau_i^{(k)} \le \frac{1}{2}\sum_{i=1}^n q_i^{(k)}\tau_i^{(k)}. \tag{179}$$

From Lemma C.1, we know that under the invariance (55) the probability of set S under q, $q(S)$, satisfies $q(S) \ge 1-\epsilon$ and $q|S$ is a $\frac{\epsilon}{1-\epsilon}$-deletion of p_S. We have

$$\begin{aligned}
\sum_{i\in S} q_i\tau_i &= \sum_{i\in S} q_i(v^\top(X_i - \mu_q))^2 && (180)\\
&= q(S)\mathbb{E}_q[(v^\top(X-\mu_q))^2|S] \\
&= q(S)(\mathbb{E}_q[(v^\top(X-\mu_{q|S}))^2|S] + (v^\top(\mu_q - \mu_{q|S}))^2) && (181)\\
&= q(S)\mathbb{E}_q[(v^\top(X-\mu_{q|S}))^2|S] + q(S)(v^\top(\mu_q - \mu_{q|S}))^2). && (182)
\end{aligned}$$

The term $q(S)(v^\top(\mu_q - \mu_{q|S}))^2$ could be upper bounded by

$$\begin{aligned}
q(S)(v^\top(\mu_q - \mu_{q|S}))^2 &\le q(S)\frac{1-q(S)}{q(S)}\sum_{i=1}^n q_i(v^\top(X_i-\mu_q))^2) && (183)\\
&\le \epsilon\sum_{i=1}^n q_i(v^\top(X_i-\mu_q))^2) && (184)
\end{aligned}$$

following Lemma C.2, $q(S) \geq 1-\epsilon$, and the fact that $q|S$ is a $1-q(S)$ deletion of q. For the first term, we have

$$q(S)\mathbb{E}_q[(v^\top(X-\mu_{q|S}))^2|S] \leq q(S)\mathbb{E}_q[(v^\top(X-\mu_{p_S}))^2|S] \tag{185}$$

$$\leq q(S)\frac{1}{1-(\epsilon/(1-\epsilon))}\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2] \tag{186}$$

$$\leq \frac{1-\epsilon}{1-2\epsilon}\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2], \tag{187}$$

where in (186) we used the inequality $q|S \leq \frac{1}{1-(\epsilon/(1-\epsilon))}p_S$ and $q(S) \leq 1$. Combining these two together, we have

$$\sum_{i\in S} q_i(v^\top(X_i-\mu_q))^2 \leq \frac{1-\epsilon}{1-2\epsilon}\mathbb{E}_{p_S}[(v^\top(X-\mu_{p_S}))^2] + \epsilon\sum_{i=1}^n q_i(v^\top(X_i-\mu_q))^2) \tag{188}$$

$$\leq \frac{1-\epsilon}{1-2\epsilon}\sigma^2 + \epsilon\sum_{i\in[n]} q_i\tau_i \tag{189}$$

$$\leq \frac{1}{2}\sum_{i\in[n]} q_i\tau_i, \tag{190}$$

since we have assumed $\|\Sigma_q\| \geq \sigma'^2$ which implies $\sum_{i\in[n]} q_i\tau_i \geq \frac{2(1-\epsilon)}{(1-2\epsilon)^2}\cdot\sigma^2$. Thus (56) holds. From Lemma 4.4 the error is bounded by (178).

□

E.4 Proof of Theorem 4.3

When both F_1, F_2 are negative, the algorithm will just output $q^{(k)}$ with the desired results. Thus it suffices to show that no matter which cases we are in (either $F_1 \geq 0$ or $F_2 \geq 0$), we will always have the invariance

$$\sum_{i\in S}(\frac{1}{n}-c_i^{(k)}) \leq \sum_{i\in[n]/S}(\frac{1}{n}-c_i^{(k)}). \tag{191}$$

To show the invariance, it suffices to show that under either $F_1 \geq 0$ or $F_1 \leq 0, F_2 \geq 0$, we have

$$\sum_{i\in S} q_i^{(k)}\tau_i^{(k)} \leq \frac{1}{2}\sum_{i=1}^n q_i^{(k)}\tau_i^{(k)}. \tag{192}$$

Now we show the first case: when $F_1 \geq 0$, the invariance holds.

Certifiable hypercontractivity. It follows from the general analysis of filter algorithms (Lemma 4.4) that it suffices to show the implication

$$\sum_{i\in S}\frac{1}{n}-c_i \leq \sum_{i\notin S}\frac{1}{n}-c_i, \quad \kappa'^2\mathbb{E}_q[(v^\top X)^2]^2 \preceq_{\text{sos}} \mathbb{E}_q[(v^\top X)^4]$$

$$\Rightarrow \sum_{i\in S} q_i E_v(v^\top X_i)^4 \leq \frac{1}{2}\sum_{i=1}^n q_i E_v(v^\top X_i)^4. \tag{193}$$

To see the implication holds, observe that

$$\sum_{i\in S} c_i E_v[(v^\top X_i)^4] \overset{(i)}{\le} \frac{\kappa^2}{1-2\kappa^2\epsilon} E_v[(\sum_{i\in S} c_i (v^\top X_i)^2)^2] \tag{194}$$

$$\overset{(ii)}{\le} \frac{\kappa^2}{1-2\kappa^2\epsilon} E_v[(\sum_{i=1}^{n} c_i (v^\top X_i)^2)^2] \tag{195}$$

$$\overset{(iii)}{\le} \frac{1}{2}\sum_{i=1}^{n} c_i E_v[(v^\top X_i)^4]. \tag{196}$$

Here (i) is by Lemma E.1 (and the fact that $E_v[p] \le E_v[q]$ if $p \preceq_{\text{sos}} q$), (ii) is by the fact that adding the $c_i(v^\top X_i)^2$ terms for $i \notin S$ is adding a sum of squares, and (iii) is by the assumption that E refutes hypercontractivity. Thus as long as $\kappa^2\epsilon \le 1$ we have the desired property.

Filtering for bounded noise Next, we show that when $F_1 \le 0, F_2 \ge 0$, the invariance still holds.

From Lemma 4.4, it suffices to show the implication

$$\sum_{i\in S}\frac{1}{n} - c_i \le \sum_{i\notin S}\frac{1}{n} - c_i, \forall v \in \boldsymbol{R}^d, \mathbb{E}_q[(Y - X^\top\theta(q))^2 (v^\top X)^2] \ge \sigma'^2 \mathbb{E}_q[(v^\top X)^2]$$

$$\Rightarrow \sum_{i\in S} q_i (Y_i - X_i^\top\theta(q))^2 (v^\top X_i)^2 \le \frac{1}{2}\sum_{i=1}^{n} q_i (Y_i - X_i^\top\theta(q))^2 (v^\top X_i)^2. \tag{197}$$

Denote $\tilde\sigma^2 = F(q)$. Then the LHS satisfies

$$\begin{aligned}
\sum_{i\in S} q_i (Y_i - X_i^\top\theta(q))^2 (v^\top X_i)^2 &\le 2\sum_{i\in S} q_i ((Y_i - X_i^\top\theta(p_S))^2 (v^\top X_i)^2 + ((\theta(p_S) - \theta(q))^\top X_i)^2 (v^\top X_i)^2) \\
&\overset{(i)}{\le} 2(\frac{1}{1-2\epsilon}\mathbb{E}_{p_S}[(Y - X^\top\theta(p_S))^2 (v^\top X)^2] \\
&\quad + \mathbb{E}_q[((\theta(p_S) - \theta(q))^\top X)^4]^{1/2} \cdot \mathbb{E}_q[(v^\top X)^4]^{1/2}) \\
&\overset{(ii)}{\le} \frac{2}{1-2\epsilon}(\sigma^2 \mathbb{E}_{p_S}[(v^\top X)^2] + 5\kappa'^2 \mathbb{E}_q[((\theta(p_S) - \theta(q))^\top X)^2] \cdot \mathbb{E}_q[(v^\top X)^2]) \\
&\overset{(iii)}{\le} \frac{2}{1-2\epsilon}(\frac{(1-2\epsilon+2\kappa\sqrt{\epsilon(1-\epsilon)})^2\sigma^2}{(1-2\epsilon-2\kappa\sqrt{\epsilon(1-\epsilon)})^2} \\
&\quad + 5\kappa'^2 \mathbb{E}_q[((\theta(p_S) - \theta(q))^\top X)^2])\mathbb{E}_q[(v^\top X)^2] \\
&\overset{(iv)}{\le} \frac{2}{1-2\epsilon}(\frac{(1-2\epsilon+2\kappa\sqrt{\epsilon(1-\epsilon)})^2\sigma^2}{(1-2\epsilon-2\kappa\sqrt{\epsilon(1-\epsilon)})} \\
&\quad + 5\kappa'^2 \mathbb{E}_q[((\theta(p_S) - \theta(q))^\top X)^2])\mathbb{E}_q[(Y - X^\top\theta(q))^2 (v^\top X)^2]/\tilde\sigma^2
\end{aligned} \tag{198}$$

Here (i) comes from that $q_i \le \frac{p_{S,i}}{1-2\epsilon}$ for all $i \in [n]$, (ii) comes from the assumption on p_S and the hypercontractivity of q, (iii) is by Lemma D.3, (iv) comes from the definition of $\tilde\sigma^2$. From (Zhu et al., 2019, Theorem 3.4), we know that when $\epsilon < 1/(1+4\kappa'^2)$

$$\mathbb{E}_q[((\theta(p_S) - \theta(q))^\top X)^2] \le \frac{2\kappa'\tilde\sigma^2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}. \tag{199}$$

Denote $\tilde{\sigma}^2 = \sup_{v\in\boldsymbol{R}^d} \mathbb{E}_q[(Y - X^\top\theta(q))^2(v^\top X)^2]/\mathbb{E}_q[(v^\top X)^2]$. Then overall, we have

$$\begin{aligned}\sum_{i\in S} q_i(Y_i - X_i^\top\theta(q))^2(v^\top X_i)^2 &\le \frac{2}{(1-2\epsilon)\tilde{\sigma}^2}\Big(\frac{(1-2\epsilon+2\kappa\sqrt{\epsilon(1-\epsilon)})^2\sigma^2}{(1-2\epsilon-2\kappa\sqrt{\epsilon(1-\epsilon)})^2}\\ &\quad + \frac{10\kappa'^3\tilde{\sigma}^2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\Big)\mathbb{E}_q[(Y - X^\top\theta(q))^2(v^\top X)^2])\\ &\le \frac{2}{(1-2\epsilon)\sigma'^2}\Big(\frac{(1-2\epsilon+2\kappa\sqrt{\epsilon(1-\epsilon)})^2\sigma^2}{(1-2\epsilon-2\kappa\sqrt{\epsilon(1-\epsilon)})^2}\\ &\quad + \frac{5\kappa'^3\sigma'^2\epsilon(1-\epsilon)}{(1-2\epsilon)^2}\Big)\mathbb{E}_q[(Y - X^\top\theta(q))^2(v^\top X)^2]). \qquad (200)\end{aligned}$$

By taking $\sigma'^2 = \frac{4\sigma^2(1-2\epsilon+2\kappa'\sqrt{\epsilon(1-\epsilon)})}{(1-2\epsilon)^3-20\kappa'^3\epsilon(1-\epsilon)}$, we know that the implication holds.

E.5 Proof of Theorem 4.4

Proof of Theorem 4.4. It follows from the general analysis of filter algorithm (Lemma 4.4) that it suffices to show the implication

$$\begin{aligned}&\sum_{i\in S}\frac{1}{n} - c_i \le \sum_{i\notin S}\frac{1}{n} - c_i, \kappa'^2\mathbb{E}_q[(v^\top(X-\mu_q))^2]^2 \preceq_{\text{sos}} \mathbb{E}_q[(v^\top(X-\mu_q))^4]\\ &\qquad\Rightarrow \sum_{i\in S} q_i E_v(v^\top(X_i-\mu_q))^4 \le \frac{1}{2}\sum_{i=1}^n q_i E_v(v^\top(X_i-\mu_q))^4. \qquad (201)\end{aligned}$$

Observe that when $\kappa^2\epsilon < 1/4$,

$$\begin{aligned}\sum_{i\in S} c_i E_v[(v^\top(X_i-\mu_q))^4] &\le \sum_{i\in S} 8c_i E_v[(v^\top(X_i-\mu_{q|S}))^4 + (v^\top(\mu_q-\mu_{q|S}))^4]\\ &\overset{(i)}{\le} \frac{8\kappa^2}{1-2\kappa^2\epsilon}E_v[(\sum_{i\in S} c_i(v^\top(X_i-\mu_{q|S}))^2)^2] + \frac{32\epsilon(1-\epsilon)^2}{(1-2\epsilon)^2}E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^2]^2]\\ &\overset{(ii)}{\le} \frac{8\kappa^2}{1-2\kappa^2\epsilon}E_v[(\sum_{i=1}^n c_i(v^\top(X_i-\mu_q))^2)^2] + \frac{32\epsilon(1-\epsilon)^2}{(1-2\epsilon)^2}E_v[\mathbb{E}_q[(v^\top(X-\mu_q))^2]^2]\\ &\le (\frac{8\kappa^2}{1-2\kappa^2\epsilon} + \frac{32\epsilon(1-\epsilon)}{(1-2\epsilon)^2})\cdot E_v[(\sum_{i=1}^n c_i(v^\top(X_i-\mu_q))^2)^2]\\ &\overset{(iii)}{\le} \frac{1}{2}\sum_{i=1}^n c_i E_v[(v^\top X_i)^4]. \qquad (202)\end{aligned}$$

Here (i) is by Lemma E.1 and Lemma D.2 (and the fact that $E_v[p] \le E_v[q]$ if $p \preceq_{\text{sos}} q$), (ii) is by the fact that adding the $c_i(v^\top X_i)^2$ terms for $i \notin S$ is adding a sum of squares, and (iii) is by the assumption that E refutes hypercontractivity. Thus as long as $\epsilon < 1/4\kappa^2$ we have the desired property.

□

E.6 Proof of Theorem 4.5

From boundedness of X_i, we know that $g(X_i; q^{(k)}) \le 4d/\epsilon$. Denote $\eta^{(k)} = \delta \cdot \frac{\epsilon}{8d}$. The algorithm has the following regret bound from (Arora et al., 2012, Theorem 2.4):

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{q^{(t)}}[(v^{(t)\top}(X_i-\mu_q))^2-1]-\mathbb{E}_{p_S}[(v^{(t)\top}(X_i-\mu_q))^2-1]\right)\\
\le&\frac{\delta}{2T}\sum_{t=1}^{T}\mathbb{E}_{p_S}[|(v^{(t)\top}(X_i-\mu_q))^2-1|]+\frac{16d}{T\delta}\\
\le&\frac{\delta}{2T}\sum_{t=1}^{T}(\mathbb{E}_{p_S}[(v^{(t)\top}(X_i-\mu_{p_S}))^2]+1+\|\mu_q-\mu_{p_S}\|_2^2)+\frac{16d}{T\delta}\\
\le&\frac{\delta}{2T}\sum_{t=1}^{T}(2+\xi+\|\mu_q-\mu_{p_S}\|_2^2)+\frac{16d}{T\delta}\\
\le&\frac{\delta}{2T}\sum_{t=1}^{T}\|\mu_q-\mu_{p_S}\|_2^2+(1+\xi/2)\delta+\frac{16d}{T\delta}
\end{aligned}\tag{203}$$

By taking $T = T_0 = \frac{8(2+\xi)d}{\xi^2}$ and taking $\delta = \frac{4\beta\sqrt{d}}{\sqrt{(1+\xi/2)T}} = 2\beta\xi/(2+\xi), \beta \in (0,1)$, we have

$$\frac{1}{T_0}\sum_{t=1}^{T_0}\left(\mathbb{E}_{q^{(t)}}[(v^\top(X_i-\mu_q))^2-1]-\mathbb{E}_{p_S}[(v^\top(X_i-\mu_q))^2-1]-\frac{\beta\xi}{2+\xi}\|\mu_q-\mu_{p_S}\|_2^2\right)\le 2\xi/\beta. \tag{204}$$

Thus there must exists some $t_0 \in [T_0]$ such that

$$\begin{aligned}
\mathbb{E}_q[(v^\top(X_i-\mu_q))^2-1] &\le \mathbb{E}_{p_S}[(v^\top(X_i-\mu_q))^2-1]+\frac{\beta\xi}{2+\xi}\|\mu_q-\mu_{p_S}\|_2^2+2\xi/\beta\\
&= \mathbb{E}_{p_S}[(v^\top(X_i-\mu_{p_S}))^2-1]+(1+\frac{\beta\xi}{2+\xi})\cdot\|\mu_q-\mu_{p_S}\|_2^2+2\xi/\beta\\
&\le (1+\frac{\beta\xi}{2+\xi})\cdot\|\mu_{p_S}-\mu_q\|^2+(1+2/\beta)\xi\\
&\le (1+\frac{\beta\xi}{2+\xi})\cdot\left(\frac{\rho}{1-2\epsilon}+\sqrt{\frac{\epsilon(\xi+\xi'+\epsilon/(1-\epsilon))}{1-2\epsilon}+\frac{\epsilon(1-\epsilon)\rho^2}{(1-2\epsilon)^2}}\right)^2+(1+2/\beta)\xi,
\end{aligned}\tag{205}$$

where we denote $\xi' = \max(\|\Sigma_q\|-1, 0)$. The last inequality comes from Lemma D.1. On the other hand, from the choice of v, we have

$$\mathbb{E}_q[(v^\top(X_i-\mu_q))^2-1] \ge \xi'(1-\gamma)-\gamma. \tag{206}$$

By combining the above two inequalies and solve them for ξ', we can see that there exists some constant C,

$$\xi' \le C\cdot\frac{(1+1/\beta)\xi+\rho^2+\epsilon}{(1-3(1+\beta\xi/(1-\gamma\epsilon))\epsilon)^2}. \tag{207}$$