

Fast algorithms for learning a Gaussian under halfspace truncation with optimal sample complexity

Haitong Liu
liuhai@student.ethz.ch
ETH Zurich

Deepak Narayanan
Sridharan
dsridharan@inf.ethz.ch
ETH Zurich

David Steurer
david.steurer@inf.ethz.ch
ETH Zurich

Manuel Wiedmer
manuel.wiedmer@inf.ethz.ch
ETH Zurich

June 25, 2026

Abstract

We study the fundamental problem of learning a high-dimensional Gaussian truncated to an unknown halfspace. Lee, Mehrotra and Zampetakis (FOCS'24) recently obtained the first polynomial time algorithm for this problem, but their resulting sample and time complexity bounds are not optimal. Under non-trivial truncation, for any target accuracy $\varepsilon > 0$ and dimension d we give an efficient algorithm that uses $n = \tilde{O}(d^2/\varepsilon^2)$ samples and learns the underlying Gaussian to error ε in total variation distance. Our algorithm is also fast: its runtime is dominated by the cost of computing the empirical covariance matrix. Both our sample and time complexity are optimal in terms of d and ε even *without* truncation: in this regard, we can learn a Gaussian under halfspace truncation for free.

The key ingredient behind our result is a novel reinterpretation of the low-degree moments of the truncated Gaussian in terms of a relative truncation parameter. This relative truncation parameter uniquely determines the parameters of the untruncated Gaussian and enables direct parameter recovery. This reinterpretation allows us to circumvent the time intensive projected stochastic gradient descent procedure that is widely used in learning under truncation.

Contents

1	Introduction	1
1.1	Main contributions	2
1.2	Related work	5
2	Technical overview	6
3	Preliminaries	12
3.1	Notation	12
3.2	Inverse Mill's ratio	12
3.3	Gaussian concentration and anti-concentration	19
3.4	Concentration bounds for sub-Gaussian random variables	19
4	Properties of halfspace truncated Gaussians	21
4.1	The first three moments	22
4.2	Sub-Gaussianity	23
4.3	Preconditioning	24
5	Proof of the main result	26
5.1	Recovering the direction of Σw	28
5.2	Relative truncation parameter recovery	31
5.3	Relative truncation parameter recovery implies parameter recovery	35
5.4	Proof of Theorem 5.1	42
6	Robustness	50
6.1	Preliminaries on robustness	50
6.2	Learning parameters under halfspace truncation and robustness	51
7	Extension to intersections of two orthogonal halfspaces	57
7.1	Moments	58
7.2	Relative truncation parameter estimation	59
7.3	Proof of Theorem 7.1	62
	Acknowledgments	65
A	Deferred proofs from the main text	65
A.1	Cumulants of the one-dimensional truncated distribution	65
A.2	Matrix and vector calculations	74
A.3	Concentration results and error analysis for sample moments	76
B	Computations	79
C	Determining w	84

1 Introduction

We study statistical estimation under truncation. In this setting, data is only observed when it falls in some survival set. There is a long line of work on learning based on truncated or censored samples; dating back to at least the work of Daniel Bernoulli [Ber60]. Bernoulli studied the effectiveness of the smallpox vaccination and encountered the problem of missing data from people who died of smallpox; thus he only had access to ‘truncated’ data. Since then, statistical estimation from truncated or censored samples has been studied by numerous scientists, including [Gal98], [Pea02], [PL08], [Lee14], [Fis31], and [BS37]; we refer the reader to the book [Coh91] for a more comprehensive treatment. Importantly however, all these works either study the one-dimensional setting or utilize methods that are not computationally efficient in higher dimensions.

In the last decade, the following setting has received renewed attention: Given samples from a multivariate Gaussian with unknown parameters truncated to some set S , learn the Gaussian in total variation (TV) distance. We note that for this, it is enough to learn the mean and covariance up to small error. [DGTZ18] showed that, given oracle access to the truncation set, the parameters can be learned in TV-distance up to error ε in polynomial time and in sample complexity $\tilde{O}(d^2/\varepsilon^2)$. This sample complexity is optimal up to logarithmic factors since already for learning the covariance of a Gaussian, even without truncation, $n = \Omega(d^2/\varepsilon^2)$ samples are information-theoretically necessary (cf. [CZZ10]).

If the truncation set is unknown and *arbitrary*, [DGTZ18] also showed that learning up to arbitrarily small error in TV-distance is impossible. [KTZ19] studied the more tractable setting where the truncation set is unknown but belongs to some well-behaved class of possible truncation sets. They are able to learn the parameters of a Gaussian with *diagonal* covariance matrix given a bound on the Gaussian surface area of the truncation set. If the Gaussian surface area is at most Γ , then they can learn the Gaussian up to TV-error ε with time and sample complexity $d^{\text{poly}(1/\varepsilon)\Gamma^2}$. Without further assumptions on the truncation set, [DKPZ24] showed that the exponential dependence on $1/\varepsilon$ is necessary in the SQ model.

More recently, [LMZ24] extended the result of [KTZ19] to general Gaussians. Furthermore, they showed that for two simple classes of truncation sets, namely, axis-aligned rectangles and halfspaces, one can even learn in time and sample complexity $\text{poly}(d/\varepsilon)$. This is the first algorithm that achieves fully polynomial runtime for learning a Gaussian under truncation to some unknown survival set.

In this work, we continue the investigation of learning from truncated samples for ‘simple’ survival sets. We focus on the case of halfspace truncation. For halfspaces, the result by [LMZ24] needs at least $\Omega(d^3)$ samples and might not be practical. This is because an important step of their algorithm is based on projected stochastic gradient descent, for which they use the ellipsoid method for the projection step. It is important to note that the projection is *not* to the survival set, but instead to some set in the parameter space. Therefore, even for simple classes it is unclear whether this step can be made fast. Concretely, a direct application of the ellipsoid method (without using any specifics about the problem) yields a lower bound of $\Omega(d^{10})$ on the runtime of their algorithm ([LMZ24, Algorithm 3]).¹ Thus, their paper leaves open the following natural question for halfspace truncation:

Are there fast algorithms with optimal sample complexity, matching the untruncated setting?

¹Algorithm 3 in [LMZ24] runs the ellipsoid method on a problem in $m = \Theta(d^2)$ dimensions. The ellipsoid method needs $\Omega(m^2)$ iterations to terminate and each iteration takes $\Omega(m^2)$ operations (as we need to update the $m \times m$ matrix describing the ellipsoid). Thus, one run of the ellipsoid method needs $\Omega((d^2)^4) = \Omega(d^8)$ time. Finally, they run the ellipsoid method $\Omega(d^2)$ times, giving a runtime lower bound of $\Omega(d^{10})$.

Naturally, the best one could hope for is to match the results of the untruncated setting — which has been studied for decades and is well-understood. There, the best known algorithm is to compute the sample mean and the sample covariance; see e.g. [CZZ10]. To achieve error ε , one needs $N = \Theta(d^2/\varepsilon^2)$ many samples. The time needed to compute this estimator is asymptotically the same as the time needed to multiply a $d \times N$ matrix with its transpose, which we denote by $T(N, d)$. Currently, the best-known runtime for this is $O(d^{3.250035}/\varepsilon^2)$ as a result of recent advances in fast matrix multiplication due to [ADVW⁺25].

1.1 Main contributions

In this paper, we answer the above question in the affirmative by designing an algorithm that achieves the following guarantees.

Theorem 1.1 (Informal, see Theorem 5.1). *There is an algorithm with the following guarantee: The algorithm takes as input N samples from a d -dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ truncated to some halfspace of mass α as well as a lower bound α_0 on α . If $N = d^2 \cdot \text{poly}(1/\varepsilon, \log(1/\alpha_0))$, the algorithm outputs $\hat{\mu}$ and $\hat{\Sigma}$ such that, with probability at least 0.99, we have:*

$$d_{\text{TV}}\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})\right) \leq \varepsilon.$$

The time complexity of the algorithm is $O(T(N, d)) \leq O(d^{3.250035}) \cdot \text{poly}(1/\varepsilon, \log(1/\alpha_0))$.

Remark. *If $\alpha \leq 0.99$, then our algorithm needs only $N = d^2/\varepsilon^2 \cdot \text{polylog}(1/\alpha_0)$ samples; otherwise it needs $N = O(d^2/\varepsilon^4)$ many samples. The algorithm is the same for both cases.*

We also note that the ‘0.99’ both in the condition on α and in the success probability can be replaced by arbitrary constants $c < 1$.

Our algorithm is the first algorithm that achieves optimal sample complexity dependence on d . The runtime of our algorithm matches the runtime of computing the sample covariance in the untruncated setting and hence is also optimal in terms of d . Furthermore, the dependence of the sample and time complexity on ε are optimal whenever there is non-trivial truncation (meaning, $\alpha \leq 0.99$).

Regarding our dependence of the sample complexity on α_0 , we remark that all prior works, including inefficient algorithms, also have a dependence on α_0 . In fact, under the assumption that the set is unknown, all previous *efficient* algorithms incur either an exponential dependence on $1/\alpha_0$ or work under the assumption that $\alpha_0 \geq \Omega(1)$ (information-theoretically, [KTZ19] showed that one can achieve polynomial dependence). Interestingly, in the case where the set is known, more recently, [KKT25] showed that one can achieve $\text{polylog}(1/\alpha_0)$ dependence on α_0 for the sample complexity (their runtime dependence on $1/\alpha_0$ is polynomial). In the case of halfspaces, our algorithm matches this guarantee even when the set is *unknown*. This is the first efficient algorithm to get better than exponential dependence on $1/\alpha_0$ when the set is unknown. An overview of the known algorithms for learning a Gaussian under halfspace truncation can be found in Table 1. We note that all of these cited results are more general than halfspaces; in order to be able to compare them to our result, the results mentioned in the table are specialized to the case of halfspace truncation.

Table 1: Overview of efficient algorithms for learning a Gaussian under halfspace truncation

Paper	Truncation set (known/unknown)	Underlying distribution	Sample complexity (in d and ε)	Dependence on α_0
[DGTZ18]	Known	$\mathcal{N}(\mu, \Sigma)$	$\tilde{O}(d^2/\varepsilon^2)$	Assumption: $\alpha_0 \geq \Omega(1)$
[KTZ19]	Unknown	$\mathcal{N}(\mu, \Sigma)$ for Σ diagonal	$d^{\text{poly}(1/\varepsilon)}$	$\exp(1/\alpha_0)$
[LMZ24]	Unknown	$\mathcal{N}(\mu, \Sigma)$	$\text{poly}(d, 1/\varepsilon)$	$\exp(1/\alpha_0)$
[KKT25]	Known	$\mathcal{N}(\mu, \Sigma)$	$O(d^2/\varepsilon^2)$	$\text{poly log}(1/\alpha_0)$
This work	Unknown	$\mathcal{N}(\mu, \Sigma)$	$O(d^2/\varepsilon^2)$ if $\alpha \leq 0.99$ $O(d^2/\varepsilon^4)$ otherwise	$\text{poly log}(1/\alpha_0)$

High-level overview of our techniques and comparison to prior work. At a high level, our algorithm is based on the method of moments. We use the first three moments of the distribution $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$ ² to estimate a *relative truncation parameter*. Our key insight is that this relative truncation parameter *fully* characterizes how truncation changes the moments compared to the untruncated distribution. This enables us to directly estimate the parameters μ and Σ of the Gaussian from the moments themselves.

We now compare our approach to [LMZ24], who gave the first efficient algorithm for learning the parameters of a Gaussian under halfspace truncation. They use the method of moments to first learn the *truncation set*. Then, they apply projected stochastic gradient descent to estimate the parameters. At a high level, their proof can be divided into the following four steps:

1. First, they use the third central moment to find a vector close to $u = \frac{\Sigma w}{\|\Sigma w\|}$.
2. Then, they compute a vector close to the vector w of the halfspace.³
3. Next, they find a value close to τ by picking the smallest threshold such that the resulting set contains all the samples.
4. Finally, given the estimate for the truncation set, they use their general projected stochastic gradient descent procedure to learn the parameters.

While our approach is also based on the method of moments, we use the moments for a different purpose: rather than estimating the truncation set as in [LMZ24], we use the moments to *directly* estimate the parameters of the underlying Gaussian. By using the method of moments for direct parameter recovery, we are able to get a fast algorithm: crucially, we can avoid the time-consuming projected stochastic gradient step (step 4) and achieve optimal runtime in terms of d .

²Here, $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$ denotes the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ truncated to the halfspace $\{x : \langle w, x \rangle \leq \tau\}$ for some unit vector w .

³In their paper, [LMZ24] claim that the third central moment M_3 is proportional to

$$M_3 := \mathbb{E}[(X - \mathbb{E}[X])^{\otimes 3}] \propto w^{\otimes 3}.$$

They use this to determine w . However, there is small gap in this argument: As we show later, the third central moment is in fact proportional to $(\Sigma w)^{\otimes 3}$ (and not $w^{\otimes 3}$); see Lemma 4.2. Since our goal is to estimate Σ , the third central moment therefore *cannot* be directly used to estimate w . Instead, one can compute a vector u proportional to Σw using the third central moment. Then, one can solve the following linear system involving the second central moment M_2 to recover w . Namely, it holds that

$$M_2^{-1}u \propto w.$$

We include a proof of this for completeness in Appendix C.

As a first step in our algorithm, we also estimate a vector \hat{u} as [LMZ24] do in step 1. However, we then do not use it to estimate the vector w . Instead, the key insight in our algorithm is that the first non-central moment and the second and third central moments all share an important structural property: They are equal to the untruncated moments plus an additional term involving *only* the following vector $\Sigma w / \|\Sigma^{1/2} w\|$ and the *same* relative truncation parameter γ . Importantly, we show that the first three moments *uniquely* determine γ and $\|\Sigma w\| / \|\Sigma^{1/2} w\|$. This allows us to use the moments to get an estimate for γ and $\|\Sigma w\| / \|\Sigma^{1/2} w\|$. Combining this with our first step enables us to directly compute estimates for the parameters. Together, these ideas yield a simple and fast algorithm for estimating the parameters of a halfspace truncated Gaussian.

Regarding the sample complexity, [LMZ24] require a high-accuracy estimate of the third moment in their step 2. This step requires at least $\Omega(d^3)$ samples. In contrast, for our algorithm, when computing a vector close to $u = \Sigma w / \|\Sigma w\|$, we use a random contraction of the third central moment and then use a spectral decomposition of the resulting matrix, which allows us to get optimal sample complexity.

Robustly learning the parameters of a Gaussian under halfspace truncation. In contrast to the algorithm by [LMZ24], we utilize the moments of the truncated distribution to directly estimate the parameters. Motivated by this, we introduce the problem of *robust estimation of a truncated Gaussian*. We consider the following corruption model: First, samples from a halfspace truncated Gaussian $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$ are drawn. Then, an adversary is allowed to modify an η -fraction of the samples arbitrarily; in particular the samples *do not* need to lie in the set $\{x \in \mathbb{R}^d \mid \langle w, x \rangle \leq \tau\}$ anymore. The adversary is computationally unbounded, knows the truncation set as well as the parameters and can adaptively change the samples after they have been drawn. We believe that this corruption model, which is analogous to the strong contamination model in robust statistics [DK23], might be of independent interest for learning under truncation.

Given that our algorithm only uses the first three moments, we can robustify it by using robust estimates for the moments instead of relying on the sample moments. Our robust algorithm uses a black-box robust estimator due to [KSS18] and the fact that the truncated distribution is sub-Gaussian. We prove the following result. To simplify the exposition, we state our result for constant α and for $\Sigma \geq \Omega(1) \cdot I_d$. The full result can be found in [Theorem 6.8](#) in [Section 6](#).

Theorem 1.2 (Informal, see [Theorem 6.8](#)). *Consider the truncated distribution $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, where the mass α of the survival set satisfies $\Omega(1) \leq \alpha \leq 0.99$. Assume $\Sigma \geq \Omega(1) \cdot I_d$. Let $0 < \delta < 1/2$ be an arbitrary constant. Let $\eta > 0$ be the corruption parameter and assume that $\eta \leq 1/d^{1-2\delta}$. For $n = \text{poly}(d)$, given an η -corruption X_1, \dots, X_n of $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, we can compute, in polynomial time in d and with probability at least 0.99, estimators $\hat{\mu}$ and $\hat{\Sigma}$ such that*

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq O_\delta \left(\sqrt{d} \cdot \eta^{1/2-\delta} \right)$$

and

$$\|\Sigma^{-1/2} \hat{\Sigma}^{-1/2} - I_d\|_F \leq O_\delta \left(\sqrt{d} \cdot \eta^{1/2-\delta} \right).$$

While our algorithm can only tolerate a very small amount of corruption, we believe that already this is non-trivial because we are *not* trying to estimate the moments of the distribution to which we have (sample) access to (which would be the typical goal in robust statistics) but instead of the underlying untruncated distribution. In particular, the adversary can place samples outside the survival set. Thus, it is unclear if one can even estimate the truncation set. We remark that previous work in truncated statistics cannot (easily) be made robust. For example, even if one

could learn the truncation set under adversarial corruption, it is unclear whether the algorithm of [LMZ24] can be made to succeed, since it is not clear how to robustify their final step involving projected stochastic gradient descent.

That being said, we believe that it is likely that the amount of corruption and the error rate can be improved. We leave it as an open question for future work whether using a more specialized robust estimator for the moments can improve Theorem 1.2. We believe that our result can act as a starting point for further investigation of robust estimation given only truncated samples. In particular, our result also leaves open what happens in weaker corruption models (e.g., a model where only oblivious corruptions are allowed).

Extensions to other truncation sets. Finally, we show that our techniques can be used beyond halfspace truncation. In Theorem 7.1, we show that using similar ideas we can also learn the mean of an identity-covariance Gaussian $\mathcal{N}(\mu, I_d)$ truncated to an *intersection of two orthogonal halfspaces*. Similar to our main result, we can achieve sample complexity $O\left(\frac{d^2}{\varepsilon^2}\right) \cdot \text{polylog}(1/\alpha_0) + O\left(\frac{d^2}{\varepsilon^4}\right)$. Our algorithm for this case also works using only the first three moments of the truncated distribution. At a very high level, instead of recovering one relative truncation parameter, for the case of truncation by an intersection of two orthogonal halfspaces, we recover two relative truncation parameters; one for each truncation direction. For details, we refer to Section 7.

While we only prove the case of truncation by an intersection of 2 orthogonal halfspaces, we believe that our proof can be extended to an intersection of k (pairwise) orthogonal halfspaces, yielding a sample complexity of $O\left(\frac{d^2 \cdot \text{poly}(k)}{\varepsilon^2}\right) \cdot \text{polylog}(1/\alpha_0) + O\left(\frac{d^2 \cdot \text{poly}(k)}{\varepsilon^4}\right)$; see Remark 7.2.

1.2 Related work

Method of Moments. Solving problems in statistical estimation using the Method of Moments (MoM) has a long history. It can be dated back to the famous statistician Karl Pearson [Pea94]. This method has led to several recent advances in high-dimensional parameter estimation tasks, including those involving learning mixture models [MV10, LL22] and independent component analysis (ICA) [BRV13, BCV14]. In this work we adopt the MoM for learning the parameters of a halfspace truncated Gaussian.

Truncated Learning. Beyond the results described earlier, estimation under truncation has seen applications to *differentially private* statistical estimation [ZZ25]. There has also been recent work on mean testing under truncation [CGWY25] and on detecting truncation [DNS23, DLNS24]. A recent work [LMZ26] studies learning under positive and imperfectly labeled samples and extends the general learning algorithm of [LMZ24] for learning under general truncation sets.

Robust Estimation. Robust parameter estimation has a long history in statistics [Hub92, HRVF04]. Recent algorithmic breakthroughs by [LRV16, DKK⁺19] have led to the new subfield of algorithmic robust statistics, which primarily deals with parameter estimation in the presence of adversarial outliers (see the textbook [DK23] for a comprehensive treatment). A very useful method that has emerged for robust statistical estimation is the powerful Sum-of-Squares hierarchy of semi-definite programs [BS16] that often efficiently achieves the best possible statistical guarantees for a large range of robust estimation tasks (see e.g. [KSS18]). In this work, we use Sum-of-Squares based robust moment estimation algorithms as a black-box to solve the robust version of the learning under truncation problem that we introduce.

2 Technical overview

In order to prove our main result [Theorem 1.1](#), we want to use the moments of the truncated distribution to estimate the parameters. Consider the distribution $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$, where w is a unit vector. Recall that our goal is to find estimates $\hat{\mu}$ and $\hat{\Sigma}$ such that

$$d_{\text{TV}}\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})\right) \leq \varepsilon.$$

To show closeness in total variation distance as above, it is enough to show

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq \varepsilon \quad \text{and} \quad \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d\|_F \leq \varepsilon,$$

see e.g. [\[DKK⁺19\]](#). To achieve this, on a high level, we first determine the moments of the truncated distribution. Then, we observe that the moments are determined by μ and Σ as well as a *relative truncation parameter* γ and a certain vector. We show that if we can determine this parameter and vector, then based on the first non-central and the second central moment of the truncated distribution, we would be able to determine the μ and Σ . Finally, we show how to determine γ and this vector based on the second and third central moment.

The moments of the truncated distribution. As a warm up, we start by considering the truncated distribution $Y \sim \mathcal{N}(0, I_d)|_{\{e_1^\top x \leq \gamma\}}$. Intuitively, the moments of this distribution should be the ones of the standard Gaussian, but shifted in the direction of $e_1^{\otimes k}$ since e_1 is the direction in which the distribution is truncated.⁴ For example, $\mathbb{E}[Y] = c_1 e_1$ since in all but the first coordinate, Y is a standard Gaussian. More precisely, $c_1 = \kappa_1(\gamma)$ is the first cumulant⁵ of a one-dimensional standard Gaussian truncated to $(-\infty, \gamma]$. More generally, higher moments can be related to the higher order cumulants of this one-dimensional distribution.

To compute the mean (and more generally the moments) of a general truncated Gaussian $X \sim \mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$, our goal is to relate it to the mean (and the corresponding moments) of Y . To this end, consider the linear transformation

$$y \mapsto \psi(y) = \mu + \Sigma^{1/2}Uy,$$

where U is a rotation matrix. For any U , this map transforms a standard Gaussian $\mathcal{N}(0, I_d)$ to $\mathcal{N}(\mu, \Sigma)$. We want to pick U and γ in a way such that the truncation set of $\psi(Y)$ is equal to the truncation set of X . This will show that $\psi(Y)$ and X have the same distribution, thus allowing us to express the moments of X in terms of the moments of Y .

First, we show how to pick U such that the direction of truncation of $\psi(Y)$ is equal to w . Since Y is truncated to $\{x \mid e_1^\top x \leq \gamma\}$, $\psi(Y)$ is truncated to the set

$$\begin{aligned} \{\psi(y) \mid y \text{ such that } e_1^\top y \leq \gamma\} &= \{x \mid e_1^\top \psi^{-1}(x) \leq \gamma\} \\ &= \{x \mid e_1^\top U^{-1}\Sigma^{-1/2}(x - \mu) \leq \gamma\} \\ &= \{x \mid e_1^\top U^\top \Sigma^{-1/2}x \leq \gamma + e_1^\top U^\top \Sigma^{-1/2}\mu\} \end{aligned} \tag{2.1}$$

⁴We remark that this is only true since the coordinates are independent. As we will see later, for $\mathcal{N}(\mu, \Sigma)|_{\{e_1^\top x \leq \gamma\}}$, the moments would be shifted in the direction of $(\Sigma e_1)^{\otimes k}$ and not $e_1^{\otimes k}$.

⁵The first and second cumulants are just the mean and the covariance. The third cumulant is the third central moment. For more details, see [Definition 3.1](#).

Thus, we want to pick U such that

$$\Sigma^{-1/2}Ue_1 \propto w \iff Ue_1 \propto \Sigma^{1/2}w.$$

As a rotation matrix, U preserves the norm and hence

$$Ue_1 = \frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|}. \quad (2.2)$$

In order to determine the value of γ needed such that $\psi(Y)$ has the same distribution as X , we want to write (2.1) in the form $\{x \mid w^\top x \leq \tau\}$. By (2.2), we can write the truncation set for Y as

$$\{\psi(y) \mid y \text{ such that } e_1^\top y \leq \gamma\} = \{x \mid w^\top x \leq \|\Sigma^{1/2}w\|\gamma + w^\top \mu\}.$$

Hence, we need that

$$\gamma = \frac{\tau - w^\top \mu}{\|\Sigma^{1/2}w\|}. \quad (2.3)$$

For U as in (2.2) and γ as in (2.3), $\psi(Y) \sim \mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$. In particular, we can now compute $\mathbb{E}[X]$ using this transformation ψ and $\mathbb{E}[Y]$:

$$\mathbb{E}[X] = \mu + \Sigma^{1/2}U\mathbb{E}[Y] = \mu + \kappa_1(\gamma) \cdot \Sigma^{1/2}Ue_1 = \mu + \kappa_1(\gamma) \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|}.$$

Analogously we can also compute the higher moments of Y and use them, together with the the transformation ψ to compute the moments of X . Namely, we have that

$$\mathbb{E}[X] = \mu + \kappa_1(\gamma) \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|} \quad (2.4)$$

$$\mathbb{E}[(X - \mathbb{E}[X])^{\otimes 2}] = \Sigma - (1 - \kappa_2(\gamma)) \cdot \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^{\otimes 2} \quad (2.5)$$

$$\mathbb{E}[(X - \mathbb{E}[X])^{\otimes 3}] = \kappa_3(\gamma) \cdot \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^{\otimes 3} \quad (2.6)$$

For more details on this computation, see [Lemma 4.2](#). In particular, these moments are determined by the parameters of the untruncated Gaussian, the vector $\Sigma w / \|\Sigma^{1/2}w\|$ and γ defined as in (2.3). We will refer to $\gamma = \gamma(\mu, \Sigma, w, \tau)$ as the *relative truncation parameter* of the distribution $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$. In particular, since the third moment is rank 1, if we would know the relative truncation parameter, then we could use (2.6) to estimate $\Sigma w / \|\Sigma^{1/2}w\|$. Once we determined this vector, we could then use (2.4) and (2.5) respectively to estimate μ and Σ . To summarize, since we can approximate the moments of $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$ by the sample moments, the main question is how to estimate the relative truncation parameter defined in (2.3).

First three moments determine the relative truncation parameter. Before showing how we estimate the relative truncation parameter γ , we want to establish that the information in the first three moments is in fact enough to uniquely determine γ . For this, consider two truncated Gaussians $\mathcal{N}(\mu_1, \Sigma_1)|_{\{w_1^\top x \leq \tau_1\}}$ and $\mathcal{N}(\mu_2, \Sigma_2)|_{\{w_2^\top x \leq \tau_2\}}$. We want to show that if the first non-central moment as well as the second and third central moments of these two distributions are identical (cf. (2.4), (2.5) and (2.6)), then we need to have $\gamma_1 = \gamma_2$. This will show that it is information-theoretically possible to determine the relative truncation parameter based only on these low-degree moments.

To formally show that the first three moments of the two distributions $\mathcal{N}(\mu_1, \Sigma_1)|_{\{w_1^\top x \leq \tau_1\}}$ and $\mathcal{N}(\mu_2, \Sigma_2)|_{\{w_2^\top x \leq \tau_2\}}$ being equal implies that $\gamma_1 = \gamma_2$, consider first the second moment. By assumption, we have that

$$\Sigma_1 - (1 - \kappa_2(\gamma_1)) \cdot \left(\frac{\Sigma_1 w_1}{\|\Sigma_1^{1/2} w_1\|} \right)^{\otimes 2} = \Sigma_2 - (1 - \kappa_2(\gamma_2)) \cdot \left(\frac{\Sigma_2 w_2}{\|\Sigma_2^{1/2} w_2\|} \right)^{\otimes 2}$$

By multiplying from the left by w_1^\top and from the right by w_2 as well as identifying $a^{\otimes 2}$ with aa^\top , we get that

$$w_1^\top \Sigma_1 w_2 - \frac{(1 - \kappa_2(\gamma_1))}{\|\Sigma_1^{1/2} w_1\|^2} \cdot w_1^\top \Sigma_1 w_1 \cdot w_1^\top \Sigma_1 w_2 = w_1^\top \Sigma_2 w_2 - \frac{(1 - \kappa_2(\gamma_2))}{\|\Sigma_2^{1/2} w_2\|^2} \cdot w_1^\top \Sigma_2 w_2 \cdot w_2^\top \Sigma_2 w_2.$$

Using that $w_i^\top \Sigma_i w_i = \|\Sigma_i^{1/2} w_i\|^2$, the above simplifies to

$$\kappa_2(\gamma_1) \cdot w_1^\top \Sigma_1 w_2 = \kappa_2(\gamma_2) \cdot w_1^\top \Sigma_2 w_2. \quad (2.7)$$

Our goal now is to relate the two numbers $w_1^\top \Sigma_1 w_2$ and $w_1^\top \Sigma_2 w_2$. We do so using the third central moment (cf. (2.6)). Since we assumed that the third moments also agrees, we have that

$$\kappa_3(\gamma_1) \cdot \left(\frac{\Sigma_1 w_1}{\|\Sigma_1^{1/2} w_1\|} \right)^{\otimes 3} = \kappa_3(\gamma_2) \cdot \left(\frac{\Sigma_2 w_2}{\|\Sigma_2^{1/2} w_2\|} \right)^{\otimes 3}$$

This is a rank-1 tensor, so by considering the diagonal elements, we get that also the corresponding vectors need to agree, i.e., that

$$\frac{\Sigma_1 w_1}{\|\Sigma_1^{1/2} w_1\|} = \left(\frac{\kappa_3(\gamma_2)}{\kappa_3(\gamma_1)} \right)^{1/3} \cdot \frac{\Sigma_2 w_2}{\|\Sigma_2^{1/2} w_2\|}.$$

By multiplying by w_2^\top from the left, we can get an expression for $w_1^\top \Sigma_1 w_2 = w_2^\top \Sigma_1 w_1$. By multiplying instead by w_1^\top from the left, we can get an expression for $w_1^\top \Sigma_2 w_2$. Using again that $w_i^\top \Sigma_i w_i = \|\Sigma_i^{1/2} w_i\|^2$ and combining these we get that

$$\left(\frac{\kappa_3(\gamma_1)}{\kappa_3(\gamma_2)} \right)^{1/3} w_2^\top \Sigma_1 w_1 = \|\Sigma_1^{1/2} w_1\| \cdot \|\Sigma_1^{1/2} w_2\| = \left(\frac{\kappa_3(\gamma_2)}{\kappa_3(\gamma_1)} \right)^{1/3} w_1^\top \Sigma_2 w_2.$$

We can now use this equality in (2.7) to get

$$\frac{\kappa_2(\gamma_1)}{\kappa_3(\gamma_1)^{2/3}} = \frac{\kappa_2(\gamma_2)}{\kappa_3(\gamma_2)^{2/3}} \iff \frac{\kappa_3(\gamma_1)}{\kappa_2(\gamma_1)^{3/2}} = \frac{\kappa_3(\gamma_2)}{\kappa_2(\gamma_2)^{3/2}}.$$

We now observe that the function $\kappa_3(\gamma)/(\kappa_2(\gamma))^{3/2}$ is the *skewness* of the one-dimensional standard Gaussian truncated to $(-\infty, \gamma]$. We denote this function by $\text{skew}(\gamma)$. Thus, we have that if the first three moments of $\mathcal{N}(\mu_1, \Sigma_1)|_{\{w_1^\top x \leq \tau_1\}}$ and $\mathcal{N}(\mu_2, \Sigma_2)|_{\{w_2^\top x \leq \tau_2\}}$ agree, then we need to have

$$\text{skew}(\gamma_1) = \text{skew}(\gamma_2).$$

The key argument in this proof now is that for our specific case this function skew is *monotonic* (cf. Lemma 3.17) and thus the above implies that also $\gamma_1 = \gamma_2$.

Estimating the relative truncation parameter algorithmically. We now want to show how we can use the proof above to determine the parameters μ and Σ given samples from the distribution $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$. We showed above that the first three moments uniquely determine $\gamma = \frac{\tau - w^\top \mu}{\|\Sigma^{1/2} w\|}$. The above also lays down a clear path: first, estimate γ and then determine the parameters μ and Σ . To do this, we first compute $\text{skew}(\gamma)$ based on the moments. Then, using the fact that skew is monotonic, we can invert this *one-dimensional* function and recover the relative truncation parameter. This then allows us to use (2.4) and (2.5) to recover the parameters.

It thus remains to show how to determine γ algorithmically based on the first three moments. Define the vector

$$u^* = \kappa_3(\gamma)^{1/3} \cdot \frac{\Sigma w}{\|\Sigma^{1/2} w\|}.$$

The third central moment equals $u^{*\otimes 3}$ and hence, we can use it to obtain a unit vector \tilde{u} in the direction of u^* , that is, $\tilde{u} = (\Sigma w)/\|\Sigma w\|$. By computing the inner product of the third central moment with the tensor $\tilde{u}^{\otimes 3}$, we can even get the norm of u^* since

$$\langle \tilde{u}^{\otimes 3}, \mathbb{E}[(X - \mathbb{E}[X])^{\otimes 3}] \rangle = \langle \tilde{u}, u^* \rangle^3 = \kappa_3(\gamma) \cdot \frac{\|\Sigma w\|^3}{\|\Sigma^{1/2} w\|^3}. \quad (2.8)$$

We now want to use the second central moment to compute a similar expression involving the same norms but $\kappa_2(\gamma)$ instead of $\kappa_3(\gamma)$. Combining this with the above then allows us to compute $\text{skew}(\gamma)$. It turns out that we can do so by computing $\tilde{u}^\top M_2^{-1} \tilde{u}$, where M_2 is the second central moment. Since M_2 differs from Σ only by a rank-1 term, we can use the Sherman-Morrison formula to compute its inverse

$$\begin{aligned} M_2^{-1} &= \Sigma^{-1} + (1 - \kappa_2(\gamma)) \cdot \frac{\Sigma^{-1} \frac{\Sigma w}{\|\Sigma^{1/2} w\|} \cdot \frac{w^\top \Sigma}{\|\Sigma^{1/2} w\|} \Sigma^{-1}}{1 - (1 - \kappa_2(\gamma)) \cdot \frac{w^\top \Sigma}{\|\Sigma^{1/2} w\|} \Sigma^{-1} \frac{\Sigma w}{\|\Sigma^{1/2} w\|}} \\ &= \Sigma^{-1} + (1 - \kappa_2(\gamma)) \cdot \frac{w w^\top}{\|\Sigma^{1/2} w\|^2 \cdot \kappa_2(\gamma)}. \end{aligned}$$

Now, left-multiplying by \tilde{u}^\top and right-multiplying by \tilde{u} , we get that

$$\begin{aligned} \tilde{u}^\top M_2^{-1} \tilde{u} &= \frac{w^\top \Sigma \Sigma^{-1} \Sigma w}{\|\Sigma w\|^2} + (1 - \kappa_2(\gamma)) \cdot \frac{w^\top \Sigma w w^\top \Sigma w}{\|\Sigma w\|^2 \cdot \|\Sigma^{1/2} w\|^2 \cdot \kappa_2(\gamma)} \\ &= \frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2} + (1 - \kappa_2(\gamma)) \cdot \frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)} \\ &= \frac{1}{\kappa_2(\gamma)} \cdot \frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2}. \end{aligned} \quad (2.9)$$

Now, combining (2.9) with (2.8), we get that

$$\langle \tilde{u}^{\otimes 3}, \mathbb{E}[(X - \mathbb{E}[X])^{\otimes 3}] \rangle \cdot (\tilde{u}^\top M_2^{-1} \tilde{u})^{3/2} = \frac{\kappa_3(\gamma)}{\kappa_2(\gamma)^{3/2}} = \text{skew}(\gamma).$$

We remark that we can compute both quantities on the left-hand side of the above equation using \tilde{u} (that we computed using the third moment) and the moments. Hence, we can compute $\text{skew}(\gamma)$.

We now crucially use the fact that skew is monotonic, which allows us to recover γ . As already discussed before, this allows us to recover the parameters: we can use the third moment to recover the vector $(\Sigma w)/\|\Sigma^{1/2}w\|$ (which we do implicitly in our algorithm) and then use the first and second moment to recover μ and Σ . Our algorithm is summarized in [Algorithm 1](#).

Algorithm 1: Parameter estimation of a halfspace-truncated Gaussian based on first three moments

Input: First non-central moment μ_t , second central moment M_2 and third central moment M_3 of the distribution $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$

Output: Mean μ and covariance Σ .

1. Compute \tilde{u} from the third central moment M_3 .
 2. Compute the relative truncation parameter: $\gamma = \text{skew}^{-1} \left((\tilde{u}^\top M_2^{-1} \tilde{u})^{3/2} \cdot \langle \tilde{u}^{\otimes 3}, M_3 \rangle \right)$.
 3. Compute the mean of the Gaussian: $\mu = \mu_t - \frac{\kappa_1(\gamma)}{\sqrt{\kappa_2(\gamma)}} (\tilde{u}^\top M_2^{-1} \tilde{u})^{-1/2} \tilde{u}$.
 4. Compute the covariance of the Gaussian: $\Sigma = M_2 + \left(\frac{1}{\kappa_2(\gamma)} - 1 \right) \cdot (\tilde{u}^\top M_2^{-1} \tilde{u})^{-1} \tilde{u} \tilde{u}^\top$.
-

We emphasize here that our particular choice of expressing both the second and third moment in (2.5) and (2.6) in terms of the *same* relative truncation parameter γ is precisely what enables *direct* parameter recovery without having to learn the truncation set — deviating from existing approaches in the truncated statistics literature [DGTZ18, LMZ24]. An important upshot is that this also enables us to design *simple* and *fast* algorithms, making learning under truncation potentially practical.

Sample complexity. In [Section 5](#), we will show how to implement the above strategy when only having access to the sample moments. We now give a brief overview how the errors in estimating the sample moments influence the sample complexity of our algorithm.

The main bottlenecks for the sample complexity are the approximation \hat{u} to \tilde{u} and the approximation \widehat{M}_2 to M_2 . We need to ensure that the expressions in (2.8) and (2.9) are close to their true values. If we have sufficiently good approximations to these, then with $O_\alpha(d^2/\varepsilon^2)$ extra samples we can get good approximations to the mean and covariance of the truncated distribution and use the strategy described above to get estimators for μ and Σ with error ε . We show in [Section 5](#) that it is sufficient to get estimates

$$\left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\|_2 \leq O_\alpha(\varepsilon) \quad \text{and} \quad \|M_2^{-1/2} \widehat{M}_2 M_2^{-1/2} - I_d\|_F \leq O_\alpha(\varepsilon). \quad (2.10)$$

We note that our algorithm contains a preconditioning step that allows us to reduce our problem to a well-conditioned one. This allows us to invert the matrix \widehat{M}_2 without incurring an error that depends on the condition number of Σ .

Achieving the estimate of \widehat{M}_2 as in (2.10) is possible with $O_\alpha(d^2/\varepsilon^2)$ samples by standard concentration bounds. For the estimate of \tilde{u} , we could estimate it by estimating the diagonal elements of the third moment tensor. However, the diagonal elements of the third moment tensor are $(u_i^*)^3$ (and not u_i^*). Using this strategy would lead to suboptimal sample complexity in d since we would need a very good approximation to each entry $(u_i^*)^3$. This is because $(u_i^*)^3$ can be small and thus taking the cube-root can amplify the error.

To improve the sample complexity, we instead compute a random contraction of the third moment tensor.⁶ Then, on the resulting matrix we can compute the largest eigenvalue (in absolute value) and compute the corresponding vector. This vector is close to \tilde{u} since in expectation (over the samples) this random contraction will yield a rank one matrix proportional to $\tilde{u}\tilde{u}^\top$. We show in [Lemmas 5.5](#) and [5.8](#) that this strategy allows us to get an estimate \hat{u} satisfying the guarantee in [\(2.10\)](#) using $O_\alpha(d^2/\varepsilon^2)$ samples. This allows us to get optimal sample complexity in d .

Putting it together, we show that using $\frac{d^2}{\varepsilon^2} \cdot \text{poly}(1/\kappa_2(\gamma), 1/\kappa_3(\gamma))$ samples, we can get estimates $\hat{\mu}$ and $\hat{\Sigma}$ such that

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq \varepsilon \quad \text{and} \quad \left\| I_d - \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} \right\|_F \leq \varepsilon.$$

We show in [Section 3.2](#) that for $\alpha \leq 0.99$ we can bound $1/\kappa_2(\gamma), 1/\kappa_3(\gamma) \leq \text{poly} \log(1/\alpha_0)$. For $\alpha \rightarrow 1$, $\kappa_3(\gamma) \rightarrow 0$, which leads to an increase in the sample complexity. If α is very close to 1 (meaning $\alpha \geq 1 - \tilde{O}(\varepsilon)$), then the sample mean and sample covariance of the truncated distribution are good enough estimates and we show that our estimators are close to these. For the in-between case (i.e., $0.99 \leq \alpha < 1 - \tilde{O}(\varepsilon)$), we can bound $\kappa_3(\gamma) \geq \Omega(\varepsilon)$, which gives the sample complexity $O(d^2/\varepsilon^4)$. This shows the sample complexity as described in [Theorem 1.1](#).

The fact that we need more samples for the case $\alpha > 0.99$ seems somewhat artificial. The problem is that, when α is close to 1, the norm of the third central moment tensor M_3 goes to 0, which makes it harder to estimate \tilde{u} . On the other hand it seems that the problem of estimating the parameters of a truncated Gaussian should become easier if there is less truncation. However, it is unclear to us whether and how the sample complexity could be improved with our approach.

Runtime. We now briefly discuss the runtime of our algorithm. As a first step, we need to compute the vector \hat{u} . For this, we need to compute a random contraction of the third central sample moment tensor. If we first perform the contraction and then average over the samples, we need to add N rank-1 matrices of size d^2 , which can be done in time $O(T(N, d))$ using fast rectangular matrix multiplication. Next, we need to compute the spectral decomposition of the resulting matrix, which can be done in time $O(d^3)$. For computing the estimate \hat{M}_2 we need to compute the sample covariance, which again can be done in time $O(T(N, d))$. We then need to invert the matrix \hat{M}_2 , which can be done in time $O(d^3)$. Next, we compute the expressions in [\(2.8\)](#) and [\(2.9\)](#). If we again first compute the inner product and then average over the samples (for [\(2.8\)](#)), these can be done in time $O(N \cdot d)$ and $O(d^3)$ respectively. Since $T(N, d) \geq \Omega(N \cdot d) \geq \Omega(d^3)$, the terms involving $T(N, d)$ dominate and we can thus compute an estimate for $\text{skew}(\gamma)$ in time $O(T(N, d))$.

Once we know an estimate of $\text{skew}(\gamma)$, we use the inverse of skew to compute $\hat{\gamma}$. We note that since this is a one-dimensional monotonic function, we can compute an approximate inverse using binary search. The computational complexity of this step is negligible relative to the primary calculations (see [Remark 5.16](#)). The final step requires adding a single term to the first and second moments, which is achievable in time $O(d^2)$.

Hence, the dominating term in the overall runtime is the computation of the sample covariance, which, as argued in [Section 1](#), is $O(T(N, d)) \leq O(d^{3.250035}) \cdot \text{poly}(1/\varepsilon, \log(1/\alpha_0))$.

Organization. In [Section 3](#), we cover the necessary preliminaries for our paper and in particular prove that the function skew is monotonic. In [Section 4](#), we compute the moments of the truncated

⁶Random contractions are typically used in decomposition of higher-order tensors and have seen applications in learning Gaussian Mixture Models, Independent Component Analysis and Dictionary Learning. The specific type of random contraction we utilize is attributed to Jennrich [[Har70](#), [LRA93](#)].

distribution and discuss our preconditioning. Then, in [Section 5](#), we give a complete proof of our main result. In [Section 6](#), we give our result on robustly estimating the parameters of a Gaussian under halfspace truncation. In [Section 7](#), we prove the extension of our result to an intersection of two orthogonal halfspaces. In [Appendices A](#) and [B](#), we prove some facts and computations needed for our results. Finally, in [Appendix C](#), we show how one could get w based on the moments. This is not needed in our proof, but we include it for completeness since this is necessary for the proof of [\[LMZ24\]](#).

3 Preliminaries

3.1 Notation

We use the following convention: We denote the unit sphere in \mathbb{R}^d by \mathcal{S}^{d-1} . For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \dots, n\}$. Unless explicitly stated, the base of the logarithm is e . For two vectors $a, b \in \mathbb{R}^d$, we denote their inner product by $\langle a, b \rangle$ or $a^\top b$. Unless otherwise specified, all vector norms ($\|\cdot\|$ or $\|\cdot\|_2$) are the Euclidean norm. For matrices, $\|\cdot\|$ denotes the spectral norm and $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_*$ denotes the nuclear norm. We use the notation $O(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$, \lesssim , \gtrsim to hide absolute constants. We use $\tilde{O}(\cdot)$ to hide logarithmic factors. We use $\mathbb{1}[\cdot]$ for the indicator variable. For positive semi-definite (PSD) matrices $A, B \in \mathbb{R}^{d \times d}$, we use $A \preceq B$ to mean that A is at most B in the Löwner order (meaning $B - A$ is PSD). We use $\lambda_{\min}(A)$ to denote the minimum eigenvalue of a square matrix A . We denote the total variation distance by d_{TV} . We denote the density function of the standard Gaussian by $\varphi(\cdot)$ and its cumulative density function by $\Phi(\cdot)$.

Tensor norm. For two order-3 tensors $T, S \in \mathbb{R}^{d \times d \times d}$, we define $\langle T, S \rangle = \sum_{ijk} T_{i,j,k} S_{i,j,k}$. Furthermore, we define the injective tensor norm of T as $\|T\| = \sup_{\|x\|=\|y\|=\|z\|=1} \{\langle x \otimes y \otimes z, T \rangle\}$. Note that if T is a symmetric tensor, as a consequence of Banach's Theorem (see e.g. [\[Fri13, Theorem 1\]](#) or [\[CHLZ12, Corollary 4.2\]](#)), it holds that

$$\|T\| = \sup_{\|x\|=\|y\|=\|z\|=1} \{\langle x \otimes y \otimes z, T \rangle\} = \sup_{\|x\|=1} \{\langle x^{\otimes 3}, T \rangle\}.$$

Truncated Gaussian. For a measurable set $S \subseteq \mathbb{R}^d$, let $\alpha > 0$ to be the probability mass of the set S under the true Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, i.e., $\alpha = \mathbb{E}_{x \sim \mathcal{N}(\mu, \Sigma)} \mathbb{1}[x \in S]$. We say that a random vector X follows a Gaussian $\mathcal{N}(\mu, \Sigma)$ truncated to S , if it follows the distribution $\mathcal{N}(\mu, \Sigma)$ conditioned on the event $x \in S$. For a halfspace $H = \{x : \langle w, x \rangle \leq \tau\}$, we denote a Gaussian $\mathcal{N}(\mu, \Sigma)$ truncated to a halfspace H by

$$\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$$

3.2 Inverse Mill's ratio

For $\gamma \in \mathbb{R}$, consider the standard Gaussian random variable $\mathcal{N}(0, 1)$ conditioned on $(-\infty, \gamma]$, denoted by Z_γ . In this section, we introduce the first three cumulants of Z_γ , show how to describe them in terms of the density function and the cumulative density function of the standard Gaussian and give several bounds that will be needed later in our proofs.

For a general random variable X on \mathbb{R} , the cumulants are defined as follows.

Definition 3.1. Let X be a random variable on \mathbb{R} . We define its cumulant generating function by $K(t) = \log \mathbb{E} e^{tX}$. If $K(t)$ exists and is analytic, we define the n -th cumulant of X , denoted by $\kappa_n(X)$ as the n -th order derivative of $K(t)$ evaluated at $t = 0$, i.e., $\kappa_n(X) = K^{(n)}(0)$.

In particular, one can compute that the first cumulant is the mean of X , the second cumulant is the variance of X and the third cumulant is the third central moment. Thus, the first three cumulants of Z_γ (which we denote by $\kappa_1(\gamma)$, $\kappa_2(\gamma)$, and $\kappa_3(\gamma)$ respectively) are as follows:

$$\begin{aligned}\kappa_1(\gamma) &= \mathbb{E} Z_\gamma, \\ \kappa_2(\gamma) &= \mathbb{E} (Z_\gamma - \mathbb{E} Z_\gamma)^2, \\ \kappa_3(\gamma) &= \mathbb{E} (Z_\gamma - \mathbb{E} Z_\gamma)^3.\end{aligned}\tag{3.1}$$

They can be characterized in terms of $\varphi(\gamma)$ and $\Phi(\gamma)$ as follows.

Fact 3.2. Let $\gamma \in \mathbb{R}$. We have that

$$\begin{aligned}\kappa_1(\gamma) &= \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t \varphi(t) dt = -\frac{\varphi(\gamma)}{\Phi(\gamma)}, \\ \kappa_2(\gamma) &= 1 - \gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} - \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2, \\ \kappa_3(\gamma) &= \frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \left(1 - \gamma^2 - 3\gamma \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)} - 2 \cdot \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2 \right).\end{aligned}$$

The proof of [Fact 3.2](#) follows by direct computation. We include a proof in [Section A.1](#).

Moreover, the *skewness* of a random variable X , denoted by $\text{skew}(X)$ on \mathbb{R} , is defined as the third moment of the normalized random variable $\frac{X-\mu}{\sigma}$, where μ denotes the mean of X and σ the standard deviation. This can also be written in terms of the cumulants. Namely, we have that

$$\text{skew}(X) = \frac{\kappa_3(X)}{\kappa_2(X)^{3/2}}.$$

We denote the skewness of Z_γ by $\text{skew}(\gamma)$. Using [Fact 3.2](#), we immediately get that

$$\text{skew}(\gamma) = \frac{\varphi(\gamma)}{\Phi(\gamma)} \left(1 - \gamma^2 - 3\gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} - 2 \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2 \right) \left(1 - \gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} - \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2 \right)^{-3/2}.\tag{3.2}$$

We now discuss some properties of $\kappa_1, \kappa_2, \kappa_3$ and give some estimates for them that we will need for our analysis later. Observe that $\kappa_1(x), \kappa_2(x), \kappa_3(x)$ can be viewed as polynomials in x and $\frac{\varphi(x)}{\Phi(x)}$.

We have that the ratio $\frac{\varphi(x)}{\Phi(x)} \approx \varphi(x)$ as $x \rightarrow +\infty$ and thus the non-trivial case is when $x \rightarrow -\infty$. To address this latter case, as in [\[LMZ24\]](#), we introduce the following function, commonly known as the *inverse Mills ratio* or *hazard function*:

$$h(x) = \frac{\varphi(x)}{1 - \Phi(x)}.$$

Observe that bounds on $h(x)$ (which typically only work for $x \geq 0$) can be related to $\frac{\varphi(-x)}{\Phi(-x)}$ as follows

$$\frac{\varphi(-x)}{\Phi(-x)} = \frac{\varphi(x)}{1 - \Phi(x)} = h(x).$$

We will need the following three facts.

Fact 3.3 ([GU14]). Define the following functions for $k \in \mathbb{N}_0$ and $x > 0$

$$r_{k+1}(x) := x + \frac{1}{x + \frac{2}{x + \frac{3}{\ddots x + \frac{k}{x}}}}$$

Then, for any $m \in \mathbb{N}$ and $x \geq 0$ we have

$$r_{2m-1}(x) < h(x) < r_{2m}(x).$$

In particular, we have

$$x < h(x) < x + \frac{1}{x}.$$

We have that $r_{2m}(x) \searrow h(x)$ and $r_{2m+1} \nearrow h(x)$ as $m \rightarrow \infty$ for all $x > 0$. In particular, the above estimates are accurate for large $x > 0$. Intuitively, this is because the higher-order term $x + \frac{k}{x} \approx x$ for large x . However, for small fixed k , the approximation quality of $r_k(x)$ can be extremely poor near $x = 0$. In particular, despite $h(0) = \sqrt{\frac{2}{\pi}}$, the upper bound $r_{2m}(x) \rightarrow \infty$ as $x \rightarrow 0$ for all fixed integers $m > 0$. For x close to 0, we can instead use the following two estimates of the inverse Mills ratio, where the second one is exact at $x = 0$.

Fact 3.4 ([Sam53, Bir42]). For $x > 0$, we have

$$\frac{1}{4}(\sqrt{x^2 + 8} + 3x) < h(x) < \frac{1}{2}(\sqrt{x^2 + 4} + x).$$

Fact 3.5 ([Boy59]). For $x > 0$, we have

$$\frac{\sqrt{(\pi - 2)^2 x^2 + 2\pi} + 2x}{\pi} < h(x) < \frac{\sqrt{x^2 + 2\pi} + (\pi - 1)x}{\pi}.$$

Based on these facts, we prove the following claims that are necessary for our analysis. We provide proofs of these claims in [Section A.1](#).

Claim 3.6. The function κ_1 as defined in (3.1) is 1-Lipschitz on \mathbb{R} .

Claim 3.7. For every $t \in \mathbb{R}$, we have that $0 < \kappa_2(t) \leq 1$.

Claim 3.8. For κ_2 and κ_3 as in (3.1), we have $\frac{d}{dt} \kappa_2(t) = -\kappa_3(t)$. Moreover, $\kappa_3(t) < 0$ for every $t \in \mathbb{R}$ and thus $\kappa_2(t)$ is monotonically increasing.

Claim 3.9. For every $t \in \mathbb{R}$, we have $|\kappa_1(t)| \cdot \sqrt{\kappa_2(t)} \leq O(1)$.

Claim 3.10. For every $t \in \mathbb{R}$, we have $|\frac{d}{dt} \kappa_2(t)| = |\kappa_3(t)| \leq O(1)$ and $0 < \kappa_1(t) \cdot \kappa_3(t) \leq O(1)$.

Claim 3.11. For $t \geq 1$, we have $|\kappa_1(t)| \leq 1 - \kappa_2(t)$.

Claim 3.12. For $t \leq -3$, we have $\kappa_2(t) \geq \frac{1}{4t^2}$.

Claim 3.13. For $t \leq -2$, we have $|\kappa_3(t)| \geq \frac{1}{50|t|^3}$.

Claim 3.14. For $t \geq 1$, we have $1 - \kappa_2(t) \leq 16 \cdot (1 - \Phi(t)) \cdot \log 1/(1 - \Phi(t))$.

Claim 3.15. For $t \geq 2$, we have $\frac{t^2}{2} \leq \log(1/(1 - \Phi(t))) \leq 2t^2$.

Claim 3.16. There are absolute constants $m, M > 0$ such that

$$m \leq \frac{|\text{skew}(t)|}{(1 - \Phi(t)) \log^{3/2}(1/(1 - \Phi(t)))} \leq M$$

for all $t \geq 0$.

The following lemma is central for our proof.

Lemma 3.17. The function skew as defined in (3.2) is monotonically increasing on \mathbb{R} .

Proof. From Claim 3.8 we know that $\frac{d}{dt} \kappa_2(t) = -\kappa_3(t)$. It follows that

$$\begin{aligned} \frac{d}{dt} \text{skew}(t) &= \frac{\left(\frac{d}{dt} \kappa_3(t)\right) \kappa_2(t)^{3/2} - \frac{3}{2} \sqrt{\kappa_2(t)} \left(\frac{d}{dt} \kappa_2(t)\right) \cdot \kappa_3(t)}{\kappa_2(t)^3} \\ &= \frac{2 \left(\frac{d}{dt} \kappa_3(t)\right) \kappa_2(t) + 3\kappa_3(t)^2}{2\kappa_2(t)^{5/2}} \end{aligned}$$

Since $\kappa_2(t) > 0$ for every $t \in \mathbb{R}$ by Claim 3.7, it suffices to show that the numerator is positive. Defining $g(t) = \frac{\Phi(t)}{\varphi(t)} \cdot (2 \left(\frac{d}{dt} \kappa_3(t)\right) \kappa_2(t) + 3\kappa_3(t)^2)$, since $\frac{\Phi(t)}{\varphi(t)} > 0$, it is enough to show $g(t) > 0$ for all $t \in \mathbb{R}$. After some calculation we obtain

$$g(t) = 2t^3 - 6t + (t^4 + 14t^2 - 5) \frac{\varphi(t)}{\Phi(t)} + (2t^3 + 20t) \left(\frac{\varphi(t)}{\Phi(t)}\right)^2 + (t^2 + 8) \left(\frac{\varphi(t)}{\Phi(t)}\right)^3. \quad (3.3)$$

In order to show that $g(t) > 0$ for all $t \in \mathbb{R}$, we consider the following function

$$p_t(x) = 2t^3 - 6t + (t^4 + 14t^2 - 5)x + (2t^3 + 20t)x^2 + (t^2 + 8)x^3.$$

Note that $g(t) = p_t\left(\frac{\varphi(t)}{\Phi(t)}\right)$. Our argument is now structured as follows: For fixed t , we will define values $x_\ell(t) \leq \frac{\varphi(t)}{\Phi(t)} \leq x_u(t)$. We then show that p_t is monotonically increasing on $[x_\ell(t), x_u(t)]$. Then we lower bound

$$g(t) = p_t\left(\frac{\varphi(t)}{\Phi(t)}\right) \geq p_t(x_\ell(t))$$

and argue that $p_t(x_\ell(t)) > 0$. In order to show that p_t is monotonically increasing, we compute the derivative of p_t with respect to x :

$$\left(\frac{\partial}{\partial x} p_t\right)(x) = t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x + 3(t^2 + 8) \cdot x^2$$

and show that this is positive for $x \in [x_\ell(t), x_u(t)]$.

We distinguish the following cases: (i) $0 < t$; (ii) $-\frac{4}{10} < t \leq 0$; and (iii) $t < -\frac{4}{10}$. For each of the cases we give values for $x_\ell(t)$ and $x_u(t)$ based on approximations for $\frac{\varphi(x)}{\Phi(x)}$ and show that $g(t) > 0$ using the strategy described above. We will defer some of the computations to Appendix B, where we give Mathematica code to verify the statements we need.

Case (i): $0 < t$. For $t > 0$, we define $x_\ell(t) = \frac{e^{-t^2/2}}{\sqrt{\pi/2+t}} > 0$. Note that since for $t > 0$ we have $\Phi(t) \leq \frac{1}{2} + \frac{t}{\sqrt{2\pi}}$, we get that $\frac{\varphi(t)}{\Phi(t)} \geq x_\ell(t)$ for every $t > 0$. Since we have $4(t^3 + 10t) > 0$ and $3(t^2 + 8) > 0$ for $t > 0$, we get that for any $x \geq x_\ell(t)$ ⁷

$$\left(\frac{\partial}{\partial x} p_t\right)(x) \geq \left(\frac{\partial}{\partial x} p_t\right)(x_\ell(t)).$$

We show in [Fact B.1](#) that

$$\left(\frac{\partial}{\partial x} p_t\right)(x_\ell(t)) > 0$$

for all $t > 0$. Hence, we get that for $t > 0$, p_t is monotonically increasing on $[x_\ell(t), \infty)$. Thus, we get that

$$g(t) = p_t\left(\frac{\varphi(t)}{\Phi(t)}\right) \geq p_t(x_\ell(t)).$$

In [Fact B.1](#), we also show that $p_t(x_\ell(t)) > 0$ for $t > 0$, which shows $g(t) > 0$ for $t > 0$.

Case (ii): $-\frac{4}{10} < t \leq 0$. In this case, we define

$$x_\ell(t) = \frac{\sqrt{(\pi - 2)^2 t^2 + 2\pi} - 2t}{\pi} \quad \text{and} \quad x_u(t) = \frac{\sqrt{t^2 + 2\pi} - (\pi - 1)t}{\pi}.$$

For $t \leq 0$, we have $x_\ell(t) > 0$. By [Fact 3.5](#) we have $x_\ell(t) \leq \frac{\varphi(t)}{\Phi(t)} \leq x_u(t)$. Since for $t \leq 0$ we have $4(t^3 + 10t) \leq 0$ and $3(t^2 + 8) > 0$, we get that, for $x_\ell(t) \leq x \leq x_u(t)$ we have⁸

$$\left(\frac{\partial}{\partial x} p_t\right)(x) \geq t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x_\ell(t)^2.$$

We show in [Fact B.2](#) that

$$t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x_\ell(t)^2 > 0$$

for $-\frac{4}{10} < t \leq 0$. Thus, we have that p_t is monotonically increasing on $[x_\ell(t), x_u(t)]$ and thus in particular

$$g(t) = p_t\left(\frac{\varphi(t)}{\Phi(t)}\right) \geq p_t(x_\ell(t)).$$

In [Fact B.2](#) we also show that $p_t(x_\ell(t)) > 0$ for $-\frac{4}{10} < t \leq 0$ and thus $g(t) > 0$ for $-\frac{4}{10} < t \leq 0$.

Case (iii): $t \leq -\frac{4}{10}$. In this case, we define $x_\ell(t) = r_{81}(-t)$ and $x_u = r_{20}(-t)$. We verify in [Fact B.3](#) that for $t \leq -\frac{4}{10}$ we have $x_\ell(t) > 0$. By [Fact 3.3](#), we have that $x_\ell(t) \leq \frac{\varphi(t)}{\Phi(t)} \leq x_u(t)$. We again get that⁸

$$\left(\frac{\partial}{\partial x} p_t\right)\left(\frac{\varphi(t)}{\Phi(t)}\right) \geq t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x_\ell(t)^2.$$

In [Fact B.3](#) we show that

$$t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x_\ell(t)^2 > 0$$

⁷Since $x_\ell(t) > 0$, this also implies $x^2 \geq x_\ell(t)^2$.

⁸We are again using that since $x_\ell(t) > 0$ we also have $x_\ell(t)^2 \leq x^2$.

for $t \leq -\frac{4}{10}$. We thus again get that p_t is monotonically increasing on $[x_\ell(t), x_u(t)]$ and thus

$$g(t) = p_t \left(\frac{\varphi(t)}{\Phi(t)} \right) \geq p_t(x_\ell(t)).$$

Finally, we also show in [Fact B.3](#) that $p_t(x_\ell(t)) > 0$ for $t \leq -\frac{4}{10}$ and thus also $g(t) > 0$ for these t . This completes the proof. \square

In [Appendix B](#), we will show the following bound on skew.

Claim 3.18. *We have $\lim_{t \rightarrow -\infty} \text{skew}(t) = -2$ and $\lim_{t \rightarrow +\infty} \text{skew}(t) = 0$. Thus, for all $t \in \mathbb{R}$ we have $\text{skew}(t) \in [-2, 0]$ and in particular $|\text{skew}(t)| \leq 2$.*

We can even strengthen [Lemma 3.17](#). In [Appendix B](#), we will show the following two claims that we need for the proofs later. For both of the claims, the idea is that the involved function are point-wise strictly positive. Thus, since the functions are continuous, it is enough to show that the limits as $t \rightarrow \pm\infty$ are positive as well to conclude that we can lower bound them by a constant $C > 0$. In [Appendix B](#), we compute these limits and thus prove these claims.

Claim 3.19. *There exists a constant $C > 0$ such that for all $t \in \mathbb{R}$ we have*

$$\frac{\frac{d}{dt} \text{skew}(t)}{\left| \frac{d}{dt} \kappa_2(t) \right|} \geq C.$$

Moreover, $|\kappa_2(a) - \kappa_2(b)| \leq \frac{1}{C} \cdot |\text{skew}(a) - \text{skew}(b)|$ for every $a, b \in \mathbb{R}$.

Claim 3.20. *Define $\psi(t) := \frac{\kappa_1(t)}{\kappa_2(t)^{1/2}}$. Then, there exists a constant $C > 0$ such that for all $t \in \mathbb{R}$ we have*

$$\frac{\frac{d}{dt} \text{skew}(t)}{\left| \frac{d}{dt} \psi(t) \right|} \geq C \cdot \kappa_2(t)^2.$$

Moreover, $|\psi(a) - \psi(b)| \leq \frac{1}{C} \cdot |\text{skew}(a) - \text{skew}(b)| \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\}$ for every $a, b \in \mathbb{R}$.

Furthermore, we can also show an analogous statement for the relation between κ_1 and κ_2 , which we will also prove in [Appendix B](#). The proof idea is the same as for the above two claims (note that κ_2 is also monotonic by [Claim 3.8](#)).

Claim 3.21. *There exists a constant $C > 0$ such that for all $t \in \mathbb{R}$ we have*

$$\frac{\frac{d}{dt} \kappa_2(t)}{\left| \frac{d}{dt} \kappa_1(t) \right|} \geq C \cdot \kappa_2(t)^2.$$

Moreover, $|\kappa_1(a) - \kappa_1(b)| \leq \frac{1}{C} \cdot |\kappa_2(a) - \kappa_2(b)| \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\}$ for every $a, b \in \mathbb{R}$.

Finally, we need tight estimates of $\kappa_2(\gamma)$ and $\kappa_3(\gamma)$ for our proofs later. In order to relate these bounds to the mass α of the truncation set (the halfspace), we use the relation $\alpha = \Phi(\gamma)$. We defer its proof to [Lemma 4.2](#).

Claim 3.22. *Let $\gamma \in \mathbb{R}$. For $\alpha = \Phi(\gamma)$, we have*

$$\kappa_2(\gamma) \geq \min \left\{ \frac{1}{20}, \frac{1}{8 \log 1/\alpha} \right\}.$$

Proof. From [Claim 3.8](#) we know that $\kappa_2(\gamma)$ monotonically increases on \mathbb{R} . For $\gamma \rightarrow \infty$ the distribution converges to standard Gaussian, and the variance goes to 1. For the case $\gamma \rightarrow -\infty$, we can use the rational approximation of the Mills ratio from [Fact 3.3](#). Concretely, we distinguish the two cases $\gamma > -3$ and $\gamma \leq -3$:

Case (i): $\gamma > -3$. For $\gamma > -3$ we have $\kappa_2(\gamma) \geq \kappa_2(-3) \geq 1/20$ by monotonicity of κ_2 .

Case (ii): $\gamma \leq -3$. For $\gamma \leq -3$, by [Claim 3.12](#) we have that $\kappa_2(\gamma) \geq \frac{1}{4\gamma^2}$. It remains to get a bound on γ in terms of α . By [Fact 3.3](#), we know that $\alpha = \Phi(\gamma) \leq \frac{\varphi(\gamma)}{-\gamma}$. This implies that

$$\frac{1}{\alpha} \geq -\sqrt{2\pi}\gamma \cdot e^{\gamma^2/2} \geq 3\sqrt{2\pi} \cdot e^{\gamma^2/2} \geq e^{\gamma^2/2}$$

and thus we get that $\log 1/\alpha \geq \gamma^2/2$. Combining this with bound on $\kappa_2(\gamma)$ from above, we get that, for $\gamma \leq -3$,

$$\kappa_2(\gamma) \geq \frac{1}{8} \cdot \frac{1}{\log 1/\alpha}.$$

Combining the two cases, we get the lower bound as claimed. \square

Claim 3.23. *Let $\gamma \in \mathbb{R}$. For $\alpha = \Phi(\gamma)$, we have*

$$|\kappa_3(\gamma)| \geq \min \left\{ \frac{1}{50 \log^{3/2} 1/\alpha}, \frac{1}{20}, \frac{1}{4\sqrt{2}}(1-\alpha) \log^{3/2} \frac{1}{1-\alpha} \right\}.$$

Proof. Unlike $\kappa_2(\gamma)$, the function $\kappa_3(\gamma)$ is not monotonic in γ . Moreover, the function approaches 0 both for $\gamma \rightarrow +\infty$ and $\gamma \rightarrow -\infty$. Thus, we split our analysis in three parts: First, we consider the two cases $\gamma \rightarrow \pm\infty$ where we use the rational approximations of the Mills ratio from [Fact 3.3](#). Finally, we bound $\kappa_2(\gamma)$ on a small interval around 0 by an explicit computation.

Case (i): $\gamma \leq -2$. By [Claim 3.13](#), for $\gamma \leq -2$, we have the lower bound $|\kappa_3(\gamma)| \geq \frac{1}{50|\gamma|^3}$. From the proof of [Claim 3.22](#) we already know that $\log 1/\alpha \geq \gamma^2/2$ for $\gamma \leq -3 \leq -2$. So it holds that $\kappa_3(\gamma) \geq \frac{1}{50 \log^{3/2} 1/\alpha}$ for $\gamma \leq -2$.

Case (ii): $\gamma \geq 2$. From [Claim 3.8](#) we know $\kappa_3(\gamma) < 0$ and thus we have that

$$|\kappa_3(\gamma)| = h(\gamma) (\gamma^2 - 1 + 3\gamma h(\gamma) + 2h(\gamma)^2).$$

Noting that $h(\gamma) > 0$ and $\Phi(\gamma) \leq 1$ for all γ , we have for $\gamma \geq 2 > \sqrt{2}$ that

$$|\kappa_3(\gamma)| \geq \varphi(\gamma)(\gamma^2 - 1) \geq \frac{1}{2}\varphi(\gamma)\gamma^2.$$

By [Fact 3.3](#), we have for $\gamma \geq 2 > 0$ that $\gamma \leq \frac{\varphi(\gamma)}{1-\Phi(\gamma)}$. Combining this with the above bound on $|\kappa_3(\gamma)|$ and using $\alpha = \Phi(\gamma)$ we get

$$|\kappa_3(\gamma)| \geq \frac{1}{2}\gamma^3(1-\alpha).$$

It again remains to give a bound on γ in terms of α . Note that, for $\gamma \geq 2$, $\alpha = \Phi(\gamma)$, by [Claim 3.15](#) we have

$$\log 1/(1-\alpha) \leq 2\gamma^2$$

Combining this with the inequality above, we get that

$$|\kappa_3(\gamma)| \geq \frac{1}{2} \left(\frac{1}{2} \log \left(\frac{1}{1-\alpha} \right) \right)^{3/2} (1-\alpha) = \frac{1}{4\sqrt{2}}(1-\alpha) \log^{3/2} \frac{1}{1-\alpha}$$

Case (iii): $\gamma \in (-2, 2)$. For the last case, we verify in [Fact B.4](#) in [Appendix B](#) that for $\gamma \in (-2, 2)$, we have $|\kappa_3(\gamma)| \geq \frac{1}{20}$.

Putting all the three cases together gives the result. \square

3.3 Gaussian concentration and anti-concentration

We will need the following two results on the concentration and anti-concentration of a Gaussian random variable. We note that the first fact follows immediately from the boundedness of the Gaussian density.

Fact 3.24 (Gaussian anti-concentration). *For every $\delta > 0$, we have that*

$$\mathbb{P}_{X \sim N(0, \sigma^2)} [-\delta \leq X \leq \delta] \leq \sqrt{\frac{2}{\pi}} \cdot \frac{\delta}{\sigma}$$

Fact 3.25 (Gaussian concentration, see e.g. [\[vH16, Theorem 3.25\]](#)). *Let $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a λ -Lipschitz function. Let $g \sim N(0, I_d)$. Then for all $t > 0$,*

$$\mathbb{P} [|\psi(g) - \mathbb{E} \psi(g)| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\lambda^2}\right)$$

3.4 Concentration bounds for sub-Gaussian random variables

In this section, we state two concentration results for sub-Gaussian random variables. To do so, we define the following (quasi)-norm on a random variable X . Let $\psi : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$ be a non-decreasing function that satisfies $\psi(0) = 0$ and $\psi(x) \rightarrow \infty$ as $x \rightarrow \infty$. Then we define the following (quasi)-norm of X as

$$\|X\|_\psi = \inf \left\{ t > 0 \mid \mathbb{E} \psi\left(\frac{|X|}{t}\right) \leq 1 \right\}$$

For a constant $\theta \in (0, 1) \cup \{2\}$, let $\psi_\theta(x) = \exp(x^\theta) - 1$. Setting $\psi = \psi_2$ in the above results in the sub-Gaussian norm. Recall that a random vector X is r -sub-Gaussian if $\langle v, X \rangle$ is r -sub-Gaussian for every unit vector v (meaning that $\|\langle v, X \rangle - \mathbb{E}[\langle v, X \rangle]\|_{\psi_2} \leq r$).⁹ The following fact provides a standard concentration bound for the covariance estimation of sub-Gaussian distributions. Note that the second part follows directly from the first one by bounding $\|A\|_F \leq \sqrt{d} \cdot \|A\|$.

Fact 3.26 ([\[Ver18, Theorem 4.6.1\]](#)). *Let $y_1, \dots, y_n \in \mathbb{R}^d$ be i.i.d. samples with covariance $\Sigma > 0$ such that $\Sigma^{-1/2} y_i$ is r -sub-Gaussian. Let the sample mean be $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and let the sample covariance be $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^{\otimes 2}$. Then, with probability at least $1 - \delta$, we have that*

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\| \lesssim r^2 \cdot \left(\sqrt{\frac{d + \log 1/\delta}{n}} + \frac{d + \log 1/\delta}{n} \right).$$

Furthermore, there exists some absolute constant $C > 0$ such that for $n \geq C(d + \log 1/\delta)$ we have with probability at least $1 - \delta$

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\|_F \lesssim r^2 \cdot \frac{d + \sqrt{d \log 1/\delta}}{\sqrt{n}}.$$

⁹Often, such a random variable is also called σ^2 -sub-Gaussian, where $\sigma = r$ is an upper bound on the ψ_2 -norm of the one-dimensional projections.

In the remainder of this section, we show a concentration result for cubes of sub-Gaussian random variables. As in [KC22], we define the function $\psi_{\theta,L}^{-1}(y) = \sqrt{\log(1+y)} + L \log(1+y)^{1/\theta}$ for all $y \geq 0$, and let $\psi_{\theta,L}$ be the inverse function of $\psi_{\theta,L}^{-1}$. Note that when $\theta \in (0,1)$, $\|\cdot\|_{\psi_{\theta,L}}$ is not a norm since it only satisfies sub-additivity up to a multiplicative constant depending on θ . The norm $\|\cdot\|_{\psi_{\theta,L}}$ can be related to the norm $\|\cdot\|_{\psi_\theta}$ via the following theorem.

Theorem 3.27 (Theorem 3.1 in [KC22]). *Fix $\theta \in (0,1)$. Let $k \geq 1$ and consider a sequence of independent zero-mean random variables $\{X_i\}_{i=1}^k$ such that $\|X_i\|_{\psi_\theta} < \infty$ for all $i \in [k]$. Then there exist constants $Q_\theta, Q'_\theta > 0$ that only depend on θ such that*

$$\left\| \sum_{i=1}^k X_i \right\|_{\psi_{\theta,L}} \leq Q_\theta \cdot \sqrt{\sum_{i=1}^k \|X_i\|_{\psi_\theta}^2}$$

where $L = \frac{Q'_\theta}{\sqrt{k}}$.

Finally, we have the following tail bound for a random variable with finite $\|\cdot\|_{\psi_{\theta,L}}$.

Fact 3.28 (Proposition A.3 in [KC22]). *Fix $\theta \in (0,1)$ and $L > 0$. Then, for a random variable X with $\|X\|_{\psi_{\theta,L}} < \infty$, we have that*

$$\mathbb{P} \left[|X| \geq \|X\|_{\psi_{\theta,L}} (\sqrt{t} + Lt^{1/\theta}) \right] \leq 2 \exp(-t)$$

for all $t > 0$.

We will now put the above results together to get the following corollary about the concentration of cubes of sub-Gaussian random variables.

Corollary 3.29 (Concentration inequality for cubes of sub-Gaussians). *Let Z_1, \dots, Z_n be i.i.d. samples of a real-valued random variable Z such that $\|Z\|_{\psi_2} \leq r$. Then there exists an absolute constant $K > 0$ such that, for every $t > 0$,*

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n Z_i^3 - \mathbb{E} Z^3 \right| \geq Kr^3 \left(\sqrt{\frac{t}{n}} + \frac{t^{3/2}}{n} \right) \right) \leq 2e^{-t}.$$

Proof. Since $\|Z\|_{\psi_2} \leq r$, we have from the definition that

$$\|Z^3\|_{\psi_{2/3}} \leq \|Z\|_{\psi_2}^3 \leq r^3.$$

Also, the $\psi_{2/3}$ -quasi-norm satisfies the triangle inequality up to an absolute constant, that is,

$$\|Z^3 - \mathbb{E} Z^3\|_{\psi_{2/3}} \lesssim \|Z^3\|_{\psi_{2/3}} + |\mathbb{E} Z^3|.$$

We will show that $|\mathbb{E} Z^3|$ is in fact bounded by $O(\|Z^3\|_{\psi_{2/3}})$. To show this we first note by definition of the $\psi_{2/3}$ norm that

$$\mathbb{E} \left[\exp \left(\frac{|Z^3|}{\|Z^3\|_{\psi_{2/3}}} \right)^{2/3} \right] \leq 2.$$

Using this and applying Markov's inequality we get

$$\begin{aligned} \mathbb{P}\left(|Z^3| \geq u \|Z^3\|_{\psi_{2/3}}\right) &= \mathbb{P}\left(\exp\left(\frac{|Z^3|}{\|Z^3\|_{\psi_{2/3}}}\right)^{2/3} \geq \exp\left(u^{2/3}\right)\right) \\ &\leq \mathbb{E}\left[\exp\left(\frac{|Z^3|}{\|Z^3\|_{\psi_{2/3}}}\right)^{2/3}\right] \cdot \exp\left(u^{-2/3}\right) \\ &\leq 2 \exp\left(u^{-2/3}\right) \end{aligned}$$

Obtaining the above bound will now enable us to apply the integral identity for expectations. We therefore have

$$\mathbb{E}|Z^3| = \int_0^\infty \mathbb{P}(|Z^3| \geq s) ds \leq \|Z^3\|_{\psi_{2/3}} \int_0^\infty 2e^{-u^{2/3}} du \lesssim \|Z^3\|_{\psi_{2/3}}.$$

where in the first inequality we used the substitution $s = u \cdot \|Z^3\|_{\psi_{2/3}}$. Hence, we have that

$$\|Z^3 - \mathbb{E}Z^3\|_{\psi_{2/3}} \lesssim r^3.$$

Now define

$$Y_i := Z_i^3 - \mathbb{E}Z_i^3.$$

Then, the Y_1, \dots, Y_n are i.i.d., mean-zero, and satisfy

$$\|Y_i\|_{\psi_{2/3}} \lesssim r^3.$$

Applying the sum-of-sub-Weibulls concentration from [Theorem 3.27](#) we get

$$\left\| \sum_{i=1}^n Y_i \right\|_{\psi_{2/3,L}} \lesssim \sqrt{n} r^3,$$

where $L = C/\sqrt{n}$ for some absolute constant $C > 0$. Using the corresponding tail inequality from [Fact 3.28](#) we have that

$$\mathbb{P}\left(\left| \sum_{i=1}^n Y_i \right| \geq K\sqrt{n} r^3 \left(\sqrt{t} + \frac{t^{3/2}}{\sqrt{n}}\right)\right) \leq 2e^{-t}.$$

Finally, dividing by n , we conclude that

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n Z_i^3 - \mathbb{E}Z^3 \right| \geq Kr^3 \left(\sqrt{\frac{t}{n}} + \frac{t^{3/2}}{n}\right)\right) \leq 2e^{-t},$$

which completes the proof. □

4 Properties of halfspace truncated Gaussians

In this section, we prove several statements that we need in the proofs later. In [Section 4.1](#), we compute the first three moments of a truncated Gaussian. In [Section 4.2](#), we show that halfspace truncated Gaussians are sub-Gaussian. And finally, in [Section 4.3](#), we show how to precondition a truncated Gaussian such that the covariance is well-conditioned.

4.1 The first three moments

In order to compute the first three moments, we need the following claim that shows that linear transformations of Gaussians truncated to a halfspace continue to remain Gaussians truncated to a halfspace albeit with different parameters. We defer the proof of this claim to [Section A.2](#).

Claim 4.1 (Linear transformations preserve halfspace truncation). *Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix and let $b \in \mathbb{R}^d$. Consider the random variable $x \sim \mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$. Then, the random variable $y = Ax + b$ follows the distribution $\mathcal{N}(A\mu + b, A\Sigma A^\top)|_{\langle w', y \rangle \leq \tau'}$, where*

$$w' = \frac{A^{-\top}w}{\|A^{-\top}w\|} \quad \text{and} \quad \tau' = \frac{\tau + \langle A^{-\top}w, b \rangle}{\|A^{-\top}w\|}.$$

Using this claim, we can now compute the structure of the first moment as well as the second and third central moments of the Gaussian distribution under arbitrary halfspace truncation. Each of them can be viewed as the corresponding moment of an untruncated Gaussian and an additive rank-1 term due to truncation.

Lemma 4.2. *Let $X \sim \mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$ for some unit vector w . Then, we have*

$$\begin{aligned} \mathbb{E}X &= \mu + \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2}w\|}, \\ M_2 &:= \mathbb{E}(X - \mathbb{E}X)^{\otimes 2} = \Sigma - (1 - \kappa_2(\gamma)) \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|} \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^\top, \\ M_3 &:= \mathbb{E}(X - \mathbb{E}X)^{\otimes 3} = \kappa_3(\gamma) \cdot \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^{\otimes 3}, \end{aligned}$$

where $\gamma = \frac{\tau - \langle w, \mu \rangle}{\|\Sigma^{1/2}w\|}$. Furthermore, we have that $\alpha = \Phi(\gamma)$.

Proof. Consider a random vector Y that follows a d -dimensional Gaussian $\mathcal{N}(0, I_d)|_{\langle e_1, y \rangle \leq \gamma}$. Since only the first coordinate of Y is truncated (and equal to Z_γ as in [Section 3.2](#)), we have that

$$\begin{aligned} \mathbb{E}Y &= \kappa_1(\gamma)e_1, \\ \mathbb{E}(Y - \mathbb{E}Y)^{\otimes 2} &= I_d - (1 - \kappa_2(\gamma)) \cdot e_1^{\otimes 2}, \\ \mathbb{E}(Y - \mathbb{E}Y)^{\otimes 3} &= \kappa_3(\gamma) \cdot e_1^{\otimes 3}. \end{aligned}$$

Consider an orthonormal matrix U such that $Ue_1 = \frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|}$ or equivalently $e_1 = U^\top \frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|}$. By [Claim 4.1](#), we have that that UY follows a standard Gaussian $\mathcal{N}(0, I_d)$ truncated to the set

$$\left\{ x : \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top x \leq \gamma \right\}.$$

By scaling UY by $\Sigma^{1/2}$ and shifting by μ , again by [Claim 4.1](#), we obtain a Gaussian $\mathcal{N}(\mu, \Sigma)$ truncated to the halfspace

$$H = \{x : \langle w, x \rangle \leq \gamma \|\Sigma^{1/2}w\| + \langle w, \mu \rangle\}.$$

Thus, we get a direct relationship between Y and X . Namely, if we have $\tau = \gamma \|\Sigma^{1/2}w\| + \langle w, \mu \rangle$ or equivalently $\gamma = \frac{\tau - \langle w, \mu \rangle}{\|\Sigma^{1/2}w\|}$, then we can get X by $X = \Sigma^{1/2}UY + \mu$. By construction, this also implies

$$\alpha = \mathbb{E}_{x \sim \mathcal{N}(\mu, \Sigma)} \mathbb{1}[x \in H] = \mathbb{E}_{y \sim \mathcal{N}(0, I_n)} \mathbb{1}[\{e_1^\top y \leq \gamma\}] = \mathbb{E}_{y_1 \sim \mathcal{N}(0, 1)} \mathbb{1}[y_1 \leq \gamma] = \Phi(\gamma).$$

Finally, from the relationship $X = \Sigma^{1/2}UY + \mu$ we compute the moments as follows. First, the mean of X is

$$\mathbb{E} X = \mathbb{E} [\mu + \Sigma^{1/2}UY] = \mu + \Sigma^{1/2}U \cdot \kappa_1(\gamma)e_1 = \mu + \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2}w\|}.$$

Next, we can compute the second central moment of X as

$$\begin{aligned} \mathbb{E} (X - \mathbb{E} X)^{\otimes 2} &= (\Sigma^{1/2}U)^{\otimes 2} \mathbb{E} (Y - \mathbb{E} Y)^{\otimes 2} \\ &= (\Sigma^{1/2}U)^{\otimes 2} (I_d - (1 - \kappa_2(\gamma))e_1^{\otimes 2}) \\ &= \Sigma - (1 - \kappa_2(\gamma)) \left(\Sigma^{1/2}Ue_1 \right)^{\otimes 2} \\ &= \Sigma - (1 - \kappa_2(\gamma)) \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^{\otimes 2}. \end{aligned}$$

Finally, the third central moment equals

$$\mathbb{E} (X - \mathbb{E} X)^{\otimes 3} = (\Sigma^{1/2}U)^{\otimes 3} \mathbb{E} (Y - \mathbb{E} Y)^{\otimes 3} = (\Sigma^{1/2}U)^{\otimes 3} \kappa_3(\gamma)e_1^{\otimes 3} = \kappa_3(\gamma) \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^{\otimes 3},$$

which completes the proof. \square

4.2 Sub-Gaussianity

For our algorithm we need high accuracy estimates of the moments of a halfspace truncated Gaussian. To get those, we first establish the following two claims about sub-Gaussianity of the truncated Gaussian.

Claim 4.3. *Suppose $X \sim \mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$. Then $\Sigma^{-1/2}X$ is 1-sub-Gaussian and $M_2^{-1/2}X$ is $1/\sqrt{\kappa_2(\gamma)}$ -sub-Gaussian.*

Corollary 4.4. *Suppose $X \sim \mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$. Then X is $\|\Sigma\|^{1/2}$ -sub-Gaussian.*

For the special case of the standard one-dimensional Gaussian, this was already shown in [BMA26]. We use this fact to prove the general case.

Fact 4.5 (cf. Theorem 2.1 in [BMA26]). *A random variable Z_γ from the standard Gaussian distribution $\mathcal{N}(0, 1)$ truncated to $(-\infty, \gamma]$ is 1-sub-Gaussian.*

Proof of Claim 4.3 and Corollary 4.4. A random vector is r -sub-Gaussian if its univariate projection along every direction v is r -sub-Gaussian. For the first claim it therefore suffices to show that the random variable $\langle v, \Sigma^{-1/2}X \rangle$ is 1-sub-Gaussian. Define $Y := U^\top \Sigma^{-1/2}X$. We first show that this is 1-sub-Gaussian. Let v be a unit vector. Then, we need to show that $\langle v, Y \rangle$ is 1-sub-Gaussian. Consider the canonical basis for \mathbb{R}^d defined using the vectors $\{e_j\}_{j=1}^d$. Then, by expressing v in this canonical basis we have

$$\langle v, Y \rangle = \sum_{j=1}^d v_j \langle e_j, Y \rangle = \sum_{j=1}^d v_j Y_j.$$

By construction, Y_1 is distributed identical to Z_τ . Since one-dimensional truncated Gaussians are 1-sub-Gaussian due to Fact 4.5, Y_1 is 1-sub-Gaussian. Now, Y_2, \dots, Y_d are independent one-dimensional standard Gaussians. Therefore they are also 1-sub-Gaussian. Since the density

function is a product distribution over the coordinates, we have independence between the $\{Y_i\}_{i=1}^n$. Using the fact that addition of independent sub-Gaussian random variables is sub-Gaussian, and the fact that v is a unit vector, it follows that Y is 1-sub-Gaussian.

Multiplying a random vector by a matrix A scales the sub-Gaussian parameter by $\|A^T\|$ (because $|\langle v, AY \rangle| \leq \|A^T\| \cdot |\langle v', Y \rangle|$ for $v' = A^T v / \|A^T v\|$). Thus, we have that $\Sigma^{-1/2}X$ is also 1-sub-Gaussian (because $\|U^T\| = 1$). Furthermore, we also get that X is $\|\Sigma^{1/2}\|$ -sub-Gaussian. Finally, by [Lemma 4.2](#), we have that

$$M_2 = \Sigma^{1/2} \cdot \underbrace{\left(I_d - (1 - \kappa_2(\gamma)) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}\|} \right)^T \right)}_{B:=} \cdot \Sigma^{1/2}.$$

We have that $B^{-1/2}\Sigma^{-1/2}X$ is $\|B^{-1/2}\|$ -sub-Gaussian. Since we get B by removing a rank-1 component from the identity matrix, the eigenvalues of B are 1 with multiplicity $d - 1$ and $\kappa_2(\gamma)$. Thus, $\|B^{-1/2}\| = 1/\sqrt{\kappa_2(\gamma)}$. Finally, note that $B^{-1/2}\Sigma^{-1/2}$ and $M_2^{-1/2}$ differ by an orthogonal matrix (namely $B^{-1/2}\Sigma^{-1/2}M_2^{1/2}$) and thus also $M_2^{-1/2}X$ is $1/\sqrt{\kappa_2(\gamma)}$ -sub-Gaussian, which completes the proof. \square

4.3 Preconditioning

For our algorithm, we need to estimate the inverse of M_2 . A straight-forward application of [Fact 3.26](#) to get an estimate \widehat{M}_2 for M_2 and then taking the inverse of \widehat{M}_2 would yield a sample complexity that depends on the condition number of M_2 . The reason for this is that from [Fact 3.26](#) we get a bound on $\|M_2^{-1/2}\widehat{M}_2M_2^{-1/2} - I_d\|$ but we will need a bound on $\|\widehat{M}_2^{-1} - M_2^{-1}\|$. We can relate these two via the following inequality

$$\|\widehat{M}_2^{-1} - M_2^{-1}\| \leq \|\widehat{M}_2^{-1}\| \cdot \|M_2^{1/2}\| \cdot \|I_d - M_2^{-1/2}\widehat{M}_2M_2^{-1/2}\| \cdot \|M_2^{-1/2}\|,$$

but in order to make this norm small, the sample complexity will depend on the condition number of M_2 , which can be arbitrarily large. To avoid this additional dependency, we compute a natural preconditioner to ensure that our problem becomes well-conditioned. The guarantee of this preconditioning is the following, where, as before, α denotes the probability mass of the halfspace under the target Gaussian distribution.

Lemma 4.6. *Without loss of generality, given $n = \Theta(d \cdot \max\{1, \log^4 1/\alpha\})$ samples from $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, we can assume that $\frac{1}{2}I_d \leq M_2 \leq 2I_d$ and $\frac{1}{2}I_d \leq \Sigma \leq \frac{2}{\kappa_2(\gamma)}I_d$ with high probability.*

Using [Lemma 4.6](#), we can obtain an inverse for M_2 without incurring a condition number factor in the sample complexity. Throughout the rest of the paper, we will work with the assumption that we have preconditioned the data as described in this lemma. The proof follows directly from the following two claims.

Claim 4.7. *Given $n = \Theta(d \cdot \max\{1, \log^4 1/\alpha\})$ i.i.d. samples from $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, with high probability the sample covariance $\widetilde{M}_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$ satisfies*

$$\left\| I_d - M_2^{-1/2} \widetilde{M}_2 M_2^{-1/2} \right\| \leq \frac{\kappa_2(\gamma)}{2}.$$

Claim 4.8. *Let $\widetilde{M}_2 > 0$ satisfy $\left\| I_d - M_2^{-1/2} \widetilde{M}_2 M_2^{-1/2} \right\| \leq \frac{\kappa_2(\gamma)}{2}$. Then, without loss of generality, we can assume that $\frac{1}{2}I_d \leq M_2 \leq 2I_d$ and $\frac{1}{2}I_d \leq \Sigma \leq \frac{2}{\kappa_2(\gamma)}I_d$.*

Proof of Claim 4.7. As long as we have $n = \Omega(d/\beta^2)$ samples, we know from [Fact 3.26](#) that

$$\|I_d - M_2^{-1/2} \tilde{M}_2 M_2^{-1/2}\| \lesssim \frac{1}{\kappa_2(\gamma)} \beta,$$

since for every sample x_i , $M_2^{-1} x_i$ is $1/\sqrt{\kappa_2(\gamma)}$ -sub-Gaussian by [Claim 4.3](#). Setting $\beta = \Omega(\kappa_2(\gamma)^2)$ and using [Claim 3.22](#) which lower bounds $\kappa_2(\gamma)$ in terms of α as $\kappa_2(\gamma) \geq \min\left\{1/20, \frac{1}{8 \log 1/\alpha}\right\}$, we obtain the result. \square

Proof of Claim 4.8. The idea of our preconditioning is to left-multiply the samples by $\tilde{M}_2^{-1/2}$, then apply our algorithm and in the end undo the preconditioning by changing the estimates appropriately. We will first argue that the moments after this preconditioning satisfy what we claim and then explain how to undo the preconditioning while keeping the error the same.

Moments are well-conditioned. By [Claim 4.1](#), the covariance of the untruncated Gaussian after this preconditioning is $\Sigma' = \tilde{M}_2^{-1/2} \Sigma \tilde{M}_2^{-1/2}$. The second central moment of the truncated Gaussian is $M'_2 = \tilde{M}_2^{-1/2} M_2 \tilde{M}_2^{-1/2}$. We want to argue that $\frac{1}{2} I_d \leq M'_2 \leq 2I_d$ and $\frac{1}{2} I_d \leq \Sigma' \leq \frac{2}{\kappa_2(\gamma)} I_d$. From our assumption, we get that (since $\kappa_2(\gamma) \leq 1$ by [Claim 3.7](#))

$$\frac{1}{2} I_d \leq M_2^{-1/2} \tilde{M}_2 M_2^{-1/2} \leq \frac{3}{2} I_d.$$

Left multiplying by $\tilde{M}_2^{-1/2} M_2^{1/2}$ and right multiplying by the transpose of this matrix (which preserves the Loewner order), we get

$$\frac{1}{2} I_d \leq \frac{2}{3} I_d \leq M'_2 = \tilde{M}_2^{-1/2} M_2 \tilde{M}_2^{-1/2} \leq 2I_d,$$

which proves that M'_2 is well-conditioned. For Σ' , recall from [Lemma 4.2](#) that

$$\begin{aligned} M_2 &= \Sigma - (1 - \kappa_2(\gamma)) \cdot \Sigma^{1/2} \left(\frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \left(\frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \right)^\top \right) \Sigma^{1/2} \\ &= \Sigma^{1/2} \left(I_d - (1 - \kappa_2(\gamma)) \cdot \frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \left(\frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \right)^\top \right) \Sigma^{1/2}. \end{aligned}$$

We have

$$\kappa_2(\gamma) I_d \leq I_d - (1 - \kappa_2(\gamma)) \cdot \frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \left(\frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \right)^\top \leq I_d$$

and thus, by left multiplying by $\tilde{M}_2^{-1/2} \Sigma^{1/2}$ and right multiplying by the transpose of this matrix, we get

$$\kappa_2(\gamma) \Sigma' \leq M'_2 \leq \Sigma'.$$

Hence, it follows that

$$\frac{1}{2} I_d \leq M'_2 \leq \Sigma' \leq \frac{1}{\kappa_2(\gamma)} M'_2 \leq \frac{2}{\kappa_2(\gamma)} I_d.$$

Undoing the preconditioning. Assume we have estimates $\hat{\mu}_T$ and $\hat{\Sigma}_T$ in the transformed space. We claim that $\hat{\mu} := \tilde{M}_2^{1/2} \hat{\mu}_T$ and $\hat{\Sigma} := \tilde{M}_2^{1/2} \hat{\Sigma}_T \tilde{M}_2^{1/2}$ have the same error in the original space as $\hat{\mu}_T$ and $\hat{\Sigma}_T$ have in the transformed space. This relies on the fact that the Mahalanobis norm, in which we measure the error, is affine-invariant, as the following claim shows. We defer the proof to [Section A.2](#).

Fact 4.9. Let $A, B \in \mathbb{R}^{d \times d}$ be positive definite matrices. Let $L \in \mathbb{R}^{d \times d}$ be an invertible matrix. Then, we have that

$$\left\| I_d - A^{-\frac{1}{2}} B A^{-\frac{1}{2}} \right\|_F = \left\| I_d - (L A L^\top)^{-\frac{1}{2}} L B L^\top (L A L^\top)^{-\frac{1}{2}} \right\|_F$$

By [Fact 4.9](#) for $A = \Sigma$, $B = \tilde{\Sigma}$ and $L = \tilde{M}^{-1/2}$, we have

$$\left\| I_d - \Sigma^{-1/2} \tilde{\Sigma} \Sigma^{-1/2} \right\|_F = \left\| I_d - \Sigma'^{-1/2} \tilde{\Sigma}_T \Sigma'^{-1/2} \right\|_F.$$

Furthermore, the mean after preconditioning is $\mu' = \tilde{M}_2^{-1/2} \mu$ by [Claim 4.1](#) and thus we also have

$$\begin{aligned} & \left\| \Sigma'^{-1/2} (\hat{\mu}_T - \mu') \right\|^2 \\ &= \left\langle \left(\tilde{M}_2^{-1/2} \Sigma \tilde{M}_2^{-1/2} \right)^{-1/2} \tilde{M}_2^{-1/2} (\hat{\mu} - \mu), \left(\tilde{M}_2^{-1/2} \Sigma \tilde{M}_2^{-1/2} \right)^{-1/2} \tilde{M}_2^{-1/2} (\hat{\mu} - \mu) \right\rangle \\ &= \left\langle \tilde{M}_2^{-1/2} \left(\tilde{M}_2^{-1/2} \Sigma \tilde{M}_2^{-1/2} \right)^{-1} \tilde{M}_2^{-1/2} (\hat{\mu} - \mu), (\hat{\mu} - \mu) \right\rangle \\ &= \langle \Sigma^{-1} (\hat{\mu} - \mu), \hat{\mu} - \mu \rangle = \left\| \Sigma^{-1/2} (\hat{\mu} - \mu) \right\|^2. \end{aligned}$$

Thus, both the mean estimator $\hat{\mu}$ and the covariance estimator $\hat{\Sigma}$ have the same error as $\hat{\mu}_T$ and $\hat{\Sigma}_T$.

Hence, we can indeed precondition the samples by left-multiplying them by $\tilde{M}_2^{-1/2}$, then run our algorithm on the well-conditioned instance and output $\hat{\mu}$ and $\hat{\Sigma}$. The error does not change by this and thus we can without loss of generality assume that our instance is well-conditioned. \square

5 Proof of the main result

In this section, we prove our main result, [Theorem 1.1](#), that we restate below.

Theorem 5.1 (Full statement of [Theorem 1.1](#)). Consider a halfspace truncated Gaussian $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, where $\Sigma \succ 0$. Let α be the mass of the truncation set. Then, given

$$n = \begin{cases} O\left(\frac{d^2}{\varepsilon^2}\right) \text{polylog}(\alpha^{-1}) & \text{if } \alpha \leq 0.99 \\ O\left(\frac{d^2}{\varepsilon^4}\right) & \text{if } \alpha > 0.99 \end{cases}$$

samples, [Algorithm 2](#) computes estimators $\hat{\mu}, \hat{\Sigma}$ such that with probability at least 0.99

$$d_{\text{TV}}\left(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)\right) \leq \varepsilon$$

The runtime of the algorithm is $O(T(n, d))$, where $T(n, d)$ is the time needed to multiply a $d \times n$ matrix with its transpose.

Remark 5.2. We do not need to know α to run our algorithm. It is sufficient that we get as input a lower bound α_0 on α . Then we can replace α by α_0 in the sample complexity bound of [Theorem 5.1](#) (which makes the sample complexity bound only larger) and run the algorithm with $O\left(d^2/\varepsilon^2 \cdot \text{polylog}(\alpha_0^{-1}) + d^2/\varepsilon^4\right)$ samples. If we additionally know whether $\alpha \leq 0.99$ or not, we can run the algorithm with $O\left(d^2/\varepsilon^2 \cdot \text{polylog}(\alpha_0^{-1})\right)$ or $O\left(d^2/\varepsilon^4\right)$ samples respectively.

Algorithm 2: Algorithm for estimating the mean and covariance of a halfspace truncated Gaussian

Input: Dimension d , sample size n , samples $\{x_i\}_{i=1}^n$, target accuracy ε .

Output: Estimated mean $\hat{\mu} \in \mathbb{R}^d$ and covariance $\widehat{\Sigma} \in \mathbb{R}^{d \times d}$.

1. Compute sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$;
 2. Compute sample second central moment: $\widehat{M}_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^{\otimes 2}$;
 3. Draw a standard Gaussian vector: $g \sim \mathcal{N}(0, I_d)$;
 4. Compute the random contraction: $R = (I_d \otimes I_d \otimes g^\top) \cdot \widehat{M}_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^{\otimes 2} \cdot \langle g, x_i - \bar{x} \rangle$.
 5. Compute \hat{u} , the eigenvector of R corresponding to the largest eigenvalue in absolute value (with unit norm).
 6. Estimate the relative truncation parameter: $\hat{\gamma} = \text{skew}^{-1} \left(- \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{3/2} \cdot \left| \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle \right| \right)$.
 7. Flip the sign of \hat{u} if $\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle < 0$.
 8. Estimate the mean of the Gaussian: $\hat{\mu} = \bar{x} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u}$.
 9. Estimate the covariance of the Gaussian: $\widehat{\Sigma} = \widehat{M}_2 + \left(\frac{1}{\kappa_2(\hat{\gamma})} - 1 \right) \cdot \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1} \hat{u} \hat{u}^\top$.
 10. Return estimates $(\hat{\mu}, \widehat{\Sigma})$.
-

Note that we argued in [Section 2](#) that if we would have access to the exact moments, then the estimators $\hat{\mu}$ and $\widehat{\Sigma}$ are equal to μ and Σ respectively. In this section, we prove that even if we only have access to the sample moments, the above estimators work. The proof of this theorem is structured as follows:

- In [Section 5.1](#), we prove that \hat{u} (after the potential sign flip) is good estimator for $-\frac{\Sigma w}{\|\Sigma w\|}$, assuming that \widehat{M}_3 is a good estimator for M_3 .
- In [Section 5.2](#), we prove that $\hat{\gamma}$ is a good estimator for γ , given that we have a good estimator for \hat{u} and that \widehat{M}_2 is a good estimator for M_2 .
- In [Section 5.3](#), we prove that $\hat{\mu}$ and $\widehat{\Sigma}$ are good estimators for μ and Σ , given that \widehat{M}_2 , \hat{u} and $\hat{\gamma}$ are good estimators.
- Finally, in [Section 5.4](#), we put everything together: We determine the sample complexity such that \widehat{M}_2 and \widehat{M}_3 are good enough estimators such that the lemmas proved in [Sections 5.1](#) to [5.3](#) imply the guarantees of [Theorem 5.1](#).

We remark that the assumption $\Sigma > 0$ is not necessary. Suppose Σ is of rank $0 < r < d$, then all samples would live in a subspace of \mathbb{R}^d of dimension r . One can identify this subspace with $O(d)$ samples with high probability, and this reduces the original problem to an r -dimensional problem via an appropriate linear transform (e.g., an orthogonal projection from \mathbb{R}^d to the subspace).

5.1 Recovering the direction of Σw

In order to find the direction of Σw , a natural attempt would be to consider the tensor analogue of low-rank approximations of the sample third moment \widehat{M}_3 . However, unlike the matrix case, it is NP-hard to find the best rank-1 approximation of a general symmetric order-3 tensor [HL13]. To bypass this computational barrier, we reduce the problem to that of finding the eigenvector corresponding to the largest absolute eigenvalue in a certain matrix. We obtain such a matrix by contracting the tensor with a random vector. Throughout this section we will use the notation $u^* = \kappa_3(\gamma)^{1/3} \frac{\Sigma w}{\|\Sigma^{1/2} w\|}$. It then follows that the population third central moment $M_3 = (u^*)^{\otimes 3}$. Our estimator \hat{u} will approximate $\frac{u^*}{\|u^*\|} = -\frac{\Sigma w}{\|\Sigma w\|}$. We will require the following claims for our proof that will follow next. We defer the proofs to Section A.2.

Claim 5.3. *Let $A, B \in \mathbb{R}^{d \times d}$ be symmetric matrices with B being an arbitrary rank-1 matrix. Consider the spectral decomposition of A ,*

$$A = \sum_{i=1}^d \lambda_i v_i v_i^\top$$

where the vectors $\{v_i\}_{i=1}^d$ form an orthonormal basis of \mathbb{R}^d and the singular values $\{\lambda_i\}_{i=1}^d$ satisfy $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_d|$. We then have that

$$\|\lambda_1 v_1 v_1^\top - B\|_F^2 \leq 2 \langle \lambda_1 v_1 v_1^\top - B, A - B \rangle.$$

Claim 5.4. *Let $a, b \in \mathbb{R}^d$. Let $\beta > 0$. Suppose that $\|aa^\top - bb^\top\|_F \leq \beta$. Then, we have that, for $u = \frac{a}{\|a\|}$ and $v = \frac{b}{\|b\|}$,*

$$\min\{\|u + v\|^2, \|u - v\|^2\} \leq 8 \frac{\beta^2}{\|a\|^4}$$

The main result of this section is the following lemma that shows that given an accurate estimate for M_3 , we can recover the direction of u^* . We will show afterwards how many samples we need to get such an estimate for M_3 .

Lemma 5.5. *Given $\beta > 0$ and a symmetric tensor $\widehat{M}_3 \in \mathbb{R}^{d \times d \times d}$ satisfying*

$$\sup_{\|v\|=1} \langle v^{\otimes 3}, M_3 - \widehat{M}_3 \rangle \leq \beta,$$

one can compute a unit vector $\hat{u} \in \mathbb{R}^d$ such that with probability at least $1 - 10^{-4} - e^{-2d}$

$$\min \left\{ \left\| \hat{u} + \frac{u^*}{\|u^*\|} \right\|_2, \left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\|_2 \right\} \leq O \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \beta \right).$$

Proof. To compute the unit vector \hat{u} , we first draw $g \sim \mathcal{N}(0, I_d)$ and compute the matrix

$$R = (I_d \otimes I_d \otimes g^\top) \cdot \widehat{M}_3$$

We then compute the eigenvector corresponding to the largest (absolute) eigenvalue of R . We will show that this gives an estimator \hat{u} with the claimed guarantees.

We begin by observing that $(I_d \otimes I_d \otimes g^\top) \cdot \widehat{M}_3$ is a symmetric matrix. Let \hat{u} be the (unit) eigenvector associated with its largest eigenvalue in absolute value, and λ be the corresponding

eigenvalue. Then $\lambda \cdot \hat{u} \hat{u}^\top$ is the best rank-1 approximation of $(I_d \otimes I_d \otimes g^\top) \cdot \widehat{M}_3$. Our goal is to show that \hat{u} is close to $\frac{u^*}{\|u^*\|}$.

First, we will show that $(I_d \otimes I_d \otimes g^\top) \cdot \widehat{M}_3$ and $(I_d \otimes I_d \otimes g^\top) \cdot M_3$ are close in Frobenius norm. Then, since $(I_d \otimes I_d \otimes g^\top) \cdot M_3 = \langle g, u^* \rangle \cdot u^*(u^*)^\top$ has rank 1, we will be able to use [Claim 5.4](#) that relates closeness in Frobenius norm of rank-1 matrices to the closeness of the vectors themselves to conclude the lemma. Invoking [Claim 5.3](#), we obtain

$$\begin{aligned} \|\lambda \hat{u} \hat{u}^\top - \langle g, u^* \rangle u^*(u^*)^\top\|_F^2 &\leq 2 \langle \lambda \hat{u} \hat{u}^\top - \langle g, u^* \rangle \cdot u^*(u^*)^\top, (I_d \otimes I_d \otimes g^\top) \cdot (\widehat{M}_3 - M_3) \rangle \\ &\stackrel{(i)}{\leq} 2 \|\lambda \hat{u} \hat{u}^\top - \langle g, u^* \rangle \cdot u^*(u^*)^\top\|_* \cdot \|(I_d \otimes I_d \otimes g^\top) \cdot (\widehat{M}_3 - M_3)\| \\ &\stackrel{(ii)}{\leq} 2\sqrt{2} \|\lambda \hat{u} \hat{u}^\top - \langle g, u^* \rangle \cdot u^*(u^*)^\top\|_F \cdot \|(I_d \otimes I_d \otimes g^\top) \cdot (\widehat{M}_3 - M_3)\| \end{aligned}$$

In inequality (i) we used matrix Hölder's inequality. In (ii) we used that $\lambda \hat{u} \hat{u}^\top - \langle g, u^* \rangle u^*(u^*)^\top$ has rank at most 2 and thus its nuclear norm is bounded by $\sqrt{2}$ times its Frobenius norm. We can do the exact same computation with $-g$ instead of g and thus we get

$$\|\lambda \hat{u} \hat{u}^\top \pm \langle g, u^* \rangle u^*(u^*)^\top\|_F \leq 2\sqrt{2} \|(I_d \otimes I_d \otimes g^\top) \cdot (\widehat{M}_3 - M_3)\|.$$

From our assumption we have

$$\|(I_d \otimes I_d \otimes g^\top) \cdot (\widehat{M}_3 - M_3)\| \leq \|g\| \cdot \sup_{\|v\|=1} \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \leq \|g\| \cdot \beta.$$

Using [Fact 3.25](#) about the norm of a Gaussian random vector we have

$$\mathbb{P} \left[\|g\| \geq 3\sqrt{d} \right] \leq \mathbb{P} \left[\|g\| - \mathbb{E}\|g\| \geq 2\sqrt{d} \right] \leq e^{-2d}.$$

Thus, conditioning on the above event we have with probability at least $1 - e^{-2d}$ that

$$\|\lambda \hat{u} \hat{u}^\top \pm \langle g, u^* \rangle u^*(u^*)^\top\|_F \leq 6\sqrt{2} \cdot \sqrt{d} \cdot \beta.$$

Dividing by $|\langle g, u^* \rangle|$ and choosing the appropriate sign, we get that

$$\left\| \frac{|\lambda|}{|\langle g, u^* \rangle|} \hat{u} \hat{u}^\top - u^*(u^*)^\top \right\|_F \leq 6\sqrt{2} \cdot \sqrt{d} \cdot \beta \cdot \frac{1}{|\langle g, u^* \rangle|}.$$

Since $\langle g, u^* \rangle$ is a Gaussian of variance $\|u^*\|^2$, [Fact 3.24](#) implies that with probability at least $1 - 10^{-4}$ we have $|\langle g, u^* \rangle| \geq 10^{-4} \|u^*\|$. Combining this with the above using a union bound, with probability $1 - e^{-2d} - 10^{-4}$ we have

$$\left\| \frac{|\lambda|}{|\langle g, u^* \rangle|} \hat{u} \hat{u}^\top - u^*(u^*)^\top \right\|_F \leq 6\sqrt{2} \cdot \sqrt{d} \cdot \beta \cdot \frac{1000}{\|u^*\|}.$$

Using [Claim 5.4](#) that relates Frobenius closeness to closeness of vectors themselves we obtain

$$\min\{\|\hat{u} - \tilde{u}\|, \|\hat{u} + \tilde{u}\|\} \leq \frac{24000\sqrt{d} \cdot \beta}{\|u^*\|^3},$$

where $\tilde{u} := \frac{u^*}{\|u^*\|}$. Now, since $\Sigma \geq \frac{1}{2}I_d$, we get that

$$\|u^*\|^3 = |\kappa_3(\gamma)| \cdot \left\| \frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right\|^3 \geq |\kappa_3(\gamma)| \cdot \lambda_{\min}(\Sigma^{1/2})^3 \geq \frac{|\kappa_3(\gamma)|}{2\sqrt{2}}.$$

Therefore we have

$$\min\{\|\hat{u} - \tilde{u}\|, \|\hat{u} + \tilde{u}\|\} \lesssim \frac{\sqrt{d} \cdot \beta}{|\kappa_3(\gamma)|},$$

completing the proof. □

So far we have shown how to recover a unit vector that is (approximately) parallel to Σw using a random contraction on the sample third moment. To this end, we required a strong assumption on the finite-sample error to which we learn the third moment in the *weaker* injective tensor norm as opposed to the Frobenius norm. In particular, we required this error to be as small as inverse polynomial in the dimension (up to κ_2, κ_3 factors) for the statement of the previous lemma. We will now show that we can indeed achieve such a high-accuracy estimate in the injective tensor norm using only $O(d^2)$ samples (up to κ_2, κ_3 factors). We start with the following claim that we will require.

Claim 5.6. *Let v be the mean of $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$. Let v be a fixed unit vector. Given n i.i.d. samples $\{x_i\}_{i=1}^n$ from $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, for every $t > 0$, we have that*

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^3 - \langle v, u^* \rangle^3 \right| \geq C \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{t}{n}} + \frac{t^{3/2}}{n} \right) \right] \leq 2 \exp(-t).$$

for an absolute constant $C > 0$.

Proof. The above claim follows by a direct application of [Corollary 3.29](#) for the concentration of averages-of-cubes of sub-Gaussian random variables that we proved earlier. Since $\langle v, x_i - v \rangle$ is sub-Gaussian with parameter $O\left(1/\sqrt{\kappa_2(\gamma)}\right)$ (after preconditioning) and $\mathbb{E}[\langle v, x_i - v \rangle^3] = \langle v, u^* \rangle^3$, we apply [Corollary 3.29](#) with $Z_i = \langle v, x_i - v \rangle$ and $\sigma = O\left(1/\sqrt{\kappa_2(\gamma)}\right)$ to obtain the result. □

We remark that the above claim argues about concentration of *zero-mean* random variables. We will now account for the error incurred by finite sample approximations to v . More formally we show the following.

Claim 5.7. *Let v be the mean of $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$. Let v be a fixed unit vector. Let $\{x_i\}_{i=1}^n$ be i.i.d samples from $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$ and let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Then, we have the following with probability at least $1 - 4e^{-10d}$*

$$\left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - \bar{x} \rangle^3 - \langle v, x_i - v \rangle^3 \right| \lesssim \kappa_2(\gamma)^{-3/2} \cdot \left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{3/2} \right).$$

We defer the proof of [Claim 5.7](#) to the [Section A.3](#). Given the above concentration results, we will now present the main sample complexity result for this section. For $n = O(d^2)$ (up to κ_2, κ_3 factors), the result below enables us to satisfy the requirements for the main result of this section [Lemma 5.5](#).

Lemma 5.8. Let $\{x_i\}_{i=1}^n$ be i.i.d. samples from $\mathcal{N}(\mu, \Sigma)_{\langle w, x \rangle \leq \tau}$. Let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Let $\widehat{M}_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^{\otimes 3}$. Then we have with probability at least $1 - 6e^{-7d}$ that the following holds

$$\sup_{\|v\|=1} \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \lesssim \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right).$$

We give a proof sketch here and show that the above holds for any fixed v with probability $1 - 6e^{-10d}$. The lemma then follows by a standard ε -net argument over the unit sphere. We defer the full proof to [Section A.3](#).

Proof sketch. For a fixed vector v , taking $t = \Theta(d)$ in [Claim 5.6](#), we have for an absolute constant $K_1 > 0$ that

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^3 - \langle v, u^* \rangle^3 \right| \geq K_1 \cdot \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right) \right] \leq 2e^{-10d}.$$

For the same fixed choice of v , we have for some other absolute constant $K_2 > 0$ from [Claim 5.7](#) that

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - \bar{x} \rangle^3 - \langle v, x_i - v \rangle^3 \right| \geq K_2 \cdot \kappa_2(\gamma)^{-3/2} \cdot \left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n} \right)^{3/2} \right) \right] \leq 4e^{-10d}.$$

Now since we have that

$$\left| \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \right| \leq \left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - \bar{x} \rangle^3 - \langle v, x_i - v \rangle^3 \right| + \left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^3 - \langle v, u^* \rangle^3 \right|,$$

we can conclude that

$$\left| \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \right| \leq K \cdot \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right)$$

with probability at least $1 - 6e^{-10d}$ via a union bound. So far, we have argued that for a fixed vector v , it is the case that $\left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle$ is well concentrated. What remains is to show that the supremum of the quantity over the unit sphere can still be controlled. To this end, we construct an ε -net, and complete the proof with a union bound argument over this net in [Section A.3](#). \square

5.2 Relative truncation parameter recovery

Our goal in the section will be to recover the relative truncation parameter γ . From the previous section, we have a good estimate for the direction of Σw . In this section we will show how we can use this vector to recover the relative truncation parameter γ . Recovering γ allows us to recover the parameters of the Gaussian *directly* from the moments. We first show how to get an estimate $\widehat{\gamma}$ for $\text{skew}(\gamma)$. Then, we show that this implies generally (meaning without any assumptions except that $\widehat{\gamma}$ and $\text{skew}(\gamma)$ are close) that also $\widehat{\gamma}$ is close to γ .

Lemma 5.9. Let $0 < \beta < c$, where c is some sufficiently small absolute constant. Assume that

$$\min \left\{ \left\| \widehat{u} + \frac{u^*}{\|u^*\|} \right\|_2, \left\| \widehat{u} - \frac{u^*}{\|u^*\|} \right\|_2 \right\} \leq \beta,$$

$$\|M_2^{-1/2}\widehat{M}_2M_2^{-1/2} - I_d\| \leq \beta$$

for a unit vector \hat{u} and a symmetric matrix $\widehat{M}_2 > 0$. Furthermore, let \widehat{M}_3 be a symmetric tensor satisfying

$$\sup_{\|v\|=1} \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \leq \beta.$$

Denoting by skew^{-1} the inverse function of skew , consider the estimator

$$\hat{\gamma} := \text{skew}^{-1} \left(- \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{3/2} \cdot |\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle| \right).$$

Then, we have that

$$|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq O \left(\beta \cdot \kappa_2(\gamma)^{-3/2} \right)$$

as well as

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \tilde{u}^\top M_2^{-1} \tilde{u} \right| \leq O(\beta).$$

At a high level, the above lemma say that we can learn the relative truncation parameter accurately given access to high-accuracy estimates for Σw , M_2 and M_3 by inverting the one-dimensional skew function. We will require the following claim that bounds $\|u^*\|$.

Claim 5.10. For $u^* := \kappa_3(\gamma)^{1/3} \frac{\Sigma w}{\|\Sigma^{1/2} w\|}$, we have $\frac{1}{\sqrt{2}} |\kappa_3(\gamma)|^{1/3} \leq \|u^*\| \leq \sqrt{2} \cdot |\text{skew}(\gamma)|^{1/3}$.

Proof. First, note that $\frac{\Sigma w}{\|\Sigma^{1/2} w\|} = \Sigma^{1/2} \cdot \frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|}$. By [Lemma 4.6](#), we have $\frac{1}{\sqrt{2}} I_d \leq \Sigma^{1/2} \leq \frac{\sqrt{2}}{\kappa_2(\gamma)^{1/2}} I_d$ and thus

$$\frac{1}{\sqrt{2}} \leq \left\| \Sigma^{1/2} \frac{\Sigma^{1/2} w}{\|\Sigma^{1/2} w\|} \right\| \leq \frac{\sqrt{2}}{\kappa_2(\gamma)^{1/2}}.$$

This proves the claim by recalling that $\text{skew}(\gamma) = \kappa_3(\gamma)/\kappa_2(\gamma)^{3/2}$. \square

Proof of [Lemma 5.9](#). The proof of the lemma has two main parts: a population level analysis and a finite-sample error analysis. We further split the finite sample error analysis into two sub-parts.

We will now begin with a population-level analysis that will show that knowing Σw , M_2 and M_3 will allow us to recover γ . Recalling the definition of the population second central moment we have

$$M_2 = \Sigma - \frac{1 - \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \cdot (\Sigma w) (\Sigma w)^\top$$

This implies that when applying the linear transformation M_2 on w , we get that

$$M_2 w = \kappa_2(\gamma) \cdot \Sigma w \implies w = \kappa_2(\gamma) \cdot M_2^{-1} \Sigma w$$

Now by left-multiplying both sides by $(\Sigma w)^\top$ and dividing by $\kappa_2(\gamma)$ we obtain that

$$(\Sigma w)^\top M_2^{-1} (\Sigma w) = \frac{\|\Sigma^{1/2} w\|^2}{\kappa_2(\gamma)}.$$

Recall that $u^* = \kappa_3(\gamma)^{1/3} \frac{\Sigma w}{\|\Sigma^{1/2} w\|}$. We define $\tilde{u} = \frac{u^*}{\|u^*\|} = -\frac{\Sigma w}{\|\Sigma w\|}$ as the unit vector along the direction of u^* (note that $\kappa_3(\gamma) < 0$ by [Claim 3.8](#)). Then we have

$$\tilde{u}^\top M_2^{-1} \tilde{u} = \frac{\|\Sigma^{1/2} w\|^2}{\kappa_2(\gamma) \|\Sigma w\|^2} = \frac{\kappa_3(\gamma)^{2/3}}{\kappa_2(\gamma) \|u^*\|^2} \quad (5.1)$$

Furthermore recalling that the population third central moment is $M_3 = (u^*)^{\otimes 3}$ and taking its projection along \tilde{u} we have that

$$\langle M_3, \tilde{u}^{\otimes 3} \rangle = \langle u^*, \tilde{u} \rangle^3 = -\kappa_3(\gamma) \left\langle \frac{\Sigma w}{\|\Sigma^{1/2} w\|}, \frac{\Sigma w}{\|\Sigma w\|} \right\rangle^3 = |\kappa_3(\gamma)| \cdot \frac{\|\Sigma w\|^3}{\|\Sigma^{1/2} w\|^3} = \|u^*\|^3.$$

This naturally gives us the following identity that we will need for building our estimator:

$$-(\tilde{u}^\top M_2^{-1} \tilde{u})^{3/2} \cdot \langle M_3, \tilde{u}^{\otimes 3} \rangle = \frac{\kappa_3(\gamma)}{\kappa_2(\gamma)^{3/2}}$$

In the rest of the proof, we will carry out finite-sample error analysis since we only have access to the sample moments. We will first deal with the first term on the left hand side above.

Error analysis for $(\tilde{u}^\top M_2^{-1} \tilde{u})^{3/2}$

Observe that, since \widehat{M}_2^{-1} is symmetric, we get that

$$\begin{aligned} \hat{u}^\top \widehat{M}_2^{-1} \hat{u} &= \tilde{u}^\top \widehat{M}_2^{-1} \tilde{u} + (\hat{u} - \tilde{u})^\top \widehat{M}_2^{-1} (\hat{u} + \tilde{u}) \\ &= \tilde{u}^\top M_2^{-1} \tilde{u} + \tilde{u}^\top (\widehat{M}_2^{-1} - M_2^{-1}) \tilde{u} + (\hat{u} - \tilde{u})^\top \widehat{M}_2^{-1} (\hat{u} + \tilde{u}) \\ &= \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} + \tilde{u}^\top (\widehat{M}_2^{-1} - M_2^{-1}) \tilde{u} + (\hat{u} - \tilde{u})^\top \widehat{M}_2^{-1} (\hat{u} + \tilde{u}) \end{aligned}$$

We will now bound the second and third term in the last equation above. For the second term, since $(1 - \beta)M_2 \leq \widehat{M}_2 \leq (1 + \beta)M_2$, it follows that $\frac{1}{1+\beta}M_2^{-1} \leq \widehat{M}_2^{-1} \leq \frac{1}{1-\beta}M_2^{-1}$. Therefore we have

$$\begin{aligned} \|M_2^{-1} - \widehat{M}_2^{-1}\| &= \|\widehat{M}_2^{-1}(\widehat{M}_2 - M_2)M_2^{-1}\| \\ &= \|\widehat{M}_2^{-1}M_2^{1/2}(M_2^{-1/2}\widehat{M}_2M_2^{-1/2} - I_d)M_2^{-1/2}\| \\ &\leq \beta \cdot \|\widehat{M}_2^{-1}\| \cdot \|M_2^{1/2}\| \cdot \|M_2^{-1/2}\| \\ &\stackrel{(i)}{\leq} \frac{4\beta}{1-\beta} \leq 8\beta, \end{aligned}$$

as long as $\beta < \frac{1}{2}$, where the inequality (i) follows from using our preconditioning guarantees proved in [Lemma 4.6](#). Therefore

$$\left| \tilde{u}^\top (\widehat{M}_2^{-1} - M_2^{-1}) \tilde{u} \right| \leq 8\beta.$$

Now for the third term we have that

$$|(\hat{u} - \tilde{u})^\top \widehat{M}_2^{-1} (\hat{u} + \tilde{u})| \leq \|\hat{u} - \tilde{u}\| \cdot \|\hat{u} + \tilde{u}\| \cdot \|\widehat{M}_2^{-1}\| \leq O(\beta)$$

whenever $\beta < 1/2$. Putting the two bounds together, we obtain,

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \tilde{u}^\top M_2^{-1} \tilde{u} \right| = \left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \right| \leq O(\beta)$$

We are almost done as we have shown closeness of the quadratic forms. We will now show that raising the quadratic forms to a higher power of $3/2$ still preserves closeness and conclude this part of the proof next. We will do this by an application of the mean value theorem from univariate calculus. To this end, we will try to obtain a bound on the range of values that the quadratic forms can take. Firstly, it is clear that

$$\hat{u}^\top \widehat{M}_2^{-1} \hat{u} > 0, \quad \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} > 0.$$

Now since $\|u^*\| \geq |\kappa_3(\gamma)^{1/3}| \cdot \lambda_{\min}(\Sigma^{1/2})$ we have that

$$\frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \leq \frac{1}{\kappa_2(\gamma) \lambda_{\min}^2(\Sigma^{1/2})} = \frac{1}{\kappa_2(\gamma) \lambda_{\min}(\Sigma)} \leq \frac{2}{\kappa_2(\gamma)} \quad (5.2)$$

where we used the fact that $\Sigma \geq \frac{1}{2}I_d$. Applying the mean value theorem for the function $f(x) = x^{3/2}$, we know that there exists a B between $a = \hat{u}^\top \widehat{M}_2^{-1} \hat{u}$ and $c = \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)}$ such that the following holds

$$|f(a) - f(c)| = f'(B) \cdot |a - c| = \frac{3}{2} \cdot \sqrt{B} \cdot |a - c|.$$

We have $B = O(1/\kappa_2(\gamma))$ since

$$\begin{aligned} B &\leq \max \left\{ \hat{u}^\top \widehat{M}_2^{-1} \hat{u}, \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \right\} \leq \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} + \left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \right| \\ &\stackrel{(i)}{\leq} \frac{2}{\kappa_2(\gamma)} + O(\beta) \stackrel{(ii)}{\leq} O\left(\frac{1}{\kappa_2(\gamma)}\right). \end{aligned}$$

where (i) follows from inequality (5.2) and (ii) holds as long as β is at most a sufficiently small constant. Therefore we have that

$$\left| \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{3/2} - \frac{|\kappa_3(\gamma)|}{\|u^*\|^3 \cdot \kappa_2(\gamma)^{3/2}} \right| \leq O(\beta \cdot \kappa_2(\gamma)^{-1/2}),$$

concluding the first part of our error analysis.

Error analysis for $\langle M_3, \tilde{u}^{\otimes 3} \rangle$

We will first consider the case that $\|\hat{u} - \tilde{u}\| \leq \beta$ and analyze the case $\|\hat{u} + \tilde{u}\| \leq \beta$ afterwards. We have that

$$\begin{aligned} \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle &= \langle \hat{u}^{\otimes 3}, \widehat{M}_3 - M_3 \rangle + \langle \hat{u}^{\otimes 3}, M_3 \rangle \\ &= \langle \hat{u}^{\otimes 3}, \widehat{M}_3 - M_3 \rangle + \|u^*\|^3 \cdot \left(1 + 3 \langle \hat{u} - \tilde{u}, \tilde{u} \rangle + 3 \langle \hat{u} - \tilde{u}, \tilde{u} \rangle^2 + \langle \hat{u} - \tilde{u}, \tilde{u} \rangle^3 \right). \end{aligned}$$

By assumption, we have that $\sup_{\|v\|=1} \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \leq \beta$. Thus, by applying Cauchy-Schwarz and using $\|\hat{u} - \tilde{u}\| \leq \beta$, we get that

$$\left| \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle - \|u^*\|^3 \right| \leq \beta + O(\beta \cdot \|u^*\|^3) \leq O(\beta),$$

where we also used that $\|u^*\|^3 \leq 2\sqrt{2}|\text{skew}(\gamma)| \leq O(1)$ by [Claims 3.18](#) and [5.10](#). Since $\|u^*\|^3 > 0$, we even get that

$$\left| \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle - \|u^*\|^3 \right| \leq O(\beta). \quad (5.3)$$

This also holds for the case $\|\hat{u} + \tilde{u}\| \leq \beta$ by repeating the above argument for $-\hat{u}$ instead of \hat{u} .

Putting the pieces together

We will now look at the total aggregate error that we incur. We begin by defining

$$S_1 = \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{3/2}, \quad S_2 = \left| \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle \right|$$

and

$$\varepsilon_1 = S_1 - \frac{|\kappa_3(\gamma)|}{\|u^*\|^3 \cdot \kappa_2(\gamma)^{3/2}}, \quad \varepsilon_2 = S_2 - \|u^*\|^3.$$

By our computations earlier, we know that $|\varepsilon_1| \leq O(\beta \cdot \kappa_2(\gamma)^{-1/2})$ and $|\varepsilon_2| \leq O(\beta)$. Now expanding $S_1 S_2$ we have

$$\begin{aligned} S_1 S_2 &= \left(\frac{|\kappa_3(\gamma)|}{\|u^*\|^3 \cdot \kappa_2(\gamma)^{3/2}} + \varepsilon_1 \right) (\|u^*\|^3 + \varepsilon_2) \\ &= \frac{|\kappa_3(\gamma)|}{\|u^*\|^3 \cdot \kappa_2(\gamma)^{3/2}} \cdot \|u^*\|^3 + \varepsilon_1 \cdot \|u^*\|^3 + \varepsilon_2 \cdot \frac{|\kappa_3(\gamma)|}{\|u^*\|^3 \cdot \kappa_2(\gamma)^{3/2}} + \varepsilon_1 \varepsilon_2. \end{aligned}$$

Now rearranging the above inequality gives us

$$\begin{aligned} |\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| &= \left| S_1 S_2 - \frac{|\kappa_3(\gamma)|}{\kappa_2(\gamma)^{3/2}} \right| \\ &\leq O\left(\beta \cdot \kappa_2(\gamma)^{-1/2}\right) \cdot \|u^*\|^3 + O(\beta) \cdot \frac{|\text{skew}(\gamma)|}{\|u^*\|^3} + O\left(\beta^2 \cdot \kappa_2(\gamma)^{-1/2}\right) \\ &\leq O\left(\beta \cdot \kappa_2(\gamma)^{-1/2}\right) + O\left(\beta \cdot \frac{|\text{skew}(\gamma)|}{|\kappa_3(\gamma)|}\right) + O\left(\beta^2 \cdot \kappa_2(\gamma)^{-1/2}\right) \\ &\leq O\left(\beta \cdot \left(\kappa_2(\gamma)^{-1/2} + \kappa_2(\gamma)^{-3/2} + \beta \cdot \kappa_2(\gamma)^{-1/2}\right)\right) \\ &\leq O\left(\beta \cdot \kappa_2(\gamma)^{-3/2}\right), \end{aligned} \quad (5.4)$$

where the second inequality follows from our bounds on $\|u^*\|$ (cf. [Claims 3.18](#) and [5.10](#)) and the last inequality follows from $|\kappa_2(\gamma)| \leq 1$ (cf. [Claim 3.7](#)) and $\beta \leq 1$ by assumption. This completes the proof of [Lemma 5.9](#). \square

5.3 Relative truncation parameter recovery implies parameter recovery

In this section, we show that given an estimate for the relative truncation parameter as well as the vector $\tilde{u} = -\frac{\sum w}{\|\sum w\|}$ and the first three moments, we can estimate the mean and covariance of the (untruncated) Gaussian.

5.3.1 Mean estimation

The following lemma shows that we can recover the mean.

Lemma 5.11. *Suppose there is $\beta > 0$ such that $\beta < c \cdot \kappa_2(\gamma)^{5/2}$, for some sufficiently small absolute constant $c > 0$, and there are \bar{x} , \widehat{M}_2 , \hat{u} , $\hat{\gamma}$ such that the following hold*

$$\left\| \bar{x} - \mu - \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} \right\| \leq \beta \quad (5.5)$$

$$\|\Sigma^{-1/2} (\widehat{M}_2 - M_2) \Sigma^{-1/2}\|_F \leq \beta \quad (5.6)$$

$$\left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\| = \left\| \hat{u} + \frac{\Sigma w}{\|\Sigma w\|} \right\| \leq \beta, \quad (5.7)$$

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \right| \leq \beta \quad (5.8)$$

$$|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{3/2}}\right). \quad (5.9)$$

Then the mean estimator

$$\hat{\mu} = \bar{x} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u}$$

satisfies

$$\|\hat{\mu} - \mu\| \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right).$$

Proof. First, using assumption (5.5), we have that

$$\begin{aligned} \|\hat{\mu} - \mu\| &= \left\| \bar{x} - \mu - \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} + \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} \right\| \\ &\leq \left\| \bar{x} - \mu - \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} \right\| + \left\| \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} \right\| \\ &\leq \beta + \underbrace{\left\| \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} \right\|}_{(*)}. \end{aligned}$$

We now want to bound (*). We can furthermore split this up as follows

$$\begin{aligned} (*) &\leq \left\| \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} \right\| \\ &\quad + \underbrace{\left\| \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} - \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} \right\|}_{a_1:=} \\ &\leq a_1 + \underbrace{\left\| \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} + \kappa_1(\gamma) \frac{\|\Sigma w\|}{\|\Sigma^{1/2} w\|} \hat{u} \right\|}_{a_2:=} \end{aligned}$$

$$+ \underbrace{\left\| \kappa_1(\gamma) \frac{\|\Sigma w\|}{\|\Sigma^{1/2} w\|} \hat{u} - \frac{\kappa_1(\gamma)}{\sqrt{\kappa_2(\gamma)}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u} \right\|}_{a_3:=}$$

Putting these together, we get that

$$\|\hat{\mu} - \mu\| \leq \beta + a_1 + a_2 + a_3$$

and it remains to bound a_1 , a_2 and a_3 . We do so now.

Bounding a_1 . We have that

$$a_1 \leq \left| \frac{\kappa_1(\gamma)}{\sqrt{\kappa_2(\gamma)}} - \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \right| \cdot \left| \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \right| \cdot \|\hat{u}\|.$$

We have that $\|\hat{u}\| \leq 1$. Thus, defining $\psi(t) = \frac{\kappa_1(t)}{\sqrt{\kappa_2(t)}}$ we get that

$$a_1 \leq \left| \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \right| \cdot |\psi(\hat{\gamma}) - \psi(\gamma)|.$$

By [Claim 3.20](#), we have that

$$|\psi(\hat{\gamma}) - \psi(\gamma)| \lesssim |\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \cdot \max\{\kappa_2(\hat{\gamma})^{-2}, \kappa_2(\gamma)^{-2}\}.$$

We can use assumption (5.9) to bound the first term, so it remains to give a lower bound on $\kappa_2(\hat{\gamma})$. It follows from [Claim 3.19](#) and assumption (5.9) that

$$|\kappa_2(\gamma) - \kappa_2(\hat{\gamma})| \lesssim |\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq \frac{\beta}{\kappa_2(\gamma)^{3/2}}$$

Moreover, by choosing a sufficiently small constant in our assumption $\beta < c \cdot \kappa_2(\gamma)^{5/2}$, we have that $\kappa_2(\hat{\gamma}) \geq \kappa_2(\gamma) - \frac{1}{2}\kappa_2(\gamma) = \frac{1}{2}\kappa_2(\gamma)$. Thus, we get that

$$|\psi(\hat{\gamma}) - \psi(\gamma)| \lesssim |\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \cdot \kappa_2(\gamma)^{-2} \leq \frac{\beta}{\kappa_2(\gamma)^{7/2}}.$$

On the other hand, by [Lemma 4.6](#), we have that $M_2 \leq 2I_d$, which together with assumption (5.6) implies

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right|^{-1/2} \leq \lambda_{\min} \left(\widehat{M}_2^{-1} \right)^{-1/2} = \left\| \widehat{M}_2 \right\|^{1/2} \leq \sqrt{(1+\beta)\|M_2\|} \leq 2,$$

as long as $\beta \leq 1$. Together with the above this implies

$$a_1 \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right).$$

Bounding a_2 . For a_2 , we have that

$$a_2 \leq |\kappa_1(\gamma)| \cdot \frac{\|\Sigma w\|}{\|\Sigma^{1/2} w\|} \cdot \left\| \frac{\Sigma w}{\|\Sigma w\|} + \hat{u} \right\|.$$

We have that $|\kappa_1(\gamma)| \leq O(\kappa_2(\gamma)^{-1/2})$ by [Claim 3.9](#) and that $\frac{\|\Sigma w\|}{\|\Sigma^{1/2}w\|} \leq \kappa_2(\gamma)^{-1/2}$ by [Lemma 4.6](#). Together with assumption [\(5.7\)](#), this implies

$$a_2 \leq O\left(\frac{\beta}{\kappa_2(\gamma)}\right).$$

Bounding a_3 . For a_3 , we have that

$$\begin{aligned} a_3 &\leq \left| \kappa_1(\gamma) \frac{\|\Sigma w\|}{\|\Sigma^{1/2}w\|} - \frac{\kappa_1(\gamma)}{\sqrt{\kappa_2(\gamma)}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \right| \cdot \|\hat{u}\| \\ &\leq \frac{|\kappa_1(\gamma)|}{\sqrt{\kappa_2(\gamma)}} \cdot \left| \sqrt{\kappa_2(\gamma)} \frac{\|\Sigma w\|}{\|\Sigma^{1/2}w\|} - \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \right|. \end{aligned}$$

Now, consider $f(x) = \sqrt{x}$. For $a, c \geq B > 0$, we have that, by the mean value theorem,

$$|f(a) - f(c)| \leq \frac{1}{2\sqrt{B}} |a - c|.$$

We want to apply this for $a = \kappa_2(\gamma) \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2}w\|^2}$ and $c = \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1}$. By [Lemma 4.6](#), we have that $a \geq \frac{1}{2} \kappa_2(\gamma)$. From assumption [\(5.8\)](#) and using also $u^* = \kappa_3(\gamma)^{1/3} \frac{\Sigma w}{\|\Sigma^{1/2}w\|}$, we furthermore have

$$c \geq \kappa_2(\gamma) \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2}w\|^2} - \beta \geq \frac{1}{2} \kappa_2(\gamma) - \beta \geq \frac{1}{4} \kappa_2(\gamma),$$

as long as $\beta \leq \frac{\kappa_2(\gamma)}{4}$. Hence, we can choose $B = \frac{1}{4} \kappa_2(\gamma)$. Thus, we get that

$$\begin{aligned} \left| \sqrt{\kappa_2(\gamma)} \frac{\|\Sigma w\|}{\|\Sigma^{1/2}w\|} - \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \right| &\leq \frac{1}{\sqrt{\kappa_2(\gamma)}} \cdot \left| \kappa_2(\gamma) \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2}w\|^2} - \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1} \right| \\ &\leq \frac{\beta}{\sqrt{\kappa_2(\gamma)}} \end{aligned}$$

by assumption [\(5.8\)](#). Together with the above as well as [Claim 3.9](#), we get that

$$a_3 \leq \beta \cdot \frac{|\kappa_1(\gamma)|}{\kappa_2(\gamma)} \leq O\left(\beta \cdot \kappa_2(\gamma)^{-3/2}\right).$$

Completing the proof. Combining the bounds on a_1 , a_2 and a_3 , we get that

$$\|\hat{\mu} - \mu\| \leq \beta + O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right) + O\left(\frac{\beta}{\kappa_2(\gamma)}\right) + O\left(\frac{\beta}{\kappa_2(\gamma)^{3/2}}\right) \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right),$$

where we used $\kappa_2(\gamma) \leq 1$ and $\kappa_3 \leq O(1)$ by [Claims 3.7](#) and [3.10](#) for the last inequality. \square

5.3.2 Covariance estimation

For covariance estimation, we need the following fact, whose proof we defer to [Section A.2](#).

Fact 5.12. For any $a, b \in \mathbb{R}^d$ we have that $\|aa^\top - bb^\top\|_F \leq \|a + b\| \cdot \|a - b\|$.

Lemma 5.13. Suppose there is $\beta > 0$ such that $\beta < c \cdot \kappa_2(\gamma)^{5/2}$, for a sufficiently small absolute constant $c > 0$, and $\widehat{M}_2, \hat{u}, \hat{\gamma}$ such that

$$\|\Sigma^{-1/2} (\widehat{M}_2 - M_2) \Sigma^{-1/2}\|_F \leq \beta \quad (5.10)$$

$$\min \left\{ \left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\|, \left\| \hat{u} + \frac{u^*}{\|u^*\|} \right\| \right\} \leq \beta, \quad (5.11)$$

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \right| \leq \beta \quad (5.12)$$

$$|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq \frac{\beta}{\kappa_2(\gamma)^{3/2}}. \quad (5.13)$$

Then the covariance estimator

$$\widehat{\Sigma} = \widehat{M}_2 + \left(\frac{1}{\kappa_2(\hat{\gamma})} - 1 \right) (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \cdot \hat{u} \hat{u}^\top$$

satisfies

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\|_F \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}} \right).$$

Proof. By the definition of our estimator, we get that

$$\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d = \Sigma^{-1/2} \widehat{M}_2 \Sigma^{-1/2} + \left(\frac{1}{\kappa_2(\hat{\gamma})} - 1 \right) (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} - I_d.$$

If all $\widehat{M}_2, \hat{\gamma}$ and \hat{u} would be equal to M_2, γ and \tilde{u} , then this norm would be zero. The proof now consists of accounting for the errors in these estimates. First, substituting \widehat{M}_2 by M_2 , we get that

$$\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d = A_1 + \Sigma^{-1/2} M_2 \Sigma^{-1/2} + \left(\frac{1}{\kappa_2(\hat{\gamma})} - 1 \right) (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} - I_d,$$

where

$$A_1 = \Sigma^{-1/2} (\widehat{M}_2 - M_2) \Sigma^{-1/2}.$$

Now, we can use the formula for M_2 from [Lemma 4.2](#) to get that

$$\begin{aligned} \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d = & \\ & A_1 + (1 - \kappa_2(\gamma)) \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma^{1/2} w\|^2} - \underbrace{\left(\frac{1}{\kappa_2(\hat{\gamma})} - 1 \right) (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2}}_{(*)}. \end{aligned}$$

We now want to show that

$$(*) \approx (1 - \kappa_2(\gamma)) \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma^{1/2} w\|^2}.$$

Note that if we would replace $\widehat{M}_2, \hat{\gamma}$ and \hat{u} by M_2, γ and \tilde{u} , then they would be equal. We will do so now and account for the error. Substituting $\hat{\gamma}$ by γ , we get that

$$(*) = A_2 + \left(\frac{1}{\kappa_2(\gamma)} - 1 \right) (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2},$$

where

$$A_2 = \left(\frac{1}{\kappa_2(\hat{\gamma})} - \frac{1}{\kappa_2(\gamma)} \right) (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2}.$$

Substituting $\hat{u}^\top \widehat{M}_2^{-1} \hat{u}$ by $\frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)}$, we get that

$$\begin{aligned} (*) &= A_2 + A_3 + \left(\frac{1}{\kappa_2(\gamma)} - 1 \right) \left(\frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \cdot \kappa_2(\gamma)} \right)^{-1} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} \\ &= A_2 + A_3 + (1 - \kappa_2(\gamma)) \frac{\|u^*\|^2}{\kappa_3(\gamma)^{2/3}} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2}, \end{aligned}$$

where

$$A_3 = \left(\frac{1}{\kappa_2(\gamma)} - 1 \right) \left((\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|u^*\|^2 \cdot \kappa_2(\gamma)}{\kappa_3(\gamma)^{2/3}} \right) \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2}.$$

Finally, substituting \hat{u} with $\tilde{u} = \frac{\Sigma w}{\|\Sigma w\|}$ and using that $\frac{\|\Sigma w\|}{\|\Sigma^{1/2} w\|} = \frac{\|u^*\|}{\kappa_3(\gamma)^{1/3}}$, we get that

$$\begin{aligned} (*) &= A_2 + A_3 + A_4 + (1 - \kappa_2(\gamma)) \frac{\|u^*\|^2}{\kappa_3(\gamma)^{2/3}} \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma w\|^2} \\ &= A_2 + A_3 + A_4 + (1 - \kappa_2(\gamma)) \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma^{1/2} w\|^2}, \end{aligned}$$

where

$$A_4 = (1 - \kappa_2(\gamma)) \left(\frac{\|u^*\|^2}{\kappa_3(\gamma)^{2/3}} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} - \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma^{1/2} w\|^2} \right).$$

Combining all of the above we get that

$$\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d = A_1 - A_2 - A_3 - A_4.$$

Thus, in order to bound $\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\|_F$, it is sufficient to bound the norms of A_1, A_2, A_3 and A_4 . We will do so now.

Bounding $\|A_1\|_F$. For A_1 , note that by (5.10) we have that $\|A_1\|_F \leq \beta$.

Bounding $\|A_2\|_F$. For A_2 , we can first bound

$$\|A_2\|_F \leq \left| \frac{1}{\kappa_2(\hat{\gamma})} - \frac{1}{\kappa_2(\gamma)} \right| \cdot \left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \right| \left\| \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} \right\|_F.$$

We bound these three terms separately. First, note that, by Lemma 4.6,

$$\left\| \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} \right\|_F = \left\| \Sigma^{-1/2} \hat{u} \right\|^2 \leq \left\| \Sigma^{-1/2} \right\|^2 \leq 2. \quad (5.14)$$

For the middle term, we first want to bound

$$\left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|u^*\|^2 \cdot \kappa_2(\gamma)}{\kappa_3(\gamma)^{2/3}} \right| = \left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \right|,$$

where we used $u^* = \kappa_3(\gamma)^{1/3} \frac{\Sigma w}{\|\Sigma^{1/2} w\|}$ (which implies $\frac{\|u^*\|^2 \cdot \kappa_2(\gamma)}{\kappa_3(\gamma)^{2/3}} = \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2}$). We have that, using assumption (5.12),

$$\left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \right| = \left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)} \right| \cdot \left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \cdot \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \right|$$

$$\leq \beta \cdot \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \cdot \|\Sigma^{1/2} w\|^2}.$$

Using (5.12) again to lower bound $\hat{u}^\top \widehat{M}_2^{-1} \hat{u}$, we can furthermore bound

$$\begin{aligned} \left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \right| &\leq \beta \cdot \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \cdot \left(\frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)} - \beta \right)^{-1} \\ &= \beta \cdot \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \cdot \frac{\frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2} \cdot \kappa_2(\gamma)}{1 - \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2} \cdot \kappa_2(\gamma) \cdot \beta}. \end{aligned}$$

By Lemma 4.6, we have that $\frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \leq 2$. Thus, assuming that $\beta < \frac{1}{2}$, we get that

$$\left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|u^*\|^2 \cdot \kappa_2(\gamma)}{\kappa_3(\gamma)^{2/3}} \right| = \left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \right| \leq O(\beta). \quad (5.15)$$

Using the triangle inequality and $\frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} \leq 2$ again, we furthermore get that

$$\left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} \right| \leq \frac{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)}{\|\Sigma^{1/2} w\|^2} + O(\beta) \leq O(1).$$

Finally, we want to bound

$$\left| \frac{1}{\kappa_2(\gamma)} - \frac{1}{\kappa_2(\hat{\gamma})} \right| = \frac{1}{\kappa_2(\gamma) \cdot \kappa_2(\hat{\gamma})} \cdot |\kappa_2(\gamma) - \kappa_2(\hat{\gamma})|.$$

As before in the proof of Lemma 5.11, it follows from Claim 3.19 and assumption (5.13) that

$$|\kappa_2(\gamma) - \kappa_2(\hat{\gamma})| \lesssim |\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq \frac{\beta}{\kappa_2(\gamma)^{3/2}}$$

As before, by our assumption $\beta < c \cdot \kappa_2(\gamma)^{5/2}$, we have that $\kappa_2(\hat{\gamma}) \geq \kappa_2(\gamma) - \frac{1}{2}\kappa_2(\gamma) = \frac{1}{2}\kappa_2(\gamma)$ and thus

$$\frac{1}{\kappa_2(\gamma) \cdot \kappa_2(\hat{\gamma})} \cdot |\kappa_2(\gamma) - \kappa_2(\hat{\gamma})| \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right).$$

Combining the above, we get that

$$\|A_2\|_F \leq 2 \cdot O(1) \cdot O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right) = O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right).$$

Bounding $\|A_3\|_F$. For A_3 , note that by (5.14) and (5.15) and using that $\kappa_2(\gamma) \leq 1$ (cf. Claim 3.7), we immediately get that

$$\begin{aligned} \|A_3\|_F &\leq \left| \frac{1}{\kappa_2(\gamma)} - 1 \right| \cdot \left| (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1} - \frac{\|u^*\|^2 \cdot \kappa_2(\gamma)}{\kappa_3(\gamma)^{2/3}} \right| \cdot \|\Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2}\| \\ &\leq \frac{1}{\kappa_2(\gamma)} \cdot O(\beta) \cdot 2 = O\left(\frac{\beta}{\kappa_2(\gamma)}\right). \end{aligned}$$

Bounding $\|A_4\|_F$. For A_4 , recalling that $\frac{\|u^*\|^2}{\kappa_3(\gamma)^{2/3}} = \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2}$ and $\kappa_2(\gamma) > 0$, we get that

$$\begin{aligned}
\|A_4\|_F &\leq |1 - \kappa_2(\gamma)| \cdot \left\| \frac{\|u^*\|^2}{\kappa_3(\gamma)^{2/3}} \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} - \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma^{1/2} w\|^2} \right\|_F \\
&\leq \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2} \cdot \left\| \Sigma^{-1/2} \hat{u} \hat{u}^\top \Sigma^{-1/2} - \frac{\Sigma^{1/2} w w^\top \Sigma^{1/2}}{\|\Sigma w\|^2} \right\|_F \\
&= \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2} \cdot \left\| \Sigma^{-1/2} \left(\hat{u} \hat{u}^\top - \frac{\Sigma w w^\top \Sigma}{\|\Sigma w\|^2} \right) \Sigma^{-1/2} \right\|_F \\
&\leq \|\Sigma^{-1/2}\|^2 \cdot \frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2} \cdot \left\| \hat{u} \hat{u}^\top - \frac{\Sigma w}{\|\Sigma w\|} \left(\frac{\Sigma w}{\|\Sigma w\|} \right)^\top \right\|_F
\end{aligned}$$

We now want to use assumption (5.11). Since $u^* \propto -\Sigma w$, this assumption implies that

$$\min \left\{ \left\| \hat{u} - \frac{\Sigma w}{\|\Sigma w\|} \right\|, \left\| \hat{u} + \frac{\Sigma w}{\|\Sigma w\|} \right\| \right\} \leq \beta$$

Using Fact 5.12 on the unit vectors \hat{u} and $\frac{\Sigma w}{\|\Sigma w\|}$ to get that

$$\left\| \hat{u} \hat{u}^\top - \frac{\Sigma w}{\|\Sigma w\|} \left(\frac{\Sigma w}{\|\Sigma w\|} \right)^\top \right\|_F \leq 2\beta.$$

Furthermore, using Lemma 4.6, we get that $\|\Sigma^{-1/2}\|^2 \leq 2$ and $\frac{\|\Sigma w\|^2}{\|\Sigma^{1/2} w\|^2} \leq \frac{1}{\kappa_2(\gamma)}$. Combining these with the above, we get that

$$\|A_4\|_F \leq 2 \cdot \frac{1}{\kappa_2(\gamma)} \cdot 2\beta = O\left(\frac{\beta}{\kappa_2(\gamma)}\right).$$

Completing the proof. Combining the bounds on A_1, A_2, A_3 and A_4 , we get that

$$\begin{aligned}
\left\| \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d \right\|_F &\leq \|A_1\|_F + \|A_2\|_F + \|A_3\|_F + \|A_4\|_F \\
&\leq \beta + O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right) + O\left(\frac{\beta}{\kappa_2(\gamma)}\right) + O\left(\frac{\beta}{\kappa_2(\gamma)}\right) \\
&\leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right),
\end{aligned}$$

where the last inequality follows because $\kappa_2(\gamma) \leq 1$. This completes the proof. \square

5.4 Proof of Theorem 5.1

In this section, we prove our main result, Theorem 5.1. We start with combining the results from Sections 5.1 to 5.3 to give a sample complexity bound of Algorithm 2.

Lemma 5.14. Consider a halfspace truncated Gaussian $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, where $\Sigma > 0$. Define $\gamma = \frac{\tau - \langle w, \mu \rangle}{\|\Sigma^{1/2} w\|}$.

Then, given

$$n = O\left(\frac{d^2}{\kappa_2(\gamma)^8 \kappa_3(\gamma)^2} + \frac{d^2}{\kappa_2(\gamma)^3 \kappa_3(\gamma)^4} + \frac{d^2}{\kappa_2(\gamma)^{10} \kappa_3(\gamma)^2 \varepsilon^2}\right)$$

samples, [Algorithm 2](#) computes estimators $\hat{\mu}, \widehat{\Sigma}$ such that with probability at least 0.99

$$d_{\text{TV}}\left(\mathcal{N}(\hat{\mu}, \widehat{\Sigma}), \mathcal{N}(\mu, \Sigma)\right) \leq \varepsilon.$$

Later, we will then bound this sample complexity in terms of α , as in [Theorem 5.1](#). Finally, we will also argue about the runtime of [Algorithm 2](#).

Proof of [Lemma 5.14](#). For any $\beta > 0$, given $n = O\left(\frac{d^2}{\kappa_2(\gamma)^3 \kappa_3(\gamma)^2 \beta^2}\right)$ samples, by [Lemma 5.8](#), we have that, with high probability,

$$\sup_{\|v\|=1} \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \leq O\left(\frac{\beta \cdot |\kappa_3(\gamma)|}{\sqrt{d}}\right) \leq O(\beta). \quad (5.16)$$

Now, we get, by [Lemma 5.5](#), with high probability,

$$\min \left\{ \left\| \hat{\mu} - \frac{u^*}{\|u^*\|} \right\|, \left\| \hat{\mu} + \frac{u^*}{\|u^*\|} \right\| \right\} \leq O(\beta). \quad (5.17)$$

Next, consider the sample covariance \widehat{M}_2 computed from $O(d^2/(\kappa_2(\gamma)^2 \beta^2))$ i.i.d. samples. From [Claim 4.3](#) we know each sample is $1/\sqrt{\kappa_2(\gamma)}$ -sub-Gaussian, hence as in [Fact 3.26](#), we have, with high probability

$$\|M_2^{-1/2} \widehat{M}_2 M_2^{-1/2} - I_d\| \leq \beta \quad \text{and} \quad \|\Sigma^{-1/2} (\widehat{M}_2 - M_2) \Sigma^{-1/2}\|_F \leq \beta, \quad (5.18)$$

where we also used that $\|\Sigma^{-1/2} M_2^{1/2}\| \leq \frac{1}{2}$ by [Lemma 4.6](#). Using equations (5.17), (5.18) and (5.16), we can apply [Lemma 5.9](#) as long as $\beta < c_1$ (for some absolute constant c_1). This shows that the estimator

$$\hat{\gamma} = \text{skew}^{-1} \left(- \left(\hat{\mu}^\top \widehat{M}_2^{-1} \hat{\mu} \right)^{3/2} \cdot \langle u^{\otimes 3}, \widehat{M}_3 \rangle \right)$$

satisfies

$$|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{3/2}}\right). \quad (5.19)$$

From [Lemma 5.9](#), we furthermore get the inequality

$$\left| \hat{\mu}^\top \widehat{M}_2^{-1} \hat{\mu} - \frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)} \right| \leq O(\beta) \quad (5.20)$$

As long as $\beta < c_2 \cdot \kappa_2(\gamma)^{5/2}$, using (5.18), (5.17), (5.20) and (5.19), we can apply [Lemma 5.13](#) to get that the covariance estimator

$$\widehat{\Sigma} = \widehat{M}_2 + \left(\frac{1}{\kappa_2(\hat{\gamma})} - 1 \right) (\hat{\mu}^\top \widehat{M}_2^{-1} \hat{\mu})^{-1} \cdot \hat{\mu} \hat{\mu}^\top$$

satisfies

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\|_F \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right). \quad (5.21)$$

For mean estimation, we have that x_i is $\sqrt{2/\kappa_2(\gamma)}$ -sub-Gaussian since $M_2^{-1/2} x_i$ is $1/\sqrt{\kappa_2(\gamma)}$ -sub-Gaussian by [Claim 4.3](#) and $\|M_2^{1/2}\| \leq \frac{1}{2}$ by [Lemma 4.6](#). Thus, \bar{x} is $(\sqrt{2/\kappa_2(\gamma)})/\sqrt{n}$ -sub-Gaussian (cf.

[Ver18, Proposition 2.7.1]). Given $n = O(d/(\kappa_2(\gamma)\beta^2))$ samples, we therefore have (see for example [Rig15, Theorem 1.19]), with high probability,

$$\left\| \bar{x} - \mu - \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} \right\| \leq \beta. \quad (5.22)$$

To apply Lemma 5.11, it remains to argue that choosing the sign of \hat{u} by computing the sign of $\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle$ indeed ensures (5.7). From the proof of Lemma 5.9, we know that

$$\|\hat{u} - \tilde{u}\| \leq O(\beta) \implies \left| \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle - \|u^*\|^3 \right| \leq O(\beta) \quad (5.23)$$

$$\|\hat{u} + \tilde{u}\| \leq O(\beta) \implies \left| \langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle + \|u^*\|^3 \right| \leq O(\beta). \quad (5.24)$$

If the term $O(\beta)$ on the right hand side is smaller than $\|u^*\|^3$, then only one of the two conclusions (5.23) or (5.24) can be true. By (5.17), one of the two must hold. Thus, exactly one of the two conclusions (5.23) or (5.24) is true. Furthermore, if the sign of $\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle$ is positive, then it must be (5.23) (and hence we want to pick \hat{u} for (5.7)). If this sign is negative, it must be (5.24) (and we want to pick $-\hat{u}$). Hence, if $\beta < c_3 \cdot |\kappa_3(\gamma)|$ (for a sufficiently small constant c_3 ensuring that $O(\beta) \leq \|u^*\|^3$ by Claim 5.10), then flipping the sign of \hat{u} if $\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \rangle < 0$ indeed ensures that (after the potential sign flip)

$$\left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\| \leq O(\beta). \quad (5.25)$$

Thus, as long as $\beta < c_4 \cdot \kappa_2(\gamma)^{5/2}$, using (5.22), (5.18), (5.25), (5.20) and (5.19), we can apply Lemma 5.11 to get that the mean estimator

$$\hat{\mu} = \bar{x} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{-1/2} \hat{u}$$

satisfies

$$\|\hat{\mu} - \mu\| \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right).$$

Furthermore, by Lemma 4.6 we have that $\|\Sigma^{-1/2}\| \leq O(1)$ and thus the above also implies

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right). \quad (5.26)$$

Thus, summarized, for $\beta < c \cdot \min\{\kappa_2(\gamma)^{5/2}, |\kappa_3(\gamma)|\}$ for a sufficiently small constant c , we have that, with high probability,¹⁰

$$d_{\text{TV}}\left(\mathcal{N}(\hat{\mu}, \widehat{\Sigma}), \mathcal{N}(\mu, \Sigma)\right) \leq O\left(\frac{\beta}{\kappa_2(\gamma)^{7/2}}\right).$$

The sample complexity is

$$n = O\left(\max\left\{\frac{d^2}{\kappa_2(\gamma)^3 \kappa_3(\gamma)^2 \beta^2}, \frac{d^2}{\kappa_2(\gamma)^2 \beta^2}, \frac{d}{\kappa_2(\gamma) \beta^2}\right\}\right) = O\left(\frac{d^2}{\kappa_2(\gamma)^3 \kappa_3(\gamma)^2 \beta^2}\right).$$

Thus, setting $\beta = O(\min\{\kappa_2(\gamma)^{5/2}, |\kappa_3(\gamma)|, \kappa_2(\gamma)^{7/2} \cdot \varepsilon\})$ for a sufficiently small constant, we get the lemma. \square

¹⁰For Gaussian, closeness in parameters implies closeness in TV-distance, see e.g. [AAL23, Theorem 1.8].

The above proof is sufficient as long as α is bounded away from 1 by a constant. As α approaches 1, we have that $\kappa_3(\gamma) \rightarrow 0$ and thus the sample complexity increases. To deal with this, we now give an alternative sample complexity bound that relies on the fact that if α is close to 1, then the sample mean and covariance are good estimators and the correction terms in [Algorithm 2](#) are small.

Lemma 5.15. *Consider a halfspace truncated Gaussian $\mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$, where $\Sigma \succ 0$. Define $\gamma = \frac{\tau - \langle w, \mu \rangle}{\|\Sigma^{1/2}w\|}$. Assume that $\alpha = \Phi(\gamma) \geq 1 - c$ for a sufficiently small constant $c > 0$. Then, given $n = O(d^2/\varepsilon^2)$ samples, [Algorithm 2](#) computes estimators $\hat{\mu}, \hat{\Sigma}$ such that with probability at least 0.99*

$$d_{\text{TV}}\left(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)\right) \leq O\left(\varepsilon + (1 - \alpha) \log\left(\frac{1}{1 - \alpha}\right)\right).$$

Proof. The strategy of the proof is as follows: We want to argue that we are essentially outputting the sample mean and covariance and that these are good estimates. First, note that given $n = O(d^2/\beta^2)$ samples, the sample moments satisfy, with high probability

$$\begin{aligned} \left\| \mu + \frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|} - \bar{x} \right\| &\leq \frac{\beta}{\sqrt{d}} && \text{(analogous to (5.22))} \\ \left\| M_2 - \widehat{M}_2 \right\|_F &\leq \beta && \text{(analogous to (5.18))} \\ \sup_{\|v\|=1} \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle &\leq \frac{\beta}{\sqrt{d}} && \text{(Lemma 5.8)} \end{aligned}$$

Note that we used that for $\alpha \geq 1 - c$, we have that $\kappa_2(\gamma) \geq \kappa_2(\Phi^{-1}(1 - c))$ can be lower bounded by a constant by [Claim 3.8](#).

Covariance estimation. For covariance estimation, we want to argue that M_2 is close to Σ , i.e. that the second central moment would be good estimate. Then, we will also argue that for α close to 1, we are essentially outputting \widehat{M}_2 .

First, note that, by the formula for M_2 of [Lemma 4.2](#), we have

$$\|M_2 - \Sigma\|_F \leq (1 - \kappa_2(\gamma)) \cdot \frac{\|\Sigma w w^\top \Sigma\|_F}{\|\Sigma^{1/2}w\|^2} \leq O\left((1 - \alpha) \log\left(\frac{1}{1 - \alpha}\right)\right),$$

where we bounded $1 - \kappa_2(\gamma)$ by [Claim 3.14](#) and used $\|\Sigma\| \leq 2/\kappa_2(\Phi^{-1}(1 - c)) \leq O(1)$ for $\alpha \geq 1 - c$ by [Claim 3.8](#) and [Lemma 4.6](#). Combined with $\|M_2 - \widehat{M}_2\|_F \leq \beta$ from above, we furthermore get that

$$\left\| \widehat{M}_2 - \Sigma \right\|_F \leq \beta + O\left((1 - \alpha) \log\left(\frac{1}{1 - \alpha}\right)\right).$$

Recall that the estimator of [Algorithm 2](#) is

$$\widehat{\Sigma} = \widehat{M}_2 + \left(\frac{1}{\kappa_2(\hat{\gamma})} - 1\right) \cdot \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u}\right)^{-1} \hat{u} \hat{u}^\top.$$

It remains to argue that the correction term $\left(\frac{1}{\kappa_2(\hat{\gamma})} - 1\right) \cdot \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u}\right)^{-1} \hat{u} \hat{u}^\top$ is small. We have

$$c_1 \leq \lambda_{\min}\left(\widehat{M}_2^{-1}\right) \leq \hat{u}^\top \widehat{M}_2^{-1} \hat{u} \leq \lambda_{\max}\left(\widehat{M}_2^{-1}\right) \leq c_2. \quad (5.27)$$

for appropriate constants $c_1, c_2 > 0$, by [Lemma 4.6](#), as long as β is at most a sufficiently small constant (such that we have $2c_1 \cdot M_2^{-1} \leq \widehat{M}_2^{-1} \leq \frac{c_2}{2} \cdot M_2^{-1}$).

Thus, it remains to lower bound $\kappa_2(\hat{\gamma})$ in terms of $1 - \alpha$. For this, we want to show that $\hat{\gamma}$ has to be large and use [Claim 3.14](#). Since skew is monotonic by [Lemma 3.17](#), it suffices to show that $\text{skew}(\hat{\gamma})$ is large or equivalently that $|\text{skew}(\hat{\gamma})|$ is small (since skew is always negative by [Claims 3.7](#) and [3.8](#)). We have that, using [\(5.27\)](#),

$$|\text{skew}(\hat{\gamma})| = \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u} \right)^{3/2} \cdot \left| \left\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \right\rangle \right| \lesssim \left| \left\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \right\rangle \right|.$$

Furthermore, we can bound the right hand side of the above as follows

$$\left| \left\langle \hat{u}^{\otimes 3}, \widehat{M}_3 \right\rangle \right| \leq \left| \left\langle \hat{u}^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \right| + \left| \left\langle \hat{u}^{\otimes 3}, M_3 \right\rangle \right| \leq \frac{\beta}{\sqrt{d}} + \|u^*\|^3 \leq \frac{\beta}{\sqrt{d}} + O(|\text{skew}(\gamma)|),$$

where we used the guarantee on $\widehat{M}_3 - M_3$, $M_3 = u^{*\otimes 3}$, as well as [Claim 5.10](#) to bound $\|u^*\|$. Combined with the above, this gives

$$|\text{skew}(\hat{\gamma})| \leq O \left(\max \left\{ \frac{\beta}{\sqrt{d}}, |\text{skew}(\gamma)| \right\} \right).$$

By choosing β to be smaller than a sufficiently small constant and c sufficiently small, we can make the right hand side arbitrarily small and get that $|\text{skew}(\hat{\gamma})| \leq |\text{skew}(2)|$, which implies $\hat{\gamma} \geq 2$ (since skew is always negative and increasing, $|\text{skew}|$ is decreasing). Using [Claim 3.16](#) for both γ and $\hat{\gamma}$, we get that

$$(1 - \hat{\alpha}) \cdot \log^{3/2} \left(\frac{1}{1 - \hat{\alpha}} \right) \leq O \left(\max \left\{ \frac{\beta}{\sqrt{d}}, (1 - \alpha) \cdot \log^{3/2} \left(\frac{1}{1 - \alpha} \right) \right\} \right) \quad (5.28)$$

for $\hat{\alpha} := \Phi(\hat{\gamma})$. Since both $\hat{\alpha}$ and α are close to 1, if the maximum is the second term, then we also have¹¹

$$(1 - \hat{\alpha}) \cdot \log \left(\frac{1}{1 - \hat{\alpha}} \right) \lesssim (1 - \alpha) \cdot \log \left(\frac{1}{1 - \alpha} \right).$$

If the maximum is the first term, using that $\log^{1/2} \left(\frac{1}{1 - \hat{\alpha}} \right) \geq 1$, we get that

$$(1 - \hat{\alpha}) \cdot \log \left(\frac{1}{1 - \hat{\alpha}} \right) \lesssim \frac{\beta}{\sqrt{d}}.$$

Combining these two, we thus get

$$(1 - \hat{\alpha}) \cdot \log \left(\frac{1}{1 - \hat{\alpha}} \right) \leq O \left(\max \left\{ \frac{\beta}{\sqrt{d}}, (1 - \alpha) \cdot \log \left(\frac{1}{1 - \alpha} \right) \right\} \right).$$

¹¹To formally prove this, define $a = \log \left(\frac{1}{1 - \alpha} \right)$, $\hat{a} = \log \left(\frac{1}{1 - \hat{\alpha}} \right)$ and $r = \frac{a}{\hat{a}}$. We have $\frac{\exp(-\hat{a})\hat{a}^{3/2}}{\exp(-a)a^{3/2}} \leq O(1)$ and we want to show $\frac{\exp(-\hat{a})\hat{a}}{\exp(-a)a} \leq O(1)$, for which it is sufficient to show $\frac{\hat{a}^{1/2}}{a^{1/2}} \leq O(1)$. If $r \leq 1$, then this is clear. Otherwise, since $\hat{\gamma} \geq 2$, we have $\hat{a} \geq \log \left(\frac{1}{1 - \Phi(2)} \right) \geq 1$. Thus, by assumption, we have

$$\exp(r - 1) \cdot r^{-3/2} \leq \exp((r - 1)\hat{a}) \cdot r^{-3/2} = \frac{\exp(-\hat{a})\hat{a}^{3/2}}{\exp(-a)a^{3/2}} \leq O(1).$$

Now, we have $\lim_{r \rightarrow \infty} \exp((r - 1)) \cdot r^{-3/2} = \infty$ and thus $\exp((r - 1)) \cdot r^{-3/2} \leq O(1)$ can only be true if we also have $r \leq O(1)$.

Thus, we have that, using [Claim 3.14](#) in the second inequality,

$$\begin{aligned} \frac{1}{\kappa_2(\hat{\gamma})} - 1 &\leq \frac{1}{\kappa_2(2)} \cdot (1 - \kappa_2(\hat{\gamma})) \\ &\leq O\left((1 - \hat{\alpha}) \cdot \log\left(\frac{1}{1 - \hat{\alpha}}\right)\right) \\ &\leq O\left(\max\left\{\frac{\beta}{\sqrt{d}}, (1 - \alpha) \cdot \log\left(\frac{1}{1 - \alpha}\right)\right\}\right). \end{aligned}$$

This implies, noting that $\|\hat{u}\hat{u}^\top\|_F = \|\hat{u}\|^2 = 1$,

$$\left\|\widehat{\Sigma} - \widehat{M}_2\right\|_F = \left\|\left(\frac{1}{\kappa_2(\hat{\gamma})} - 1\right) \cdot \left(\hat{u}^\top \widehat{M}_2^{-1} \hat{u}\right)^{-1} \hat{u}\hat{u}^\top\right\|_F \leq O\left(\frac{\beta}{\sqrt{d}} + (1 - \alpha) \cdot \log\left(\frac{1}{1 - \alpha}\right)\right).$$

Combined with the bound on $\|\widehat{M}_2 - \Sigma\|_F$ we get that the estimator from [Algorithm 2](#) satisfies

$$\left\|\widehat{\Sigma} - \Sigma\right\|_F \leq O\left(\beta + (1 - \alpha) \cdot \log\left(\frac{1}{1 - \alpha}\right)\right)$$

Finally, using [Lemma 4.6](#), we get that

$$\left\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I_d\right\|_F \leq \|\Sigma^{-1/2}\|^2 \cdot \left\|\widehat{\Sigma} - \Sigma\right\|_F \leq O\left(\beta + (1 - \alpha) \cdot \log\left(\frac{1}{1 - \alpha}\right)\right).$$

Mean estimation. For mean estimation, we want to similarly argue that the sample mean would be a good estimator and that we are outputting something close to the sample mean. First, recall that we have that the sample mean is close to the true mean, i.e.,

$$\left\|\mu + \frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|} - \bar{x}\right\| \leq \beta.$$

By picking c sufficiently small, we can achieve $\gamma \geq 2$. Then, we furthermore get that

$$\left\|\frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|}\right\| \leq \frac{1}{\Phi(2)} \cdot \|\Sigma^{1/2}\|^2 \cdot \varphi(\gamma) \leq O(\varphi(\gamma)),$$

where we again used [Lemma 4.6](#) and the fact that $\kappa_2(\gamma) \geq \kappa_2(2)$ to bound $\|\Sigma^{1/2}\|$. Using [Fact 3.3](#) and [Claim 3.15](#), we can bound

$$\varphi(\gamma) \leq \left(\gamma + \frac{1}{\gamma}\right) \cdot (1 - \Phi(\gamma)) \leq 2\gamma \cdot (1 - \Phi(\gamma)) \leq 2\sqrt{2} \cdot (1 - \Phi(\gamma)) \cdot \log^{1/2}\left(\frac{1}{1 - \Phi(\gamma)}\right).$$

Thus, we get that

$$\left\|\frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \frac{\Sigma w}{\|\Sigma^{1/2}w\|}\right\| \leq (1 - \alpha) \cdot \log^{1/2}\left(\frac{1}{1 - \Phi(\gamma)}\right)$$

and hence

$$\|\mu - \bar{x}\| \leq \beta + (1 - \alpha) \cdot \log^{1/2}\left(\frac{1}{1 - \Phi(\gamma)}\right).$$

Recall that the mean estimator of [Algorithm 2](#) is

$$\hat{\mu} = \bar{x} + \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1/2} \hat{u}.$$

It remains to argue that the correction term $\frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1/2} \hat{u}$ is small. We can bound this, using [\(5.27\)](#), $\hat{\gamma} \geq 2$ and [Claim 3.8](#), as follows

$$\left\| \frac{\kappa_1(\hat{\gamma})}{\sqrt{\kappa_2(\hat{\gamma})}} (\hat{u}^\top \widehat{M}_2^{-1} \hat{u})^{-1/2} \hat{u} \right\| \lesssim \frac{1}{\sqrt{\kappa_2(2)}} \cdot |\kappa_1(\hat{\gamma})| \lesssim |\kappa_1(\hat{\gamma})| = \frac{\varphi(\hat{\gamma})}{\Phi(\hat{\gamma})} \leq \frac{\varphi(\hat{\gamma})}{\Phi(2)} \lesssim \varphi(\hat{\gamma}).$$

As above for γ , using [Fact 3.3](#) and [Claim 3.15](#), we can bound

$$\varphi(\hat{\gamma}) \leq 2\sqrt{2} \cdot (1 - \Phi(\hat{\gamma})) \cdot \log^{1/2} \left(\frac{1}{1 - \Phi(\hat{\gamma})} \right) = 2\sqrt{2} \cdot (1 - \hat{\alpha}) \cdot \log^{1/2} \left(\frac{1}{1 - \hat{\alpha}} \right).$$

Using the same argument as in [Footnote 11](#), [\(5.28\)](#) also implies

$$(1 - \hat{\alpha}) \cdot \log^{1/2} \left(\frac{1}{1 - \hat{\alpha}} \right) \leq O \left(\max \left\{ \frac{\beta}{\sqrt{d}}, (1 - \alpha) \cdot \log^{1/2} \left(\frac{1}{1 - \alpha} \right) \right\} \right).$$

Thus, it follows that

$$\|\hat{\mu} - \bar{x}\| \leq O \left(\frac{\beta}{\sqrt{d}} + (1 - \alpha) \cdot \log^{1/2} \left(\frac{1}{1 - \alpha} \right) \right).$$

Combining this with the bound on $\|\mu - \bar{x}\|$ and using [Lemma 4.6](#) again to bound $\|\Sigma^{-1/2}\|$, we get that

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq O \left(\beta + (1 - \alpha) \cdot \log^{1/2} \left(\frac{1}{1 - \alpha} \right) \right).$$

Conclusion. So far, we have shown that for β smaller than a sufficiently small constant c_3 , we get that, using $O(d^2/\beta^2)$ samples, the estimators in [Algorithm 2](#) satisfy

$$d_{\text{TV}} \left(\mathcal{N}(\hat{\mu}, \widehat{\Sigma}), \mathcal{N}(\mu, \Sigma) \right) \leq O \left(\beta + (1 - \alpha) \log \left(\frac{1}{1 - \alpha} \right) \right).$$

Note that without loss of generality we can assume that $\varepsilon \leq 1$, otherwise any output is good (the TV-distance is at most 1 regardless of the output). Finally, picking $\beta = \min\{c_3, \varepsilon\}$ proves the lemma. The claimed sample complexity follows since this β satisfies $\beta \geq \Omega(\varepsilon)$ (since $\varepsilon \leq 1$). \square

Using [Lemmas 5.14](#) and [5.15](#), we can now prove [Theorem 5.1](#).

Proof of [Theorem 5.1](#). Correctness follows from [Lemmas 5.14](#) and [5.15](#) if the number of samples used by the algorithm is large enough. To determine the sample complexity, we distinguish the three cases (i) $\alpha \leq \alpha_1$, (ii) $\alpha \in (\alpha_1, \alpha_2)$ and (iii) $\alpha \geq \alpha_2$, where $\alpha_1 = 1 - c$ and $\alpha_2 = \max \left\{ 1 - c, 1 - \frac{\varepsilon}{\log(1/\varepsilon)} \right\}$ for c as in [Lemma 5.15](#). Again, we note that without loss of generality we can assume $\varepsilon \leq 1$, otherwise the guarantees hold trivially.

Case (i): Sample complexity for $\alpha \leq \alpha_1$. In this case, we want to use [Lemma 5.14](#). By [Claims 3.22](#) and [3.23](#) we get that

$$\kappa_2(\gamma) \geq \Omega(\log^{-1}(1/\alpha)) \quad \text{and} \quad |\kappa_3(\gamma)| \geq \Omega(\log^{-3/2}(1/\alpha)).$$

Thus, the sample complexity in this case is

$$n = O\left(\frac{d^2}{\varepsilon^2} \cdot \log^{14}(1/\alpha)\right).$$

Case (ii): Sample complexity for $\alpha \in (\alpha_1, \alpha_2)$. In this case, we again use [Lemma 5.14](#). We can bound $\kappa_2(\gamma) \geq \Omega(1)$ by [Claim 3.22](#). For κ_3 , we have that $1 - \alpha \geq 1 - \alpha_2 \geq c' \frac{\varepsilon}{\log(1/\varepsilon)}$. Thus, using [Claim 3.23](#), we have that

$$|\kappa_3(\gamma)| \geq \Omega\left((1 - \alpha) \cdot \log^{3/2}\left(\frac{1}{1 - \alpha}\right)\right) \geq \Omega(\varepsilon).$$

Thus, the sample complexity in this case (again from [Lemma 5.14](#)) is

$$n = O\left(\frac{d^2}{\varepsilon^4}\right).$$

Case (iii): Sample complexity for $\alpha \geq \alpha_2$. In this case, we claim that we can use [Lemma 5.15](#). The guarantee from this lemma is that the TV-distance is at most

$$O\left(\varepsilon + (1 - \alpha) \log\left(\frac{1}{1 - \alpha}\right)\right)$$

and we need to show that this is at most ε . Using that $x \mapsto (1 - x) \cdot \log(1/(1 - x))$ is decreasing for x close to 1, we get that the TV-distance is at most

$$O\left(\varepsilon + \frac{\varepsilon}{\log(1/\varepsilon)} \log\left(\frac{\log(1/\varepsilon)}{\varepsilon}\right)\right) \leq O(\varepsilon).$$

The sample complexity by [Lemma 5.15](#) is $n = O(d^2/\varepsilon^2)$.

These three cases together prove correctness of [Algorithm 2](#) given the sample complexity claimed in [Theorem 5.1](#).¹²

Runtime. Finally, it remains to prove the runtime of the algorithm. We do so analogous to the arguments presented in [Section 2](#). The runtime to compute the sample mean is $O(n \cdot d)$. The runtime to compute the sample covariance is $O(T(n, d))$. If we first do the contraction and then average over the samples, then also computing the random contraction R of the third sample moment can be computed in time $T(n, d)$. Computing a spectral decomposition of this matrix can then be done in time $O(d^3)$, e.g. by computing a singular value decomposition. Computing the sign of $\langle u^{\otimes 3}, \widehat{M}_3 \rangle$ can be done in time $O(n \cdot d)$ since

$$\langle u^{\otimes 3}, \widehat{M}_3 \rangle = \sum_{i=1}^n \langle u, x_i - \bar{x} \rangle^3.$$

Finally, computing the estimates can be done in time $O(d^2)$. Since $n \geq \Omega(d^2)$, we have that $T(n, d) \geq \Omega(n \cdot d) \geq \Omega(d^3)$ and thus the dominating term in the runtime is $O(T(n, d))$, which completes the proof. \square

¹²The above argument shows that the TV-distance is $O(\varepsilon)$. We can make the TV-distance ε instead of $O(\varepsilon)$ by doing the exact same argument for $\varepsilon' = \frac{\varepsilon}{O(1)}$.

Also note that we can simplify the sample complexity to be $O((d^2/\varepsilon^2) \cdot \text{polylog}(1/\alpha))$ for $\alpha \leq 1 - c$ (or in fact for α being at most any constant smaller than 1) and $O(n^2/\varepsilon^4)$ for the remaining two cases. Thus, the proof above does indeed show correctness of the algorithm for the sample complexity as in [Theorem 5.1](#).

Remark 5.16. In [Lemmas 5.14](#) and [5.15](#) above we have been assuming that the $\text{skew}^{-1}(\cdot)$ function can be evaluated exactly and efficiently. In practice, we would instead only find an approximate inverse using binary search. We need to ensure that we get an estimate $\hat{\gamma}$ such that

$$|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq \varepsilon \cdot \text{poly}(\min\{O(1), 1/\log(1/\alpha)\}).$$

We can restrict the search to $\hat{\gamma} \geq \Omega(\sqrt{\log 1/\alpha})$ (the lower bound on γ). Furthermore, if $\gamma \geq \tilde{O}(\sqrt{\log(1/\varepsilon)})$ (this implies $\alpha \geq 1 - \frac{\varepsilon}{\log 1/\varepsilon}$), then we showed in [Lemma 5.15](#) and the proof of [Theorem 5.1](#) that $\hat{\gamma} = \infty$ is a good enough estimate and thus we can restrict the binary search to the interval $[\Omega(\sqrt{\log 1/\alpha}), \tilde{O}(\sqrt{\log(1/\varepsilon)})]$. The derivative of skew is upper bounded by a constant and thus it is sufficient to get an estimate of the inverse up to

$$\varepsilon \cdot \text{poly}(\min\{O(1), 1/\log(1/\alpha)\}).$$

Hence, we need to do binary search on the interval $[\Omega(\sqrt{\log 1/\alpha}), \tilde{O}(\sqrt{\log(1/\varepsilon)})]$ up to the above accuracy, which can be done in time

$$\log(1/\varepsilon) \cdot \text{poly}(\log \log(1/\varepsilon), \log \log(1/\alpha)).$$

We remark that this time is negligible with respect to the other computation.

6 Robustness

In this section, we show how to extend our result for halfspaces to the robust setting. We will present the results for the non-trivial truncation regime of α bounded away from 1. Throughout this section as well we will assume $\Sigma > 0$.

6.1 Preliminaries on robustness

We consider the following definition of robustness analogous to the well-studied strong contamination model in algorithmic statistics literature. We denote a distribution \mathcal{D} truncated to the set S by $\mathcal{D}|_S$.

Definition 6.1 (η -corruption of $\mathcal{D}|_S$). Samples X_1, \dots, X_n are called an η -corruption of $\mathcal{D}|_S$ if they are obtained as follows:

1. First, we draw $Y_1, \dots, Y_n \sim \mathcal{D}|_S$.
2. Then, an adversary picks $X_1, \dots, X_n \in \mathbb{R}^d$ such that $|\{i \in [n] : X_i \neq Y_i\}| \leq \eta n$.

Observe that the adversary can remove all information about the truncation set.

In particular the adversary is allowed to also pick X_i such that $X_i \in \mathbb{R}^d \setminus S$.

We require the following definition of certifiably sub-Gaussian distributions, see e.g. [\[KSS18, Definition 1.1\]](#) or [\[DHPT25, Definition 1.3\]](#).

Definition 6.2 (Certifiably sub-Gaussian distribution). A distribution \mathcal{D} with mean μ is called s -certifiably sub-Gaussian if for all even $m \in \mathbb{N}$, there is a sum-of-squares proof of the following inequality in indeterminate v

$$\mathbb{E}_{X \sim \mathcal{D}} [\langle X - \mu, v \rangle^m] \leq (Cs\sqrt{m})^m \|v\|^m,$$

where C is a universal constant.

We also require the following two theorems that we will utilize to obtain robust estimates of the moments.

Theorem 6.3 ([KSS18, Theorems 2.2 and 2.3]). *Let $k \geq 6$ be an even constant. Let $C > 0$. There is a polynomial-time algorithm with the following guarantees: Let \mathcal{D} be a C -certifiably sub-Gaussian distribution on \mathbb{R}^d with mean μ and covariance Σ . Let $M_{3,\text{nc}}$ be the raw third moment of \mathcal{D} . Let $\eta > 0$ such that $Ck \cdot \eta^{1-2/k}$ is at most a sufficiently small constant. Then given as input an η -corrupted sample of \mathcal{D} of size $n \geq (C + d)^{O(k)}$, the algorithm outputs estimates of the moments of \mathcal{D} with the following guarantees:*

$$\begin{aligned} \|\mu - \hat{\mu}\| &\leq O(Ck)^{1/2} \cdot \eta^{1-1/k} \cdot \|\Sigma\|^{1/2} \\ (1 - O(Ck)^{1/2} \cdot \eta^{1-2/k})\Sigma &\leq \widehat{\Sigma} \leq (1 + O(Ck)^{1/2} \cdot \eta^{1-2/k})\Sigma \\ \forall v \in \mathbb{R}^d : \left\langle M_{3,\text{nc}} - \widehat{M}_{3,\text{nc}}, v^{\otimes 3} \right\rangle^2 &\leq O(Ck)^{3/2} \cdot \eta^{1-3/k} \left\langle M_{2,\text{nc}}, v^{\otimes 2} \right\rangle^3 \end{aligned}$$

where $M_{2,\text{nc}}$ corresponds to the raw second moment of \mathcal{D} .

Theorem 6.4 ([DHPT25, Theorem 1.6]). *There exists a universal constant $C' > 0$ such that any s -sub-Gaussian distribution is C' - s -certifiably sub-Gaussian.*

Fact 6.5. *The distribution $\mathcal{D} := \mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$ is $O(\|\Sigma\|^{1/2})$ -certifiably sub-Gaussian.*

We note that this follows directly by combining [Corollary 4.4](#) and [Theorem 6.4](#). As a direct consequence, we can apply [Theorem 6.3](#) for estimating the truncated moments, even under *adversarial* corruption.

6.2 Learning parameters under halfspace truncation and robustness

As a first step we begin by obtaining a robust preconditioner for our setup.

Lemma 6.6. *Let $\eta > 0$ be at most a sufficiently small constant. Given an η -corruption X_1, \dots, X_n of $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$, we can efficiently compute a matrix \widetilde{M}_2 such that*

$$(1 - \zeta)M_2 \leq \widetilde{M}_2 \leq (1 + \zeta)M_2$$

whenever $\|\Sigma\|^{1/2} k \cdot \eta^{1-2/k} \leq O(1)$ and $\zeta = \|\Sigma\|^{1/4} k^{1/2} \eta^{1-2/k}$ where k is as in the above theorem of [KSS18].

The proof of [Lemma 6.6](#) follows directly by an application of [Theorem 6.3](#). Once we have access to such an \widetilde{M}_2 , if $\zeta \leq \frac{\min\{\kappa_2(\gamma), 1\}}{2}$, then as discussed in [Claim 4.8](#) in [Section 4.3](#), we can assume that

$$\begin{aligned} 0.5I_d &\leq M_2 \leq 2I_d \\ 0.5I_d &\leq \Sigma \leq \frac{2}{\kappa_2(\gamma)}I_d \end{aligned}$$

After transforming the samples we apply [Theorem 6.3](#) (again) to obtain robust estimates for the second and third moment.

Lemma 6.7. *Let $\delta > 0$ be an arbitrary constant. Let $0 < \eta \leq O(1)$ for a sufficiently small constant. Given an η -corruption X_1, \dots, X_n of $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$ (preconditioned as discussed above), where $n = \text{poly}(d)$, we can efficiently compute estimates for the first non-central moment $\hat{\mu}_{\text{trunc}}$ as well as the second and third (central) moments \widehat{M}_2 and \widehat{M}_3 of $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$ with the following guarantees*

$$\|\mu_{\text{trunc}} - \hat{\mu}_{\text{trunc}}\| \leq O_\delta \left(\log^{1/4}(1/\alpha) \cdot \eta^{1-\delta} \right)$$

$$\begin{aligned} \left(1 - O_\delta \left(\log^{1/4}(1/\alpha) \cdot \eta^{1-\delta}\right)\right) M_2 &\leq \widehat{M}_2 \leq \left(1 + O_\delta \left(\log^{1/4}(1/\alpha) \cdot \eta^{1-\delta}\right)\right) M_2 \\ \forall v \in \mathcal{S}^{d-1} : \langle M_3 - \widehat{M}_3, v^{\otimes 3} \rangle &\leq O_\delta \left(\max \left(\log(1/\alpha)^{1/4}, \log(1/\alpha)^{7/8}\right) \cdot \eta^{1/2-\delta}\right) \end{aligned}$$

whenever $\eta^{1-\delta} \leq O_\delta(\log^{-1/2}(1/\alpha))$.

Proof. We observe that for the estimate for μ_{trunc} and M_2 it suffices to apply [Theorem 6.3](#) (where C is now at most $O(\kappa_2(\gamma)^{-1/2})$) and pick k such that $\delta \geq 3/k$ for some constant k to obtain the result. Note that, by [Claim 3.22](#), we have that $\kappa_2(\gamma)^{-1} \leq O(\log(1/\alpha))$.

For the estimate for the third central moments M_3 , we cannot directly apply [Theorem 6.3](#) since they only give an estimate for the non-central moments. Instead, we again use the estimate $\hat{\mu}_{\text{trunc}}$ for the truncated mean μ_{trunc} . We note that this satisfies, as discussed before,

$$\|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\| \leq O_\delta \left(\log^{1/4}(1/\alpha) \cdot \eta^{1-\delta}\right).$$

Then, using new samples X_i (such that they are independent of $\hat{\mu}_{\text{trunc}}$), we apply the algorithm from [Theorem 6.3](#) to the samples $X'_i = X_i - \hat{\mu}_{\text{trunc}}$ to get an estimate $M'_{3,\text{nc}}$ of the third non-central moment of the uncorrupted $Y'_i = Y_i - \hat{\mu}_{\text{trunc}}$ ¹³. We use this estimator $\widehat{M}'_{3,\text{nc}}$ to estimate the third central moment M_3 of the uncorrupted Y_i . We have

$$\begin{aligned} M_3 &= \mathbb{E}[(Y_i - \mu_{\text{trunc}})^{\otimes 3}] \\ &= \mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}} + \hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})^{\otimes 3}] \\ &= \mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 3} + \text{Sym}((Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 2} \otimes (\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})) \\ &\quad + \text{Sym}((Y_i - \hat{\mu}_{\text{trunc}}) \otimes (\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})^{\otimes 2}) + (\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})^{\otimes 3}], \end{aligned}$$

where we use Sym to denote all permutations (i.e., $\text{Sym}(a^{\otimes 2} \otimes b) = a \otimes a \otimes b + a \otimes b \otimes a + b \otimes a \otimes a$). Now, [Theorem 6.3](#) gives an estimate to $M'_{3,\text{nc}} := \mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 3}]$ with the following guarantee

$$\langle M'_{3,\text{nc}} - \widehat{M}'_{3,\text{nc}}, v^{\otimes 3} \rangle \leq O_\delta \left(\log(1/\alpha)^{3/8} \cdot \eta^{1/2-\delta} \cdot \|\mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 2}]\|^{3/2}\right) \quad (6.1)$$

for any unit vector v . We can bound

$$\begin{aligned} \|\mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 2}]\| &= \|\mathbb{E}[(Y_i - \mu_{\text{trunc}})^{\otimes 2}] + (\mu_{\text{trunc}} - \hat{\mu}_{\text{trunc}})^{\otimes 2}\| \\ &\leq \|\mathbb{E}[(Y_i - \mu_{\text{trunc}})^{\otimes 2}]\| + \|(\mu_{\text{trunc}} - \hat{\mu}_{\text{trunc}})^{\otimes 2}\| \\ &\leq 2 + \|\mu_{\text{trunc}} - \hat{\mu}_{\text{trunc}}\|^2, \end{aligned} \quad (6.2)$$

where we used in the last step that the first term is just the second central moment of the (truncated) Y_i , which satisfies $M_2 \leq 2I_d$. We now get the following for the error of our estimate, where v is again a unit vector,

$$\langle M_3 - \widehat{M}'_{3,\text{nc}}, v^{\otimes 3} \rangle = \langle M_3 - M'_{3,\text{nc}}, v^{\otimes 3} \rangle + \langle M'_{3,\text{nc}} - \widehat{M}'_{3,\text{nc}}, v^{\otimes 3} \rangle.$$

We can bound the second term by (6.1). For the first term, note that

$$\langle M_3 - M'_{3,\text{nc}}, v^{\otimes 3} \rangle = \langle \text{Sym}(\mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 2}] \otimes (\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})), v^{\otimes 3} \rangle$$

¹³Recall that Y corresponds to uncorrupted samples and X corresponds to corrupted samples.

$$\begin{aligned}
& + \langle \text{Sym}(\mathbb{E}[Y_i - \hat{\mu}_{\text{trunc}}] \otimes (\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})^{\otimes 2}), v^{\otimes 3} \rangle \\
& + \langle (\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}})^{\otimes 3}, v^{\otimes 3} \rangle \\
& \leq 4 \|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\|^3 + 3 \|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\| \cdot \|\mathbb{E}[(Y_i - \hat{\mu}_{\text{trunc}})^{\otimes 2}]\| \\
& \leq 7 \|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\|^3 + 6 \|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\|,
\end{aligned}$$

where in the last step we used (6.2). Combining this with (6.1), we get

$$\begin{aligned}
\langle M_3 - \widehat{M}'_{3,\text{nc}}, v^{\otimes 3} \rangle & \leq O_\delta \left(\log(1/\alpha)^{3/8} \cdot \eta^{1/2-\delta} \cdot (2 + \|\mu_{\text{trunc}} - \hat{\mu}_{\text{trunc}}\|^2) \right) \\
& \quad + 7 \|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\|^3 + 6 \|\hat{\mu}_{\text{trunc}} - \mu_{\text{trunc}}\| \\
& \leq O_\delta \left(\max \left(\log(1/\alpha)^{1/4}, \log(1/\alpha)^{7/8} \right) \cdot \eta^{1/2-\delta} \right),
\end{aligned}$$

where we used $\eta \leq 1$ for the last step. □

We now prove the following theorem.

Main robustness result

Theorem 6.8 (Full statement of [Theorem 1.2](#)). *Let $0 < \delta < \frac{1}{2}$ be an arbitrary constant and let $\alpha \leq 0.99$. Define*

$$L := \max \left(\log^{1/4}(1/\alpha), \log^{7/8}(1/\alpha) \right)$$

and

$$\beta_{\text{robust}} := \max \left\{ 1, \frac{1}{|\kappa_3(\gamma)|} \right\} \cdot \sqrt{d} \cdot L \cdot \eta^{1/2-\delta}.$$

Let c_δ be a sufficiently small constant depending only on δ . Assume that

$$0 < L^{\frac{2}{1-2\delta}} \cdot \eta \leq O_\delta \left(\left(\frac{\text{poly}(\kappa_2(\gamma), \kappa_3(\gamma))}{\sqrt{d}} \right)^{\frac{2}{1-2\delta}} \right). \quad (6.3)$$

The polynomial in (6.3) is chosen sufficiently small so that

$$L\eta^{1-\delta} \leq O_\delta(\beta_{\text{robust}}), \quad \beta_{\text{robust}} \leq c_\delta \cdot \left(\min \{ \kappa_2(\gamma)^{5/2}, |\kappa_3(\gamma)| \} \right),$$

and so that [Lemma 6.7](#) applies. Furthermore, assume that

$$\eta^{1-\delta} \leq \min \left\{ O_\delta(\log^{-1/2}(1/\alpha)), c_\delta \cdot \|\Sigma\|^{-1/2}, c_\delta \cdot \|\Sigma\|^{-1/4} \cdot \kappa_2(\gamma) \right\}. \quad (6.4)$$

Given an η -corruption X_1, \dots, X_n of $\mathcal{N}(\mu, \Sigma)|_{\{w^\top x \leq \tau\}}$, where $n = \text{poly}(d)$, with high probability, we can, in time $\text{poly}(d)$, compute estimators $\hat{\mu}$ and $\widehat{\Sigma}$ such that

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq O_\delta \left(\frac{1}{\kappa_2(\gamma)^{7/2}} \cdot \beta_{\text{robust}} \right)$$

and

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\|_F \leq O_\delta \left(\frac{1}{\kappa_2(\gamma)^{7/2}} \cdot \beta_{\text{robust}} \right).$$

Proof. We follow the proof of [Theorem 5.1](#) but replace the sample moments with our robust estimates from [Lemma 6.7](#). Note that we can do the preconditioning as discussed above by (6.4). Also note that we can apply [Lemma 6.7](#) by (6.4). By the robust preconditioning step and affine invariance, it suffices to prove the theorem in the preconditioned coordinates, where

$$\frac{1}{2}I_d \leq M_2 \leq 2I_d, \quad \frac{1}{2}I_d \leq \Sigma \leq \frac{2}{\kappa_2(\gamma)}I_d.$$

Covariance estimation.

In order to apply [Lemma 5.13](#) for the covariance estimation, we need to compute \hat{u} and $\hat{\gamma}$ and bound the following four quantities ((5.10), (5.11), (5.13) and (5.12) respectively):

1. $\|\Sigma^{-1/2}(\widehat{M}_2 - M_2)\Sigma^{-1/2}\|_F,$
2. $\min \left\{ \left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\|, \left\| \hat{u} + \frac{u^*}{\|u^*\|} \right\| \right\},$
3. $|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)|,$
4. $\left| \hat{u}^T \widehat{M}_2^{-1} \hat{u} - \frac{\|\Sigma^{1/2}w\|^2}{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)} \right|.$

We remark that showing 3. is enough to show (5.13) by the proof of [Lemma 5.9](#).

Bounding 1. From [Lemma 6.7](#), we immediately get

$$\|\Sigma^{-1/2}(\widehat{M}_2 - M_2)\Sigma^{-1/2}\|_F \leq \sqrt{d} \cdot \|\Sigma^{-1/2}(\widehat{M}_2 - M_2)\Sigma^{-1/2}\| \quad (6.5)$$

$$\leq \sqrt{d} \cdot O_\delta(\eta^{1-\delta} \cdot L). \quad (6.6)$$

The first inequality holds because for any matrix A we have that the Frobenius norm satisfies $\|A\|_F \leq \sqrt{\text{rank}(A)} \cdot \|A\|$ and the last inequality follows from [Lemma 6.7](#). As shown in [[KSS18](#)] the polynomial dependency on the dimension is *necessary* in general for certifiably sub-Gaussian distributions.

Bounding 2. It suffices that we get a good estimate of M_3 . Indeed, we are promised such a good estimate from [Lemma 6.7](#). This directly implies that we can find a unit vector $\hat{u} \in \mathbb{R}^d$ such that

$$\min \left\{ \left\| \hat{u} + \frac{u^*}{\|u^*\|} \right\|_2, \left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\|_2 \right\} \leq O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \quad (6.7)$$

Bounding 3. From (6.3) we have that

$$\max \left\{ O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right), O_\delta(\eta^{1-\delta} \cdot L), O_\delta(\eta^{1/2-\delta} \cdot L) \right\} = O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \leq c,$$

where c is a sufficiently small constant¹⁴. This enables us to use [Lemma 5.9](#) and compute an estimate $\hat{\gamma}$ with β in that lemma set to $O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right)$. Therefore we have

$$|\text{skew}(\hat{\gamma}) - \text{skew}(\gamma)| \leq O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \cdot \kappa_2(\gamma)^{-3/2}. \quad (6.8)$$

¹⁴Note that this is *weaker* than the condition that our theorem statement imposes on η .

Bounding 4. The second conclusion of [Lemma 5.9](#), applied with

$$\beta = O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right),$$

gives

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \tilde{u}^\top M_2^{-1} \tilde{u} \right| \leq O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right).$$

Using [\(5.1\)](#), where $\tilde{u} = u^*/\|u^*\| - \Sigma w/\|\Sigma w\|$

$$\tilde{u}^\top M_2^{-1} \tilde{u} = \frac{\|\Sigma^{1/2} w\|^2}{\kappa_2(\gamma) \|\Sigma w\|^2} = \frac{\kappa_3(\gamma)^{2/3}}{\|u^*\|^2 \kappa_2(\gamma)}.$$

Therefore

$$\left| \hat{u}^\top \widehat{M}_2^{-1} \hat{u} - \frac{\|\Sigma^{1/2} w\|^2}{\|\Sigma w\|^2 \cdot \kappa_2(\gamma)} \right| \leq O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \quad (6.9)$$

Conclusion for covariance estimation. Combining [\(6.5\)](#), [\(6.7\)](#), [\(6.8\)](#) and [\(6.9\)](#) we can apply [Lemma 5.13](#) with

$$\beta = O_\delta \left(\max \left\{ 1, \frac{1}{|\kappa_3(\gamma)|} \right\} \cdot \sqrt{d} \cdot L \cdot \eta^{1/2-\delta} \right)$$

Note that by [\(6.3\)](#) we have $\beta \lesssim \kappa_2(\gamma)^{5/2}$. This gives us an estimator $\widehat{\Sigma}$ with error

$$\|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d\|_F \leq O_\delta \left(\frac{1}{\kappa_2(\gamma)^{7/2}} \cdot \max \left\{ 1, \frac{1}{|\kappa_3(\gamma)|} \right\} \cdot \sqrt{d} \cdot L \cdot \eta^{1/2-\delta} \right).$$

Mean estimation.

For mean estimation, given [Lemma 5.11](#), we only need to argue that we can additionally bound [\(5.5\)](#) and [\(5.7\)](#)

5. $\left\| \bar{x} - \mu - \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|} \right\|$
6. $\left\| \hat{u} - \frac{u^*}{\|u^*\|} \right\|$

Bounding 5. By [Lemma 6.7](#), we are able to get an estimate $\hat{\mu}_{\text{trunc}}$ for the mean of the truncated distribution $\mu_{\text{trunc}} = \mu + \kappa_1(\gamma) \frac{\Sigma w}{\|\Sigma^{1/2} w\|}$. This estimate satisfies

$$\|\mu_{\text{trunc}} - \hat{\mu}_{\text{trunc}}\| \leq O_\delta (L \cdot \eta^{1-\delta}). \quad (6.10)$$

Thus we can take $\bar{x} = \hat{\mu}_{\text{trunc}}$ and we are done.

Bounding 6. Given that we already bounded 2., we only need to determine the sign of $\frac{u^*}{\|u^*\|}$. We can do so using the third moment tensor. Let

$$\tilde{u} := \frac{u^*}{\|u^*\|}.$$

Since $M_3 = (u^*)^{\otimes 3}$, we have

$$\langle M_3, \tilde{u}^{\otimes 3} \rangle = \langle (u^*)^{\otimes 3}, \tilde{u}^{\otimes 3} \rangle = \langle u^*, \tilde{u} \rangle^3 = \|u^*\|^3 > 0.$$

Let $a \in \{\pm 1\}$ be the sign such that

$$\|a\hat{u} - \tilde{u}\| \leq O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right).$$

from (6.7). We can compute

$$\begin{aligned} \langle \widehat{M}_3, (a\hat{u})^{\otimes 3} \rangle &= \langle \widehat{M}_3 - M_3, (a\hat{u})^{\otimes 3} \rangle + \langle M_3, (a\hat{u})^{\otimes 3} \rangle \\ &= \langle \widehat{M}_3 - M_3, (a\hat{u})^{\otimes 3} \rangle + \langle M_3, (a\hat{u})^{\otimes 3} - \tilde{u}^{\otimes 3} \rangle + \langle M_3, \tilde{u}^{\otimes 3} \rangle. \end{aligned}$$

We bound the first term by [Lemma 6.7](#):

$$\left| \langle \widehat{M}_3 - M_3, (a\hat{u})^{\otimes 3} \rangle \right| \leq O_\delta \left(L \cdot \eta^{1/2-\delta} \right).$$

For the second term, using $M_3 = (u^*)^{\otimes 3}$ and the bound on $\|a\hat{u} - \tilde{u}\|$, we get

$$\begin{aligned} \left| \langle M_3, (a\hat{u})^{\otimes 3} - \tilde{u}^{\otimes 3} \rangle \right| &= \left| \langle u^*, a\hat{u} \rangle^3 - \langle u^*, \tilde{u} \rangle^3 \right| \\ &\leq O_\delta \left(\|u^*\|^3 \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \right), \end{aligned}$$

where we used that $a\hat{u}$ and \tilde{u} are unit vectors and $O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \leq O(1)$. Therefore,

$$\langle \widehat{M}_3, (a\hat{u})^{\otimes 3} \rangle \geq \|u^*\|^3 - O_\delta \left(L\eta^{1/2-\delta} \right) - O \left(\|u^*\|^3 (\sqrt{d}/|\kappa_3(\gamma)|) L\eta^{1/2-\delta} \right).$$

Hence, since by (6.3) and

$$\|u^*\|^3 \gtrsim |\kappa_3(\gamma)|$$

in the precondition coordinates, we have

$$O_\delta \left(L\eta^{1/2-\delta} \right) + O_\delta \left(\|u^*\|^3 \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right) \right) < \frac{1}{2} \|u^*\|^3,$$

and we get

$$\langle \widehat{M}_3, (a\hat{u})^{\otimes 3} \rangle > 0.$$

Since the tensor is order three,

$$\langle \widehat{M}_3, (-a\hat{u})^{\otimes 3} \rangle = -\langle \widehat{M}_3, (a\hat{u})^{\otimes 3} \rangle < 0.$$

Thus, we can determine the correct sign for \hat{u} by choosing the sign for which

$$\langle \widehat{M}_3, \hat{u}^{\otimes 3} \rangle > 0.$$

Recall that we need to find a vector parallel to $-\Sigma w$ and u^* is a vector that points in the direction $-\Sigma w$. After this sign choice, we have

$$\left\| \hat{u} + \frac{\Sigma w}{\|\Sigma w\|} \right\| \leq O_\delta \left(\frac{\sqrt{d}}{|\kappa_3(\gamma)|} \cdot L \cdot \eta^{1/2-\delta} \right).$$

Conclusion for mean estimation. Combining (6.10), (6.5), (6.7), (6.9) and (6.8), we can apply Lemma 5.11

$$\beta = O_\delta \left(\max \left\{ 1, \frac{1}{|\kappa_3(\gamma)|} \right\} \cdot \sqrt{d} \cdot L \cdot \eta^{1/2-\delta} \right).$$

Note that by (6.3) we have $\beta \lesssim \kappa_2(\gamma)^{5/2}$. This gives us an estimator $\hat{\mu}$ with error

$$\begin{aligned} \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| &\leq O(\|\hat{\mu} - \mu\|) \\ &\leq O_\delta \left(\frac{1}{\kappa_2(\gamma)^{7/2}} \cdot \max \left\{ 1, \frac{1}{|\kappa_3(\gamma)|} \right\} \cdot \sqrt{d} \cdot L \cdot \eta^{1/2-\delta} \right). \end{aligned}$$

□

7 Extension to intersections of two orthogonal halfspaces

In this section, we show how to extend our algorithm to truncations sets that are the intersection of two orthogonal halfspaces. Suppose $d \geq 2$ and consider a random vector X that follows a d -dimensional Gaussian $\mathcal{N}(\mu, I_d)$ truncated to the intersection of H_1 and H_2 , where $H_1 = \{x : \langle w_1, x \rangle \leq \tau_1\}$ and $H_2 = \{x : \langle w_2, x \rangle \leq \tau_2\}$ are halfspaces with $\|w_1\| = \|w_2\| = 1$ and $w_1 \perp w_2$. As before, we define $\alpha > 0$ to be the probability measure of $H_1 \cap H_2$ under the underlying Gaussian. We will prove the following result. For simplicity, we assume that the underlying Gaussian has identity covariance and we only aim to estimate the mean.

Theorem 7.1. *There is an algorithm (Algorithm 3) that takes $n = O\left(\frac{d^2}{\varepsilon^2} \text{poly log}\left(\frac{1}{\alpha}\right) + \frac{d^2}{\varepsilon^4}\right)$ i.i.d. samples from $\mathcal{N}(\mu, I_d)$ truncated to $H_1 \cap H_2$ as input and efficiently computes an estimate $\hat{\mu}$, such that with probability at least 0.99 it satisfies*

$$d_{\text{TV}}(\mathcal{N}(\hat{\mu}, I_d), \mathcal{N}(\mu, I_d)) \leq \varepsilon.$$

The runtime of the algorithm is $O(T(n, d))$, where $T(n, d)$ is the time needed to multiply a $d \times n$ matrix with its transpose.

Remark 7.2. Before proving Theorem 7.1, we want to briefly discuss potential further extensions:

1. While we only prove our extension for mean estimation, it seems plausible that one can also do covariance estimation. In this case, we would want that w_1 and w_2 are ‘ Σ -orthogonal’, meaning that Σw_1 and Σw_2 are orthogonal. One thing that is unclear however is whether and how we could do preconditioning in this case to avoid a dependence on the condition number (as preconditioning need not preserve orthogonality).
2. Finally, the correction terms in each direction are (almost) independent of each other due to orthogonality. This should enable us to generalize our estimator to an intersection of k pairwise orthogonal halfspaces (for $k \leq d$). The required sample and time complexity in this case would depend on the minimum eigenvalue gap after the random contraction, which asymptotically could be as small as $\tilde{O}\left(\frac{1}{k^2}\right)$ and the sample complexity should increase by a factor that is polynomial in k . If there is non-trivial truncation in each direction, we would thus expect a sample complexity of $O(d^2/\varepsilon^2) \cdot \text{polylog}(1/\alpha) \cdot \text{poly}(k)$.

Algorithm 3: Algorithm for estimating the mean of an identity-covariance Gaussian truncated by two orthogonal halfspaces

Input: Dimension d , sample size n , samples $\{x_i\}_{i=1}^n$, target accuracy ε .

Output: Estimated mean $\hat{\mu} \in \mathbb{R}^d$.

1. Compute sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$;
 2. Compute sample second central moment: $\widehat{M}_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^{\otimes 2}$;
 3. Draw a standard Gaussian vector: $g \sim \mathcal{N}(0, I_d)$;
 4. Compute the random contraction: $R = (I_d \otimes I_d \otimes g^\top) \cdot \widehat{M}_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^{\otimes 2} \cdot \langle g, x_i - \bar{x} \rangle$.
 5. Compute \hat{w}_1 and \hat{w}_2 , the eigenvectors of R corresponding to the two largest eigenvalue in absolute value (with unit norm).
 6. For $\ell \in \{1, 2\}$, flip the sign of \hat{w}_ℓ if $\langle \hat{w}_\ell^{\otimes 3}, \widehat{M}_3 \rangle > 0$.
 7. For $\ell \in \{1, 2\}$, estimate the relative truncation parameters γ_ℓ : $\hat{\gamma}_\ell = \kappa_2^{-1} \left(\hat{w}_\ell^\top \widehat{M}_2 \hat{w}_\ell \right)$.
 8. Estimate the mean of the Gaussian: $\hat{\mu} = \bar{x} - \kappa_1(\hat{\gamma}_1) \cdot \hat{w}_1 - \kappa_1(\hat{\gamma}_2) \cdot \hat{w}_2$.
 9. Return estimate $\hat{\mu}$.
-

7.1 Moments

As for our main result, we first compute the first three moments of $\mathcal{N}(\mu, I_d)|_{H_1 \cap H_2}$. The values γ_1, γ_2 will be defined analogously to the relative truncation parameter in the main result.

Lemma 7.3. *For a random vector X defined as above, there are real numbers γ_1, γ_2 such that*

$$\begin{aligned} \mathbb{E} X &= \mu + \kappa_1(\gamma_1) \cdot w_1 + \kappa_1(\gamma_2) \cdot w_2 \\ M_2 &:= \mathbb{E}(X - \mathbb{E} X)^{\otimes 2} = I_d - (1 - \kappa_2(\gamma_1)) \cdot w_1 w_1^\top - (1 - \kappa_2(\gamma_2)) \cdot w_2 w_2^\top \\ M_3 &:= \mathbb{E}(X - \mathbb{E} X)^{\otimes 3} = \kappa_3(\gamma_1) \cdot (w_1)^{\otimes 3} + \kappa_3(\gamma_2) \cdot (w_2)^{\otimes 3} \end{aligned}$$

Moreover, we have

$$\alpha = \Phi(\gamma_1) \cdot \Phi(\gamma_2).$$

Proof. For $d \geq 2$, since w_1, w_2 are perpendicular unit vectors, there exists an orthonormal matrix U such that $Ue_1 = w_1$ and $Ue_2 = w_2$. As before, we consider $Y = U^\top(X - \mu)$. This random variable follows a standard Gaussian $\mathcal{N}(0, I_n)$ truncated as follows:

$$\begin{aligned} w_1^\top X \leq \tau_1 &\iff (U^\top w_1)^\top Y \leq \tau_1 - w_1^\top \mu \\ w_2^\top X \leq \tau_2 &\iff (U^\top w_2)^\top Y \leq \tau_2 - w_2^\top \mu. \end{aligned}$$

Since $Ue_1 = w_1$ and $Ue_2 = w_2$, we get a simple characterization of Y : Defining $\gamma_1 = \tau_1 - \langle w_1, \mu \rangle$ and $\gamma_2 = \tau_2 - \langle w_2, \mu \rangle$, Y_1 and Y_2 are truncated standard Gaussian with thresholds γ_1 and γ_2 , all other coordinates are standard Gaussian and all coordinates are mutually independent. Thus, as in [Lemma 4.2](#), we get that

$$\begin{aligned} \mathbb{E} Y &= \kappa_1(\gamma_1) e_1 + \kappa_1(\gamma_2) e_2, \\ \mathbb{E}(Y - \mathbb{E} Y)^{\otimes 2} &= I_d - (1 - \kappa_2(\gamma_1)) e_1 e_1^\top - (1 - \kappa_2(\gamma_2)) e_2 e_2^\top, \end{aligned}$$

$$\mathbb{E}(Y - \mathbb{E}Y)^{\otimes 3} = \kappa_3(\gamma_1)e_1^{\otimes 3} + \kappa_3(\gamma_2)e_2^{\otimes 3}.$$

Using this, we can compute the moments of X as follows. First, the mean of X is

$$\mathbb{E}X = U \mathbb{E}Y + \mu = \mu + \kappa_1(\gamma_1) \cdot Ue_1 + \kappa_1(\gamma_2) \cdot Ue_2 = \mu + \kappa_1(\gamma_1) \cdot w_1 + \kappa_1(\gamma_2) \cdot w_2.$$

The covariance of X is

$$\begin{aligned} M_2 &= \mathbb{E}(X - \mathbb{E}X)^{\otimes 2} = U^{\otimes 2} \mathbb{E}(Y - \mathbb{E}Y)^{\otimes 2} \\ &= U (I_d - (1 - \kappa_2(\gamma_1)) \cdot e_1e_1^\top - (1 - \kappa_2(\gamma_2)) \cdot e_2e_2^\top) U^\top \\ &= I_d - (1 - \kappa_2(\gamma_1)) \cdot w_1w_1^\top - (1 - \kappa_2(\gamma_2)) \cdot w_2w_2^\top. \end{aligned}$$

And the third central moment of X equals

$$\begin{aligned} M_3 &= \mathbb{E}(X - \mathbb{E}X)^{\otimes 3} = U^{\otimes 3} \mathbb{E}(Y - \mathbb{E}Y)^{\otimes 3} \\ &= U^{\otimes 3} (\kappa_3(\gamma_1) \cdot e_1^{\otimes 3} + \kappa_3(\gamma_2) \cdot e_2^{\otimes 3}) \\ &= \kappa_3(\gamma_1) \cdot w_1^{\otimes 3} + \kappa_3(\gamma_2) \cdot w_2^{\otimes 3}. \end{aligned}$$

Finally, define $H'_1 = \{y \in \mathbb{R}^d \mid y_1 \leq \gamma_1\}$ and $H'_2 = \{y \in \mathbb{R}^d \mid y_2 \leq \gamma_2\}$. Then, we have that

$$\begin{aligned} \alpha &= \mathbb{E}_{X \sim \mathcal{N}(\mu, I_d)} [\mathbb{1}[X \in H_1 \cap H_2]] \\ &= \mathbb{E}_{Y \sim \mathcal{N}(0, I_d)} [\mathbb{1}[Y \in H'_1 \cap H'_2]] \\ &= \mathbb{P}_{Y_1 \sim \mathcal{N}(0, 1)} [Y_1 \leq \gamma_1] \cdot \mathbb{P}_{Y_2 \sim \mathcal{N}(0, 1)} [Y_2 \leq \gamma_2] \\ &= \Phi(\gamma_1) \cdot \Phi(\gamma_2), \end{aligned}$$

which completes the proof. \square

Since we assume that the covariance of the underlying Gaussian is the identity, we don't need to precondition the samples. Instead, we directly get that

$$\min\{\kappa_2(\gamma_1), \kappa_2(\gamma_2)\} \leq M_2 \leq I_d. \quad (7.1)$$

7.2 Relative truncation parameter estimation

By orthogonality between w_1, w_2 , we could isolate them from the second central moment as long as γ_1, γ_2 are *constantly-separated*. However, this is not true in general. It might be that $\gamma_1 = \gamma_2$, which makes it information-theoretically impossible to identify w_1, w_2 from M_2 already when $d = 2$. Therefore, as in the case of a single halfspace, we use the third central moment M_3 . Note that with probability 1, we have that $\kappa_3(\gamma_1) \cdot \langle g, w_1 \rangle$ and $\kappa_3(\gamma_2) \cdot \langle g, w_2 \rangle$ are distinct. Furthermore, we will be able to bound the distance between these eigenvalues using [Fact 3.24](#). We will use the following theorem from matrix perturbation theory (see also [[Bha97](#), Corollary III.2.6 and Theorem VII.3.2]).

Theorem 7.4 ([[DK70](#), [Wey12](#)]). *Let $A, E \in \mathbb{R}^{d \times d}$ be symmetric matrices. Consider the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ of A with associated orthonormal eigenvectors v_i and the eigenvalues $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_d$ and orthonormal eigenvectors \tilde{v}_i of $\tilde{A} = A + E$. Define $\delta_i = \min_{j \neq i} |\lambda_j - \lambda_i|$. Then, if $\delta_i > 0$, we have*

$$\min \{ \|\tilde{v}_i + v_i\|_2, \|\tilde{v}_i - v_i\|_2 \} \leq C \frac{\|E\|}{\delta_i}$$

for an absolute constant $C > 0$. Furthermore, we have that

$$|\lambda_i - \tilde{\lambda}_i| \leq \|E\|.$$

We now show the following lemma, which is analogous to [Section 5.1](#). We note that after step 6 of [Algorithm 3](#), the estimates \hat{w}_1 and \hat{w}_2 are exactly the ones from [Lemma 7.5](#).

Lemma 7.5. *Let $\beta > 0$. Given n i.i.d. samples from $\mathcal{N}(\mu, \Sigma)$ truncated to $H_1 \cap H_2$, we can compute orthogonal unit vectors \hat{w}_1, \hat{w}_2 such that the following hold for $\ell \in \{1, 2\}$: If*

$$n = O\left(\frac{d^2}{\beta^2 \kappa_3(\gamma_\ell)^2} + \frac{d^2}{\kappa_3(\gamma_\ell)^4}\right),$$

then we have with high probability that $\|\hat{w}_\ell - w_\ell\| \leq \beta$.

Proof. We want to use [Theorem 7.4](#) for the matrix $A = (I_d \otimes I_d \otimes g^\top)M_3$ and $\tilde{A} = (I_d \otimes I_d \otimes g^\top)\widehat{M}_3$. Let the spectral decomposition of $(I_d \otimes I_d \otimes g^\top)\widehat{M}_3$ be

$$(I_d \otimes I_d \otimes g^\top)\widehat{M}_3 = \sum_{j=1}^d \tilde{\lambda}_j \tilde{v}_j \tilde{v}_j^\top,$$

where $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_d$. Since this matrix is symmetric, we can assume that the \tilde{v}_j are an orthonormal basis of \mathbb{R}^d . To apply [Theorem 7.4](#), we need to determine the eigenvalues and eigenvectors of $(I_d \otimes I_d \otimes g^\top)M_3$. We have, since $\Sigma = I_d$,

$$\begin{aligned} M_3 &= \kappa_3(\gamma_1)w_1^{\otimes 3} + \kappa_3(\gamma_2)w_2^{\otimes 3}, \\ (I_d \otimes I_d \otimes g^\top)M_3 &= \kappa_3(\gamma_1) \cdot \langle g, w_1 \rangle \cdot w_1 w_1^\top + \kappa_3(\gamma_2) \cdot \langle g, w_2 \rangle \cdot w_2 w_2^\top. \end{aligned}$$

Thus, the eigenvalues of $(I_d \otimes I_d \otimes g^\top)M_3$ are $\kappa_3(\gamma_1) \cdot \langle g, w_1 \rangle$, $\kappa_3(\gamma_2) \cdot \langle g, w_2 \rangle$ and 0 with multiplicity $n - 2$. Furthermore, the corresponding eigenvectors are w_1 and w_2 (and an orthogonal basis of the complement). Let $i_1, i_2 \in [d]$ be such that $v_{i_1} = w_1$ and $v_{i_2} = w_2$. We will define our estimates as the two eigenvectors of \tilde{A} corresponding to the largest eigenvalues in absolute value (up to a sign choice that we discuss later). This already implies that they are orthogonal.

If we want one of our two estimates to be a good estimate for w_ℓ , we need to ensure that even after the perturbation, the eigenvalue corresponding to w_ℓ is one of the two largest in absolute value (i.e. it should be later than the perturbation of the 0 eigenvalues). We can ensure this as follows: The eigenvalues of A that are 0 correspond to eigenvalues of \tilde{A} of absolute value at most $\|A - \tilde{A}\|$. Thus, if the eigenvalue for w_ℓ is larger than $2 \cdot \|A - \tilde{A}\|$ in absolute value, even after perturbation it will belong to the largest two (hence, one of our estimates will be close to w_ℓ up to sign). Hence, if we ensure that $\|A - \tilde{A}\| \leq \frac{1}{2} \cdot |\kappa_3(\gamma_\ell) \cdot \langle g, w_\ell \rangle|$, it suffices to argue that one of $\|w_\ell \pm \tilde{v}_{i_\ell}\|$ is small and how to recover the sign.

In order to apply [Theorem 7.4](#), we also need to compute the values of δ_{i_1} and δ_{i_2} . For this, note that $\langle g, \kappa_3(\gamma_1) \cdot w_1 - \kappa_3(\gamma_2) \cdot w_2 \rangle$ is distributed identical to the Gaussian $\mathcal{N}(0, \kappa_3(\gamma_1)^2 + \kappa_3(\gamma_2)^2)$. Thus, by Gaussian anti-concentration (cf. [Fact 3.24](#)), we get that, with high probability,

$$|\kappa_3(\gamma_1) \cdot \langle g, w_1 \rangle - \kappa_3(\gamma_2) \cdot \langle g, w_2 \rangle| \geq \Omega\left(\sqrt{\kappa_3(\gamma_1)^2 + \kappa_3(\gamma_2)^2}\right).$$

Similarly, we get

$$|\kappa_3(\gamma_1) \cdot \langle g, w_1 \rangle| \geq \Omega(|\kappa_3(\gamma_1)|) \quad \text{and} \quad |\kappa_3(\gamma_2) \cdot \langle g, w_2 \rangle| \geq \Omega(|\kappa_3(\gamma_2)|).$$

Thus, we can lower bound

$$\delta_{i_1} \geq \Omega(|\kappa_3(\gamma_1)|) \quad \text{and} \quad \delta_{i_2} \geq \Omega(|\kappa_3(\gamma_2)|).$$

It remains to bound $\|(I_d \otimes I_d \otimes g^\top) \widehat{M}_3 - (I_d \otimes I_d \otimes g^\top) M_3\|$. First, analogous to [Lemma 5.8](#), we get that, with high probability,

$$\sup_{\|v\|=1} \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \lesssim \sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n}.$$

Note that we get a slightly better result here since the samples x_i are in fact 1-sub-Gaussian in this case (because $\Sigma = I_d$) and not only $O(1/\sqrt{\kappa_2(\gamma)})$ -sub-Gaussian as before. Now, observe that with high probability,

$$\begin{aligned} \left\| (I_d \otimes I_d \otimes g^\top) (\widehat{M}_3 - M_3) \right\| &\leq \|g\| \sup_{\|v\|=1} \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \\ &\leq 3\sqrt{d} \cdot \sup_{\|v\|=1} \left\langle v^{\otimes 3}, \widehat{M}_3 - M_3 \right\rangle \\ &\lesssim \frac{d}{\sqrt{n}} + \frac{d^2}{n}. \end{aligned}$$

Applying [Theorem 7.4](#), we conclude that, for $\ell \in \{1, 2\}$,

$$\min_{b \in \{\pm 1\}} \{ \|b \cdot \tilde{v}_{i_\ell} - w_\ell\| \} \lesssim \kappa_3(\gamma_\ell)^{-1} \cdot \left(\frac{d}{\sqrt{n}} + \frac{d^2}{n} \right).$$

Let $\ell \in \{1, 2\}$. It remains to argue that with the claimed sample complexity, we can make this error at most β and that we can recover the sign. Note that for the claimed sample complexity we indeed have, using the above bound (and assuming $\beta \leq O(1)$), that

$$\|(I_d \otimes I_d \otimes g^\top) (\widehat{M}_3 - M_3)\| \leq \frac{1}{2} \cdot |\kappa_3(\gamma_\ell) \cdot \langle g, w_\ell \rangle|,$$

so \tilde{v}_{i_ℓ} will correspond to one of the two largest eigenvalues of \tilde{A} up to sign. If

$$n = O\left(\frac{d^2}{\beta'^2 \cdot \kappa_3(\gamma_\ell)^2}\right),$$

then we get that

$$\min_{b \in \{\pm 1\}} \{ \|b \cdot \tilde{v}_{i_\ell} - w_\ell\| \} \leq \beta'.$$

To determine the sign, we again check the sign of $\langle \tilde{v}_{i_\ell}^{\otimes 3}, \widehat{M}_3 \rangle$. We have $\langle w_\ell^{\otimes 3}, M_3 \rangle = \kappa_3(\gamma_\ell) < 0$. Denote by $b_\ell \in \{\pm 1\}$ the correct sign for \tilde{v}_{i_ℓ} (i.e. the one for which $\|b_\ell \tilde{v}_{i_\ell} - w_\ell\| \leq \beta$). Then, we have that

$$\left| \kappa_3(\gamma_\ell) - \left\langle b_\ell \tilde{v}_{i_\ell}^{\otimes 3}, \widehat{M}_3 \right\rangle \right| \leq \left| \langle (w_\ell - b_\ell \tilde{v}_{i_\ell})^{\otimes 3}, M_3 \rangle \right| + \left| \left\langle b_\ell \tilde{v}_{i_\ell}^{\otimes 3}, M_3 - \widehat{M}_3 \right\rangle \right|.$$

We can bound the first term using $M_3 = \kappa_3(\gamma_1) w_1^{\otimes 3} + \kappa_3(\gamma_2) w_2^{\otimes 3}$ as follows, assuming $\beta' \leq O(1)$ (such that we can bound $\beta'^3 \leq O(\beta')$),

$$\left| \langle (w_\ell - b_\ell \tilde{v}_{i_\ell})^{\otimes 3}, M_3 \rangle \right| \lesssim (|\kappa_3(\gamma_1)| + |\kappa_3(\gamma_2)|) \cdot \|w_\ell - b_\ell \tilde{v}_{i_\ell}\|^3 \lesssim \beta'.$$

For the second term, note that the sample complexity above implies¹⁵

$$\left| \left\langle b_\ell \tilde{v}_{i_\ell}^{\otimes 3}, M_3 - \widehat{M}_3 \right\rangle \right| \lesssim \beta'.$$

¹⁵We could even get $|\kappa_3(\gamma_\ell)| \cdot \frac{\beta'}{\sqrt{d}}$, but the bounding this by β' is sufficient.

Thus, we get that

$$\left| \kappa_3(\gamma_\ell) - \langle b_\ell \tilde{v}_{i_\ell}^{\otimes 3}, \widehat{M}_3 \rangle \right| \lesssim \beta'.$$

If $\beta' \leq c \cdot \kappa_3(\gamma_\ell)$ for a sufficiently small constant c , then $\langle b_\ell \tilde{v}_{i_\ell}^{\otimes 3}, \widehat{M}_3 \rangle < 0$ and we can determine the sign b_ℓ from $\langle \tilde{v}_{i_\ell}^{\otimes 3}, \widehat{M}_3 \rangle$. Summarized, for $\beta' \leq c \cdot \kappa_3(\gamma_\ell)$ (which also implies $\beta' \leq O(1)$ by [Claim 3.10](#)), we can identify the sign b_ℓ such that $\|b_\ell \tilde{v}_{i_\ell} - w_\ell\| \leq \beta'$ if we get

$$n = O\left(\frac{d^2}{\beta'^2 \cdot \kappa_3(\gamma_\ell)^2}\right)$$

samples. We can now pick $\beta' = \min\{\beta, c \cdot \kappa_3(\gamma_\ell)\}$ to get the result. \square

7.3 Proof of [Theorem 7.1](#)

With the learned normal vectors \hat{w}_1 and \hat{w}_2 , we can compute $\hat{w}_\ell^\top \widehat{M}_2 \hat{w}_\ell$, which yields an approximation to $\kappa_2(\gamma_\ell)$. By monotonicity of κ_2 (cf. [Claim 3.8](#)), we obtain estimation for the relative truncation parameters $\hat{\gamma}_1$ and $\hat{\gamma}_2$, which allows us to compute an estimator for μ . This allows us to prove [Theorem 7.1](#). As before, we need to be careful if α_ℓ is close to 1 as then $\kappa_3(\gamma_\ell) \rightarrow 0$.

Proof of [Theorem 7.1](#). Recall from [Lemma 7.3](#) that, for $X \sim \mathcal{N}(\mu, I_d)|_{H_1 \cap H_2}$ we have

$$\mathbb{E}[X] = \mu + \kappa_1(\gamma_1) \cdot w_1 + \kappa_1(\gamma_2) \cdot w_2.$$

We thus get

$$\begin{aligned} \|\hat{\mu} - \mu\| &= \|\bar{x} - \kappa_1(\hat{\gamma}_1)\hat{w}_1 - \kappa_1(\hat{\gamma}_2)\hat{w}_2 - \mu\| \\ &= \|\bar{x} - \mathbb{E}[X] + \kappa_1(\hat{\gamma}_1)\hat{w}_1 + \kappa_1(\hat{\gamma}_2)\hat{w}_2 - \kappa_1(\gamma_1)w_1 - \kappa_1(\gamma_2)w_2\| \\ &\leq \|\bar{x} - \mathbb{E}[X]\| + \|\kappa_1(\hat{\gamma}_1)\hat{w}_1 - \kappa_1(\gamma_1)w_1\| + \|\kappa_1(\hat{\gamma}_2)\hat{w}_2 - \kappa_1(\gamma_2)w_2\|. \end{aligned}$$

We have that x_i is 1-sub-Gaussian and thus given $n = O(d/\beta^2)$ samples, as in [Lemma 5.14](#), we have $\|\bar{x} - \mathbb{E}[X]\| \leq \beta$. For the other two terms, we need to distinguish whether α_ℓ is close to 1 or not. Both terms are the same, so we only show how to bound $\|\kappa_1(\hat{\gamma}_1)\hat{w}_1 - \kappa_1(\gamma_1)w_1\|$. As before in [Theorem 5.1](#), we distinguish two cases: If $\alpha_1 \leq \alpha_* := \max\{1 - c, \frac{1-\beta}{\log(1/\beta)}\}$ for a sufficiently small constant $c > 0$, then we use [Lemma 7.5](#) to get an estimate $\|\hat{w}_\ell - w_\ell\| \leq \beta$ and use this to bound $\|\kappa_1(\hat{\gamma}_1)\hat{w}_1 - \kappa_1(\gamma_1)w_1\|$. If $\alpha_1 \geq \alpha_*$, then we show that this norm is small by arguing that both $\kappa_1(\hat{\gamma}_1)$ and $\kappa_1(\gamma)$ are small.

Case (i): $\alpha_1 \leq \alpha_*$. Using

$$n = O\left(\frac{d^2}{\beta^2 \kappa_3(\gamma_1)^2} + \frac{d^2}{\kappa_3(\gamma_1)^4}\right)$$

samples, the estimate \hat{w}_1 from [Lemma 7.5](#) satisfies $\|\hat{w}_1 - w_1\| \leq \beta$. Let $\widehat{M}_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ be the sample covariance. We have

$$\begin{aligned} \hat{w}_1^\top \widehat{M}_2 \hat{w}_1 &= w_1^\top \widehat{M}_2 w_1 + (\hat{w}_1 - w_1)^\top \widehat{M}_2 (\hat{w}_1 + w_1) \\ &= w_1^\top M_2 w_1 + w_1^\top (\widehat{M}_2 - M_2) w_1 + (\hat{w}_1 - w_1)^\top \widehat{M}_2 (\hat{w}_1 + w_1) \\ &= \kappa_2(\gamma_1) + w_1^\top (\widehat{M}_2 - M_2) w_1 + (\hat{w}_1 - w_1)^\top \widehat{M}_2 (\hat{w}_1 + w_1). \end{aligned}$$

By [Lemma 7.5](#), we have

$$\left\| (\hat{w}_1 - w_1)^\top \widehat{M}_2 (\hat{w}_1 + w_1) \right\| \leq 2 \left\| \widehat{M}_2 \right\| \beta.$$

Furthermore, analogous to [Claim 4.3](#), we get that the $M_2^{-1}x_i$ are $O(1/\min\{\sqrt{\kappa_2(\gamma_1)}, \sqrt{\kappa_2(\gamma_2)}\})$ -sub-Gaussian. Thus, using

$$n = O\left(\frac{d}{\beta^2 \cdot \min\{\kappa_2(\gamma_1), \kappa_2(\gamma_2)\}}\right)$$

samples, we get by [Fact 3.26](#) that $\|M_2 - \widehat{M}_2\| \leq \beta$. Since $M_2 \leq I_d$, this also implies that $\|\widehat{M}_2\| \leq 2$, as long as $\beta \leq 1$. Putting these together, we get that

$$\left|\widehat{w}_1^\top \widehat{M}_2 \widehat{w}_1 - \kappa_2(\gamma_1)\right| \leq O(\beta).$$

Define $\hat{\gamma}_1 = \kappa_2^{-1}(\widehat{w}_1^\top \widehat{M}_2 \widehat{w}_1)$ (this inverse function exists since κ_2 is monotonic by [Claim 3.8](#)). By [Claim 3.21](#), we have

$$|\kappa_1(\hat{\gamma}_1) - \kappa_1(\gamma_1)| \leq |\kappa_2(\hat{\gamma}_1) - \kappa_2(\gamma_1)| \cdot \max\{\kappa_2(\gamma_1)^{-2}, \kappa_2(\hat{\gamma}_1)^{-2}\}.$$

Now assume $\beta < c' \cdot \kappa_2(\gamma_1)$ (for a sufficiently small constant c'). Then, we get $\kappa_2(\hat{\gamma}_1) \geq \frac{1}{2}\kappa_2(\gamma_1)$ and thus

$$|\kappa_1(\hat{\gamma}_1) - \kappa_1(\gamma_1)| \leq \beta \cdot \kappa_2(\gamma_1)^{-2}.$$

We thus have that

$$\begin{aligned} \|\kappa_1(\hat{\gamma}_1)\widehat{w}_1 - \kappa_1(\gamma_1)w_1\| &\leq \|\kappa_1(\hat{\gamma}_1)\widehat{w}_1 - \kappa_1(\gamma_1)\widehat{w}_1\| + \|\kappa_1(\gamma_1)\widehat{w}_1 - \kappa_1(\gamma_1)w_1\| \\ &\leq O\left(\frac{\beta}{\kappa_2(\gamma_1)^2}\right) + |\kappa_1(\gamma_1)| \cdot \|\widehat{w}_1 - w_1\| \\ &\leq \frac{1}{\kappa_2(\gamma_1)^2} \cdot O(\beta), \end{aligned}$$

where we used [Claim 3.9](#) for $|\kappa_1(\gamma_1)| \lesssim \kappa_2(\gamma_1)^{-1/2}$. This holds for all $\beta < c' \cdot \kappa_2(\gamma_1)$ and the sample complexity needed is

$$n = O\left(\frac{d^2}{\beta^2 \kappa_3(\gamma_1)^2} + \frac{d^2}{\kappa_3(\gamma_1)^4} + \frac{d}{\beta^2 \cdot \min\{\kappa_2(\gamma_1), \kappa_2(\gamma_2)\}}\right).$$

Case (ii): $\alpha_1 \geq \alpha_*$. In this case, we want to argue that both $|\kappa_1(\gamma_1)|$ and $|\kappa_1(\hat{\gamma}_1)|$ are small. Note that by picking c sufficiently small, we can achieve $\gamma_1 \geq 2$ and get

$$\begin{aligned} |\kappa_1(\gamma_1)| &\lesssim \varphi(\gamma_1) && \text{since } \Phi(\gamma_1) \geq \Phi(2) \\ &\leq \left(\gamma_1 + \frac{1}{\gamma_1}\right) \cdot (1 - \Phi(\gamma_1)) && \text{by Fact 3.3} \\ &\leq 2\gamma_1 \cdot (1 - \Phi(\gamma_1)) && \text{since } \gamma_1 \geq 1 \\ &\leq 2\sqrt{2} \cdot (1 - \Phi(\gamma_1)) \cdot \log^{1/2}\left(\frac{1}{1 - \Phi(\gamma_1)}\right) && \text{by Claim 3.15} \\ &\lesssim \beta && \text{since } \alpha_1 = \Phi(\gamma_1) \geq \frac{1 - \beta}{\log(1/\beta)}. \end{aligned}$$

Similarly, using [Claim 3.14](#), we get that $1 - \kappa_2(\gamma_1) \lesssim \beta$. It remains to argue that $|\kappa_1(\hat{\gamma}_1)|$ is also at most $O(\beta)$. By [Claim 3.11](#), it suffices to show that $\kappa_2(\hat{\gamma}_1) \geq 1 - \Omega(\beta)$, as long as β is at most a sufficiently small constant (such that $\kappa_2(\hat{\gamma}) \geq \kappa_2(1)$, which implies $\hat{\gamma}_1 \geq 1$). Recall that

$$M_2 = I_d - (1 - \kappa_2(\gamma_1)) \cdot w_1 w_1^\top - (1 - \kappa_2(\gamma_1)) \cdot w_2 w_2^\top.$$

Thus, we have

$$\hat{w}_1^\top M_2 \hat{w}_1 = \kappa_2(\gamma_1) \cdot \langle w_1, \hat{w}_1 \rangle^2 + \kappa_2(\gamma_2) \cdot \langle w_2, \hat{w}_1 \rangle^2 + \left(1 - \langle w_1, \hat{w}_1 \rangle^2 - \langle w_2, \hat{w}_1 \rangle^2\right).$$

Now, we have $\kappa_2(\gamma_1) \geq 1 - \Omega(\beta)$. If $\alpha_2 \geq \alpha_*$, then we also have $\kappa_2(\gamma_2) \geq 1 - \Omega(\beta)$ and thus $\hat{w}_1^\top M_2 \hat{w}_1 \geq 1 - \Omega(\beta)$. If $\alpha_2 \leq \alpha_*$, then we have $\|\hat{w}_2 - w_2\| \leq \beta$ and since \hat{w}_1 and \hat{w}_2 are orthogonal, we get

$$|\hat{w}_1^\top w_2| = |\hat{w}_1^\top (w_2 - \hat{w}_2)| \leq \beta.$$

Thus, we have

$$\hat{w}_1^\top M_2 \hat{w}_1 \geq (1 - \Omega(\beta)) \langle w_1, \hat{w}_1 \rangle^2 + (1 - \langle w_1, \hat{w}_1 \rangle^2 - \beta^2) \geq 1 - \Omega(\beta)$$

in this case as well (assuming $\beta \leq 1$). As before, using

$$n = O\left(\frac{d}{\beta^2 \cdot \min\{\kappa_2(\gamma_1), \kappa_2(\gamma_2)\}}\right)$$

samples, we get $\|M_2 - \widehat{M}_2\| \leq \beta$ and thus also $\kappa_2(\hat{\gamma}_1) = \hat{w}_1^\top \widehat{M}_2 \hat{w}_1 \geq 1 - \Omega(\beta)$. Hence, we have $|\kappa_1(\hat{\gamma}_1)| \leq O(\beta)$ by [Claim 3.11](#). Together, this shows that for β at most a sufficiently small constant, we get that

$$\|\kappa_1(\hat{\gamma}_1) \hat{w}_1 - \kappa_1(\gamma_1) w_1\| \leq O(\beta).$$

Sample complexity. Summarized, we showed that for $\beta < \min\{c' \cdot \kappa_2(\gamma_1), c' \cdot \kappa_2(\gamma_2), c''\}$ (for a sufficiently small constant c'') we have

$$\|\hat{\mu} - \mu\| \leq \left(\frac{1}{\kappa_2(\gamma_1)^2} + \frac{1}{\kappa_2(\gamma_2)^2}\right) \cdot O(\beta).$$

The sample complexity needed for this consists of different terms. We always have

$$O\left(\frac{d}{\beta^2 \cdot \min\{\kappa_2(\gamma_1), \kappa_2(\gamma_2)\}}\right)$$

and for both $\ell \in \{1, 2\}$ we need to add

$$O\left(\frac{d^2}{\beta^2 \kappa_3(\gamma_\ell)^2} + \frac{d^2}{\kappa_3(\gamma_\ell)^4}\right)$$

if $\alpha_\ell \leq \alpha_*$. It remains to bound $\kappa_i(\gamma_\ell)$. By [Claim 3.22](#), we have

$$\kappa_2(\gamma_\ell) \geq \Omega(\min\{1, \log^{-1}(1/\alpha_\ell)\}) \geq \Omega(\min\{1, \log^{-1}(1/\alpha)\}).$$

Similarly, if $\alpha_\ell \leq \alpha_*$, then we can bound $\kappa_3(\gamma_\ell)$ using [Claim 3.23](#) as follows

$$\kappa_3(\gamma_\ell) \geq \Omega(\min\{\log^{-3/2}(1/\alpha), \beta\}).$$

Putting these together, the sample complexity becomes

$$n = O\left(\frac{d^2}{\beta^2} \text{poly log}(1/\alpha) + \frac{d^2}{\beta^4}\right).$$

Note that we can again assume with loss of generality that $\varepsilon \leq 1$ (otherwise the conclusion is trivial). Thus, picking (for a sufficiently small constant c''')

$$\beta = \min\{c' \cdot \kappa_2(\gamma_1), c' \cdot \kappa_2(\gamma_2), c'', c''' \cdot \min\{\kappa_2(\gamma_1), \kappa_2(\gamma_2)\} \cdot \varepsilon\}$$

proves the theorem.¹⁶ Note that the runtime analysis of [Algorithm 3](#) is exactly the same as the one for [Algorithm 2](#) in the proof of [Theorem 5.1](#). \square

¹⁶Note that we do not need a $\text{poly log}(1/\alpha)$ for the d^2/ε^4 term. The reason is the following: We only need this term if α_ℓ is close to 1 (say larger than some constant). But in this case $\kappa_2(\gamma_\ell)$ is a constant and thus $O(d^2/\varepsilon^4)$ is sufficient.

Acknowledgments

This research was supported by the Swiss National Science Foundation (SNSF), grant no. 10004947. We thank the anonymous reviewers for their helpful comments.

A Deferred proofs from the main text

In this appendix we provide proofs of some claims that were deferred from the main text.

A.1 Cumulants of the one-dimensional truncated distribution

In this section we prove several facts from [Section 3.2](#) on the first three cumulants of the one-dimensional random variable $Z_\gamma \sim \mathcal{N}(0, 1)|_{(-\infty, \gamma]}$. We first prove [Fact 3.2](#) on the values of the cumulants that we restate below.

Fact (Restatement of [Fact 3.2](#)). *Let $\gamma \in \mathbb{R}$. We have that*

$$\begin{aligned}\kappa_1(\gamma) &= \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t\varphi(t)dt = -\frac{\varphi(\gamma)}{\Phi(\gamma)}, \\ \kappa_2(\gamma) &= 1 - \gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} - \left(\frac{\varphi(\gamma)}{\Phi(\gamma)}\right)^2, \\ \kappa_3(\gamma) &= \frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \left(1 - \gamma^2 - 3\gamma \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)} - 2 \cdot \left(\frac{\varphi(\gamma)}{\Phi(\gamma)}\right)^2\right).\end{aligned}$$

Proof. Recall that the first three cumulants can be written in terms of the moments of Z_γ . In particular, recall from [\(3.1\)](#) that

$$\begin{aligned}\kappa_1(\gamma) &= \mathbb{E}[Z_\gamma], \\ \kappa_2(\gamma) &= \mathbb{E}\left[(Z_\gamma - \mathbb{E}Z_\gamma)^2\right], \\ \kappa_3(\gamma) &= \mathbb{E}\left[(Z_\gamma - \mathbb{E}Z_\gamma)^3\right].\end{aligned}$$

Based on this, we can now write κ_i in terms of $\varphi(\gamma), \Phi(\gamma)$. For κ_1 , observe that

$$\kappa_1(\gamma) = \mathbb{E}[Z_\gamma] = \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t\varphi(t)dt = -\frac{\varphi(\gamma)}{\Phi(\gamma)}, \tag{A.1}$$

where for the last equality we used $\frac{d}{dt}\varphi(t) = -t\varphi(t)$. For κ_2 , we first compute the second moment of Z_γ as follows:

$$\begin{aligned}\mathbb{E}[Z_\gamma^2] &= \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t^2\varphi(t)dt \\ &= \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t \cdot \frac{d}{dt}(-\varphi(t))dt \\ &\stackrel{(i)}{=} \frac{1}{\Phi(\gamma)} \left([-t\varphi(t)]_{-\infty}^{\gamma} + \int_{-\infty}^{\gamma} \varphi(t)dt \right)\end{aligned}$$

$$= -\gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} + 1$$

where in equality (i) we used integration by parts. Combined with (A.1) this implies that

$$\kappa_2(\gamma) = \mathbb{E} [Z_\gamma^2] - (\mathbb{E} [Z_\gamma])^2 = 1 - \gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} - \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2. \quad (\text{A.2})$$

Similarly, we can compute the third moment

$$\begin{aligned} \mathbb{E} [Z_\gamma^3] &= \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t^3 \varphi(t) dt \\ &= \frac{1}{\Phi(\gamma)} \cdot \int_{-\infty}^{\gamma} t^2 \cdot \frac{d}{dt} (-\varphi(t)) dt \\ &\stackrel{(i)}{=} \frac{1}{\Phi(\gamma)} \left([-t^2 \varphi(t)]_{-\infty}^{\gamma} + 2 \int_{-\infty}^{\gamma} t \varphi(t) dt \right) \\ &\stackrel{(ii)}{=} -\gamma^2 \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)} - 2 \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)}, \end{aligned} \quad (\text{A.3})$$

where in (i) we again used integration by parts and in (ii) we use (A.1). Thus, we can compute κ_3 as follows

$$\begin{aligned} \kappa_3(\gamma) &= \mathbb{E} \left[(Z_\gamma - \mathbb{E} [Z_\gamma])^3 \right] \\ &= \mathbb{E} [Z_\gamma^3] - 3 \mathbb{E} [Z_\gamma^2] \cdot \mathbb{E} [Z_\gamma] + 3 \mathbb{E} [Z_\gamma] \cdot (\mathbb{E} [Z_\gamma])^2 - (\mathbb{E} [Z_\gamma])^3 \\ &= \mathbb{E} [Z_\gamma^3] - 3 \mathbb{E} [Z_\gamma] \cdot \left(\mathbb{E} [Z_\gamma^2] - (\mathbb{E} [Z_\gamma])^2 \right) - (\mathbb{E} [Z_\gamma])^3 \\ &= -\gamma^2 \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)} - 2 \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)} - 3 \cdot \left(-\frac{\varphi(\gamma)}{\Phi(\gamma)} \right) \cdot \left(1 - \gamma \frac{\varphi(\gamma)}{\Phi(\gamma)} - \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2 \right) - \left(-\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^3, \end{aligned}$$

where in the last step we used (A.1), (A.2) and (A.3). After simplification we have that

$$\kappa_3(\gamma) = \frac{\varphi(\gamma)}{\Phi(\gamma)} \cdot \left(1 - \gamma^2 - 3\gamma \cdot \frac{\varphi(\gamma)}{\Phi(\gamma)} - 2 \cdot \left(\frac{\varphi(\gamma)}{\Phi(\gamma)} \right)^2 \right),$$

which completes the proof. \square

For several of the following proofs, we use the approximation of the inverse Mills ratio $h(t) = \frac{\varphi(t)}{1-\Phi(t)}$ from Fact 3.3. We will in particular need the following bounds from Fact 3.3 that hold for all $x > 0$:

$$\begin{aligned} h(x) &> r_1(x) = x && (k = 0) \\ h(x) &< r_2(x) = \frac{x^2 + 1}{x} && (k = 1) \\ h(x) &> r_3(x) = \frac{x(x^2 + 3)}{x^2 + 2} && (k = 2) \\ h(x) &< r_4(x) = \frac{x^4 + 6x^2 + 3}{x^3 + 5x} && (k = 3) \end{aligned}$$

$$h(x) > r_5(x) = \frac{x(x^4 + 10x^2 + 15)}{x^4 + 9x^2 + 8} \quad (k = 4).$$

Using these, we now prove the following two claims about κ_1 and κ_2 .

Claim (Restatement of [Claim 3.6](#)). *The function κ_1 as defined in (3.1) is 1-Lipschitz on \mathbb{R} .*

Claim (Restatement of [Claim 3.7](#)). *For every $t \in \mathbb{R}$, we have that $0 < \kappa_2(t) \leq 1$.*

We first prove [Claim 3.7](#) and then use it to prove [Claim 3.6](#).

Proof of Claim 3.7. First, note that $\kappa_2(\gamma) = \mathbb{E}[(Z_\gamma - \mathbb{E}[Z_\gamma])^2] > 0$. To see $\kappa_2(t) \leq 1$, recall from [\(A.2\)](#) that

$$1 - \kappa_2(t) = t \frac{\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)} \right)^2.$$

For $t \geq 0$, both terms are non-negative and thus $1 - \kappa_2(t) \geq 0$. For $t < 0$, we get from [Fact 3.3](#) that $\frac{\varphi(t)}{\Phi(t)} = \frac{\varphi(-t)}{1 - \Phi(-t)} \geq -t$ and thus

$$1 - \kappa_2(t) = \frac{\varphi(t)}{\Phi(t)} \left(t + \frac{\varphi(t)}{\Phi(t)} \right) \geq \frac{\varphi(t)}{\Phi(t)} (t - t) = 0$$

in this case as well. □

Proof of Claim 3.6. Observe that, since $\kappa_1(t) = -\frac{\varphi(t)}{\Phi(t)}$ by [Fact 3.2](#), we get that

$$\frac{d}{dt} \kappa_1(t) = t \frac{\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 = 1 - \kappa_2(t) \leq 1,$$

where we used again [Fact 3.2](#) and that $\kappa_2(t) \geq 0$ by [Claim 3.7](#). □

Next, we prove the following claim that shows that the function κ_2 is monotonically increasing.

Claim (Restatement of [Claim 3.8](#)). *For κ_2 and κ_3 as in (3.1), we have $\frac{d}{dt} \kappa_2(t) = -\kappa_3(t)$. Moreover, $\kappa_3(t) < 0$ for every $t \in \mathbb{R}$ and thus $\kappa_2(t)$ is monotonically increasing.*

Proof of Claim 3.8. Recall from [Fact 3.2](#) that $\kappa_2(t) = 1 - t \frac{\varphi(t)}{\Phi(t)} - \left(\frac{\varphi(t)}{\Phi(t)} \right)^2$. Hence, we get that

$$\begin{aligned} \frac{d}{dt} \kappa_2(t) &= -\frac{\varphi(t)}{\Phi(t)} - t \frac{d}{dt} \frac{\varphi(t)}{\Phi(t)} - 2 \left(\frac{\varphi(t)}{\Phi(t)} \right) \cdot \frac{d}{dt} \frac{\varphi(t)}{\Phi(t)} \\ &= -\frac{\varphi(t)}{\Phi(t)} + \left(t + 2 \frac{\varphi(t)}{\Phi(t)} \right) \cdot \frac{d}{dt} \frac{-\varphi(t)}{\Phi(t)} \\ &= -\frac{\varphi(t)}{\Phi(t)} + \left(t + 2 \frac{\varphi(t)}{\Phi(t)} \right) \left(t \frac{\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 \right) \\ &= (t^2 - 1) \frac{\varphi(t)}{\Phi(t)} + 3t \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 + 2 \left(\frac{\varphi(t)}{\Phi(t)} \right)^3 = -\kappa_3(t) \end{aligned}$$

by [Fact 3.2](#). Thus, $\frac{d}{dt} \kappa_2(t) > 0$ if and only if $f(t) = t^2 - 1 + 3t \cdot \frac{\varphi(t)}{\Phi(t)} + 2 \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 > 0$. We show the latter by distinguishing the three cases (i) $1 \leq t$, (ii) $0 \leq t < 1$ and (iii) $t < 0$.

Case (i): $1 \leq t$. In this case, $t^2 - 1 \geq 0$ and the other two terms are positive. Thus, if $t \geq 1$, then $f(t) > 0$ holds.

Case (ii): $0 \leq t < 1$. A direct computation shows $f(0) = \frac{4}{\pi} - 1 > 0$. For $t \in (0, 1)$, since $\varphi(t) = \frac{d}{dt}\Phi(t)$ is decreasing, we can upper bound $\Phi(t)$ by its tangent line at $t = 0$, i.e.,

$$\Phi(t) \leq \frac{1}{2} + \frac{t}{\sqrt{2\pi}}.$$

Therefore, we obtain

$$\frac{\varphi(t)}{\Phi(t)} \geq \frac{\exp(-t^2/2)}{\sqrt{\pi/2} + t} > 0$$

and thus

$$f(t) \geq t^2 - 1 + \frac{\exp(-t^2/2)}{\sqrt{\pi/2} + t} \cdot \left(3t + 2 \frac{\exp(-t^2/2)}{\sqrt{\pi/2} + t}\right).$$

We show in [Fact B.5](#) in [Appendix B](#) that this lower bound is monotonic in $(0, 1)$. Thus, the lower bound attains its minimum at $t = 0$ and we have for any $t \in (0, 1)$ that $f(t) \geq \frac{4}{\pi} - 1 > 0$.

Case (iii): $t < 0$. Finally, for $t < 0$ consider the quadratic function

$$q_t(x) = t^2 - 1 + 3tx + 2x^2.$$

We have $\frac{d}{dx}q_t(x) = 3t + 4x$ and thus $q_t(x)$ is increasing in x for $x > -\frac{3}{4}t$. Note that we have

$$f(t) = q_t\left(\frac{\varphi(t)}{\Phi(t)}\right) = q_t(h(-t))$$

for $h(-t)$ the inverse Mills ratio as in [Section 3.2](#) (note that $-t > 0$). By [Fact 3.4](#), we have

$$h(-t) > \frac{1}{4} \left(\sqrt{t^2 + 8} - 3t\right) > -\frac{3}{4}t$$

and hence, we get that

$$\begin{aligned} q_t(h(-t)) &> q_t\left(\frac{1}{4} \left(\sqrt{t^2 + 8} - 3t\right)\right) \\ &= t^2 - 1 + 3t \cdot \frac{1}{4} \left(\sqrt{t^2 + 8} - 3t\right) + 2 \left(\frac{1}{4} \left(\sqrt{t^2 + 8} - 3t\right)\right)^2 \\ &= t^2 - 1 + \frac{3}{4}t\sqrt{t^2 + 8} - \frac{9}{4}t^2 + \frac{1}{8} \left(t^2 + 8 - 6t\sqrt{t^2 + 8} + 9t^2\right) = 0, \end{aligned}$$

which shows that $f(t) > 0$ also holds for $t < 0$. This completes the proof. \square

Next, we prove the following claim that gives a uniform upper bound on $\kappa_1(t)\sqrt{\kappa_2(t)}$.

Claim (Restatement of [Claim 3.9](#)). *For every $t \in \mathbb{R}$, we have $|\kappa_1(t)| \cdot \sqrt{\kappa_2(t)} \leq O(1)$.*

Proof. Recall from [Fact 3.2](#) that $\kappa_1(t) = -\frac{\varphi(t)}{\Phi(t)}$ and $\kappa_2(t) = 1 - t \frac{\varphi(t)}{\Phi(t)} - \left(\frac{\varphi(t)}{\Phi(t)}\right)^2$. By [Claim 3.7](#), we also have that $0 < \kappa_2(t) \leq 1$. To show that $|\kappa_1(t)| \cdot \sqrt{\kappa_2(t)}$ is uniformly bounded, we distinguish the two cases (i) $t \geq -4$ and (ii) $t < -4$.

Case (i): $t \geq -4$. In this case we have $\Phi(t) \geq \Phi(-4) > 0$. Thus, we have, using that $\kappa_2(t) \leq 1$ for all t ,

$$|\kappa_1(t)| \cdot \sqrt{\kappa_2(t)} = \frac{\varphi(t)}{\Phi(t)} \cdot \sqrt{\kappa_2(t)} \leq \frac{\varphi(t)}{\Phi(-4)} \leq \frac{1}{\sqrt{2\pi}\Phi(-4)}.$$

Case (ii): $t < -4$. For this case, we want to use [Fact 3.3](#) to bound $\kappa_1(t)$ and $\kappa_2(t)$. First, for κ_1 we have $|\kappa_1(t)| = h(-t) \leq r_2(-t) = -t - 1/t \leq 2|t|$ for $t < -4 < -1$, where, as before, $h(-t)$ is the inverse Mills ratio as in [Section 3.2](#).

For κ_2 , consider the quadratic function $q_t(x) = -x^2 - tx + 1$. For $t < -4 < 0$, we can write

$$\kappa_2(t) = 1 - th(-t) - (h(-t))^2 = q_t(h(-t)).$$

We have $\frac{d}{dx}q_t(x) = -2x - t$ and hence $q_t(x)$ is increasing in x on the interval $(-\frac{t}{2}, \infty)$. Now, using [Fact 3.3](#), we get

$$h(-t) \geq r_5(-t) = \frac{-t(15 + 10t^2 + t^4)}{8 + 9t^2 + t^4} \geq -\frac{t}{2},$$

where we used $-t > 0$ and $t^2 - 1 \geq 0$ for the last step. Hence, we have $q_t(h(-t)) \leq q_t(r_5(-t))$ and thus

$$\kappa_2(t) = q_t(h(-t)) \leq q_t(r_5(-t)) = \frac{t^6 + 12t^4 + 39t^2 + 64}{(t^4 + 9t^2 + 8)^2} \leq \frac{4t^6}{t^8} = \frac{4}{t^2},$$

where we used $t < -4$ to get $12t^4 + 39t^2 + 64 \leq 3t^6$.

Finally, combining the bounds on $\kappa_1(t)$ and $\kappa_2(t)$, we get that, for all $t < -4$,

$$|\kappa_1(t)| \cdot \sqrt{\kappa_2(t)} \leq 2|t| \cdot \frac{2}{|t|} = 4,$$

which completes the proof. □

The following claim shows both $\kappa_3(t)$ and $\kappa_1(t) \cdot \kappa_3(t)$ are also uniformly bounded.

Claim (Restatement of [Claim 3.10](#)). *For every $t \in \mathbb{R}$, we have $|\frac{d}{dt}\kappa_2(t)| = |\kappa_3(t)| \leq O(1)$ and $0 < \kappa_1(t) \cdot \kappa_3(t) \leq O(1)$.*

Proof. We first show that $\kappa_3(t)$ is bounded on \mathbb{R} . By [Claim 3.7](#) we have $\kappa_3(t) < 0$ and thus, using [Fact 3.2](#), we get

$$|\kappa_3(t)| = \frac{\varphi(t)}{\Phi(t)} \cdot \left(2 \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 + 3t \frac{\varphi(t)}{\Phi(t)} + t^2 - 1 \right).$$

We distinguish the two cases (i) $t > -1$ and (ii) $t \leq -1$.

Case (i): $t > -1$. Note that $\varphi(t)$, $t\varphi(t)$ and $t^2\varphi(t)$ are all uniformly bounded over $t \in \mathbb{R}$. For $t > -1$ we furthermore have $\frac{1}{\Phi(t)} \leq \frac{1}{\Phi(-1)} \leq O(1)$. Putting these together, we get that for $t > -1$ we have $|\kappa_3(t)| \leq O(1)$.

Case (ii): $t \leq -1$. For this case, we want to again use [Fact 3.3](#). For $t < -1 < 0$, we again use that $\frac{\varphi(t)}{\Phi(t)} = h(-t)$. By [Fact 3.3](#), we get that

$$\frac{-t \cdot (t^2 + 3)}{t^2 + 2} = r_3(-t) < h(-t) = \frac{\varphi(t)}{\Phi(t)} < r_2(-t) = -t - \frac{1}{t}.$$

We thus get that

$$|\kappa_3(t)| = h(-t) \cdot (2h(-t)^2 + 3t \cdot h(-t) + t^2 - 1)$$

$$\begin{aligned}
&\leq \left(-t - \frac{1}{t}\right) \cdot \left(2\left(-t - \frac{1}{t}\right)^2 + 3t \cdot \frac{-t \cdot (t^2 + 3)}{t^2 + 2} + t^2 - 1\right) \\
&= \left(-t - \frac{1}{t}\right) \cdot \frac{2(-t^2 - 1)^2(t^2 + 2) - 3t^4(t^2 + 3) + t^2(t^2 - 1)(t^2 + 2)}{t^2(t^2 + 2)} \\
&= \left(-t - \frac{1}{t}\right) \cdot 4 \cdot \frac{2t^2 + 1}{t^2(t^2 + 2)} \\
&\leq -2t \cdot \frac{8}{t^2} \leq 16,
\end{aligned}$$

where we used $-t - 1/t \leq -2t$ and $2t^2 + 1 \leq 2(t^2 + 2)$ for $t \leq -1$.

Hence, $|\kappa_3(t)|$ is uniformly bounded over $t \in \mathbb{R}$. For $\kappa_1(t) \cdot \kappa_3(t)$, first note that by [Fact 3.2](#) we have $\kappa_1(t) = -\frac{\varphi(t)}{\Phi(t)} < 0$ and thus $\kappa_1(t) \cdot \kappa_3(t) > 0$. To show a uniform upper bound, we distinguish the same two cases as above and use a similar strategy. Note that we can use the same bounds and multiply them by $\frac{\varphi(t)}{\Phi(t)}$. If $t > -1$, we have

$$\frac{\varphi(t)}{\Phi(t)} \leq \frac{\varphi(0)}{\Phi(-1)} \leq O(1),$$

and thus $\kappa_1(t) \cdot \kappa_3(t) \leq O(1)$ for these t . If $t \leq -1$, again by [Fact 3.3](#), we get

$$\frac{\varphi(t)}{\Phi(t)} = h(-t) \leq -t - \frac{1}{t} \leq -2t \leq 2$$

and thus $\kappa_1(t) \cdot \kappa_3(t) \leq 32$ for these t . Hence, also $\kappa_1(t) \cdot \kappa_3(t)$ is uniformly bounded, which completes the proof. \square

We now prove how to lower bound κ_1 in terms of $1 - \kappa_2$.

Claim (Restatement of [Claim 3.11](#)). *For $t \geq 1$, we have $|\kappa_1(t)| \leq 1 - \kappa_2(t)$.*

Proof. Using [Fact 3.2](#), we have

$$|\kappa_1(t)| = \frac{\varphi(t)}{\Phi(t)} \leq t \cdot \frac{\varphi(t)}{\Phi(t)} \leq t \cdot \frac{\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)}\right)^2 = 1 - \kappa_2(t),$$

which completes the proof. \square

Next, we prove the following lower bound on κ_2 .

Claim (Restatement of [Claim 3.12](#)). *For $t \leq -3$, we have $\kappa_2(t) \geq \frac{1}{4t^2}$.*

Proof. We use an analogous strategy as in the second part of the proof of [Claim 3.9](#), where we showed an upper bound on $\kappa_2(t)$ for $t < -4$. Using [Fact 3.3](#) for $-t$, we get that, for $t \leq -3 < 0$,

$$-t = r_1(-t) < h(-t) < r_4(-t) = -\frac{t^4 + 6t^2 + 3}{t(t^2 + 5)}.$$

Now, consider again the quadratic function $q_t(x) = -x^2 - tx + 1$, which decreases on the interval $(-t/2, \infty)$. As before, we have $\kappa_2(t) = q_t(h(-t))$ and thus we get that

$$\kappa_2(t) = q_t(h(-t)) \geq q_t(r_4(-t)) = -\left(-\frac{t^4 + 6t^2 + 3}{t(t^2 + 5)}\right)^2 - t \left(-\frac{t^4 + 6t^2 + 3}{t(t^2 + 5)}\right) + 1.$$

Simplifying this expression, we can conclude that

$$\kappa_2(t) \geq \frac{t^4 + 4t^2 - 9}{t^6 + 10t^4 + 25t^2} \geq \frac{t^4}{4t^6} = \frac{1}{4t^2},$$

where we used $4t^2 - 9 \geq 0$ and $10t^4 + 25t^2 \leq 3t^6$ for $t \leq -3$. □

Similarly, we prove the following lower bound on κ_3 .

Claim (Restatement of [Claim 3.13](#)). *For $t \leq -2$, we have $|\kappa_3(t)| \geq \frac{1}{50|t|^3}$.*

Proof. We refine the argument from [Claim 3.8](#), where we showed that $\kappa_3(t) > 0$. [Claim 3.8](#), together with [Fact 3.2](#), shows that

$$|\kappa_3(t)| = -\kappa_3(t) = h(-t) \left(t^2 - 1 + 3th(-t) + 2(h(-t))^2 \right),$$

where $h(-t)$ is again the inverse Mills ratio. From [Fact 3.3](#) we have

$$h(-t) > r_5(-t) = -\frac{t(t^4 + 10t^2 + 15)}{t^4 + 9t^2 + 8} \geq -t.$$

Consider the quadratic function $q_t(x) = 2x^2 + 3tx + t^2 - 1$. We have $|\kappa_3(t)| = h(-t)q_t(h(-t))$. As already argued in the proof of [Claim 3.8](#), this function is increasing in x for $x > -\frac{3}{4}t$. Thus, we get that

$$q_t(h(-t)) \geq q_t(r_5(-t)) = \frac{2t^4 + 10t^2 - 64}{(t^4 + 9t^2 + 8)^2} \geq \frac{1}{50t^4},$$

where we used $10t^2 - 64 \geq -\frac{3}{2}t^4$ and $9t^2 + 8 \leq 4t^4$ for $t \leq -2$. Finally, again by [Fact 3.3](#), we have that $h(-t) \geq -t$ and thus

$$|\kappa_3(t)| = h(-t)q_t(h(-t)) \geq -t \cdot \frac{1}{50t^4} = \frac{1}{50|t|^3}$$

since $t < 0$ and thus $-t^3 = |t|^3$. □

Next, we prove the following lower bound on $\kappa_2(t)$.

Claim (Restatement of [Claim 3.14](#)). *For $t \geq 1$, we have $1 - \kappa_2(t) \leq 16 \cdot (1 - \Phi(t)) \cdot \log 1/(1 - \Phi(t))$.*

Proof. Note that for $t \geq 1$, we have $\Phi(t) \geq \frac{1}{2}$ as well as $2\varphi(t) \leq \sqrt{\frac{2}{\pi}} \leq t$. It follows that, using [Fact 3.2](#),

$$1 - \kappa_2(t) = t \frac{\varphi(t)}{\Phi(t)} + \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 = \frac{\varphi(t)}{\Phi(t)} \left(t + \frac{\varphi(t)}{\Phi(t)} \right) \leq 2\varphi(t) \cdot (t + 2\varphi(t)) \leq 4t \cdot \varphi(t).$$

Applying [Fact 3.3](#), we have $\frac{\varphi(t)}{1 - \Phi(t)} \leq t + \frac{1}{t} \leq 2t$ and thus $\varphi(t) \leq 2t(1 - \Phi(t))$. We get that

$$1 - \kappa_2(t) \leq 8t^2 \cdot (1 - \Phi(t)).$$

Applying [Fact 3.3](#) again, we have $\frac{\varphi(t)}{1 - \Phi(t)} \geq t$ and hence

$$\frac{1}{1 - \Phi(t)} \geq \frac{t}{\varphi(t)} \geq t\sqrt{2\pi} \cdot e^{t^2/2} \geq e^{t^2/2} \implies \log \left(\frac{1}{1 - \Phi(t)} \right) \geq t^2/2.$$

Combining this with the above, we get

$$1 - \kappa_2(t) \leq 16 \cdot (1 - \Phi(t)) \cdot \log\left(\frac{1}{1 - \Phi(t)}\right),$$

as claimed. □

We now prove the following claim bounding $\log(1/(1 - \Phi(t)))$.

Claim (Restatement of [Claim 3.15](#)). *For $t \geq 2$, we have $\frac{t^2}{2} \leq \log(1/(1 - \Phi(t))) \leq 2t^2$.*

Proof. Using [Fact 3.3](#) we have

$$\frac{1}{1 - \Phi(t)} \leq \frac{1}{\varphi(t)} \cdot \left(t + \frac{1}{t}\right) \leq \frac{1}{\varphi(t)} \cdot (t + 1) = \sqrt{2\pi}e^{t^2/2} \cdot (t + 1),$$

where we used $\frac{1}{t} \leq 1$ for $t \geq 2 \geq 1$. Thus, we get that

$$\begin{aligned} \log\left(\frac{1}{1 - \Phi(t)}\right) &\leq \frac{t^2}{2} + \log(t + 1) + \log(\sqrt{2\pi}) \\ &\leq \frac{t^2}{2} + t + \log(\sqrt{2\pi}) \\ &\leq 2t^2, \end{aligned}$$

where we used $t > 0$ for $\log(t + 1) \leq t$ and $t \geq 2$ for $t + \log(\sqrt{2\pi}) \leq \frac{3}{2}t^2$. The other direction was already proved in [Claim 3.14](#) for $t \geq 1$. □

Finally, we want to prove the following claim about the asymptotics of $\text{skew}(t)$.

Claim (Restatement of [Claim 3.16](#)). *There are absolute constants $m, M > 0$ such that*

$$m \leq \frac{|\text{skew}(t)|}{(1 - \Phi(t)) \log^{3/2}(1/(1 - \Phi(t)))} \leq M$$

for all $t \geq 0$.

In order to prove this claim, we first prove the following upper bound on $\kappa_3(t)$.

Claim A.1 (Upper bound on $\kappa_3(t)$). *For $t \geq 1$ and $\delta = 1 - \Phi(t)$, we have that*

$$|\kappa_3(t)| \leq 40\sqrt{2} \cdot \delta \cdot \log^{3/2}(1/\delta).$$

Proof of Claim A.1. For $t \geq 1$, we have as before (cf. [Fact 3.2](#) and [Claim 3.8](#))

$$|\kappa_3(t)| = \frac{\varphi(t)}{\Phi(t)} \left(t^2 - 1 + 3t \cdot \frac{\varphi(t)}{\Phi(t)} + 2 \left(\frac{\varphi(t)}{\Phi(t)} \right)^2 \right).$$

For $t \geq 1$, we have $\Phi(t) \geq \Phi(0) \geq \frac{1}{2}$ and thus

$$|\kappa_3(t)| \leq 2\varphi(t) (t^2 - 1 + 6t \cdot \varphi(t) + 8\varphi(t)^2)$$

$$\begin{aligned} &\leq 2\varphi(t) \left(t^2 - 1 + 6t \cdot \frac{1}{\sqrt{2\pi}} + 8 \cdot \frac{1}{2\pi} \right) \\ &\leq 10\varphi(t)t^2, \end{aligned}$$

where we used in the last step that $\frac{6}{\sqrt{2\pi}}t + \frac{8}{2\pi} \leq 4t^2$ for $t \geq 1$. Now, by [Fact 3.3](#), we have that $h(t) \leq t + 1/t \leq 2t$ for $t \geq 1$. Since $h(t) = \frac{\varphi(t)}{1-\Phi(t)} = \frac{\varphi(t)}{\delta}$, it follows that

$$|\kappa_3(t)| \leq 20 \cdot \delta \cdot t^3.$$

It remains to bound t in terms of δ . Again by [Fact 3.3](#), we have $h(t) \geq t \geq 1$ for $t \geq 1$ and thus

$$\delta \leq \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq e^{-\frac{t^2}{2}}$$

or equivalently

$$t \leq \sqrt{2 \log(1/\delta)}.$$

Combining this with the above we get that

$$|\kappa_3(t)| \leq 20 \cdot \delta \cdot (2 \log(1/\delta))^{3/2} = 40\sqrt{2} \cdot \delta \cdot \log^{3/2}(1/\delta),$$

which completes the proof. \square

Proof of [Claim 3.16](#). First, from [Claims 3.7](#) and [3.8](#) we know that $0 < \kappa_2(t) \leq 1$ and $\kappa_3(t) < 0$ for all $t \in \mathbb{R}$. Hence, we have in particular that

$$\text{skew}(t) = \frac{\kappa_3(t)}{\kappa_2(t)^{3/2}} < 0.$$

Define $\delta = 1 - \Phi(t)$. By [Claim 3.23](#), we have that, for $t \geq C$ for some constant C (such that the minimum in [Claim 3.23](#) is indeed the third term)

$$|\kappa_3(t)| \geq \frac{1}{4\sqrt{2}} \cdot \delta \cdot \log^{3/2}(1/\delta).$$

Using $\kappa_2(t) \leq 1$, we get that

$$\frac{|\text{skew}(t)|}{\delta \cdot \log^{3/2}(1/\delta)} \geq \frac{1}{4\sqrt{2}}$$

On the interval $t \in [0, C]$, the function $\frac{|\text{skew}(t)|}{\delta \cdot \log^{3/2}(1/\delta)}$ is continuous and strictly positive and thus there is $m' > 0$ such that

$$\frac{|\text{skew}(t)|}{\delta \cdot \log^{3/2}(1/\delta)} \geq m'$$

for $t \in [0, C]$. Defining $m = \min \left\{ m', \frac{1}{4\sqrt{2}} \right\}$ shows the lower bound. For the upper bound, we get by [Claim A.1](#) that

$$|\kappa_3(t)| \leq 40\sqrt{2} \cdot \delta \cdot \log^{3/2}(1/\delta)$$

for $t \geq 1$. Using [Claim 3.8](#), we get that $0 < \kappa_2(0) \leq \kappa_2(t)$ for all $t \geq 0$ and thus

$$\frac{|\text{skew}(t)|}{\delta \cdot \log^{3/2}(1/\delta)} \leq \frac{40\sqrt{2}}{\kappa_2(0)^{3/2}}$$

for $t \geq 1$. As before, the function $\frac{|\text{skew}(t)|}{\delta \cdot \log^{3/2}(1/\delta)}$ is continuous for $t \in [0, 1]$ and thus has a uniform upper bound M' . We get the claimed upper bound by defining $M = \max \left\{ M', \frac{40\sqrt{2}}{\kappa_2(0)^{3/2}} \right\}$. \square

A.2 Matrix and vector calculations

In this section, we prove several linear algebra facts that we deferred from [Sections 4](#) and [5](#). We first show that a linear transformation of a Gaussian truncated by a halfspace is again a (different) Gaussian truncated by a (different) halfspace.

Claim (Restatement of [Claim 4.1](#)). *Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix and let $b \in \mathbb{R}^d$. Consider the random variable $x \sim \mathcal{N}(\mu, \Sigma)|_{\langle w, x \rangle \leq \tau}$. Then, the random variable $y = Ax + b$ follows the distribution $\mathcal{N}(A\mu + b, A\Sigma A^\top)|_{\langle w', y \rangle \leq \tau'}$, where*

$$w' = \frac{A^{-\top}w}{\|A^{-\top}w\|} \quad \text{and} \quad \tau' = \frac{\tau + \langle A^{-\top}w, b \rangle}{\|A^{-\top}w\|}.$$

Proof. We prove this claim by computing the density of y explicitly. Note that the density of x is

$$f(x) \propto \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right) \mathbb{1}[w^\top x \leq \tau].$$

Hence, the density of y is

$$\begin{aligned} g(y) &= \frac{1}{|\det(A)|} f(A^{-1}(y - b)) \\ &\propto \exp\left(-\frac{1}{2}(A^{-1}(y - b) - \mu)^\top \Sigma^{-1}(A^{-1}(y - b) - \mu)\right) \mathbb{1}[w^\top (A^{-1}(y - b)) \leq \tau] \\ &= \exp\left(-\frac{1}{2}(y - b - A\mu)^\top (A^{-\top} \Sigma^{-1} A^{-1})(y - b - A\mu)\right) \mathbb{1}[(A^{-\top}w)^\top (y - b) \leq \tau] \\ &= \exp\left(-\frac{1}{2}(y - (A\mu + b))^\top (A\Sigma A^\top)^{-1}(y - (A\mu + b))\right) \mathbb{1}[(A^{-\top}w)^\top y \leq \tau + \langle A^{-\top}w, b \rangle] \end{aligned}$$

Thus, $g(y)$ is the density of the Gaussian $\mathcal{N}(A\mu + b, A\Sigma A^\top)$ truncated to a halfspace with normal $A^{-\top}w$. To conclude the lemma, we normalize the normal vector. Thus, the halfspace is

$$H' = \left\{ y : \left\langle \frac{A^{-\top}w}{\|A^{-\top}w\|}, y \right\rangle \leq \frac{\tau + \langle A^{-\top}w, b \rangle}{\|A^{-\top}w\|} \right\},$$

which completes the proof. \square

Next, we show that the Frobenius norm of $I_d - A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$ is invariant under the transformation $A \mapsto LAL^\top, B \mapsto LBL^\top$ for an invertible matrix L .

Fact (Restatement of [Fact 4.9](#)). *Let $A, B \in \mathbb{R}^{d \times d}$ be positive definite matrices. Let $L \in \mathbb{R}^{d \times d}$ be an invertible matrix. Then, we have that*

$$\left\| I_d - A^{-\frac{1}{2}}BA^{-\frac{1}{2}} \right\|_F = \left\| I_d - (LAL^\top)^{-\frac{1}{2}}LBL^\top(LAL^\top)^{-\frac{1}{2}} \right\|_F$$

Proof. Let $X = LA^{\frac{1}{2}}$ and $R = (XX^\top)^{-\frac{1}{2}}X$. Note that $LAL^\top = XX^\top$ and observe that $RR^\top = R^\top R = I_d$, i.e., R is orthonormal. The right-hand side of the identity can now be rewritten as

$$\left\| I_d - (LAL^\top)^{-\frac{1}{2}}LBL^\top(LAL^\top)^{-\frac{1}{2}} \right\|_F = \left\| I_d - (XX^\top)^{-\frac{1}{2}}XA^{-\frac{1}{2}}BA^{-\frac{1}{2}}X^\top(XX^\top)^{-\frac{1}{2}} \right\|_F$$

$$\begin{aligned}
&= \left\| I_d - RA^{-\frac{1}{2}}BA^{-\frac{1}{2}}R^\top \right\|_F \\
&= \left\| R \left(I_d - A^{-\frac{1}{2}}BA^{-\frac{1}{2}} \right) R^\top \right\|_F.
\end{aligned}$$

Let $M = I_d - A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$. Using that the trace is cyclic, we get that

$$\|RMR^\top\|_F^2 = \text{tr} \left(RMR^\top (RMR^\top)^\top \right) = \text{tr} (MM^\top) = \|M\|_F^2,$$

which completes the proof. \square

We now prove the following claim that shows that the best rank-1 approximation of a matrix A that is close to some matrix B is also a good approximation to B .

Claim (Restatement of [Claim 5.3](#)). *Let $A, B \in \mathbb{R}^{d \times d}$ be symmetric matrices with B being an arbitrary rank-1 matrix. Consider the spectral decomposition of A ,*

$$A = \sum_{i=1}^d \lambda_i v_i v_i^\top$$

where the vectors $\{v_i\}_{i=1}^d$ form an orthonormal basis of \mathbb{R}^d and the singular values $\{\lambda_i\}_{i=1}^d$ satisfy $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_d|$. We then have that

$$\|\lambda_1 v_1 v_1^\top - B\|_F^2 \leq 2 \langle \lambda_1 v_1 v_1^\top - B, A - B \rangle.$$

Proof. As $\lambda_1 v_1 v_1^\top$ is the best rank-1 approximation of A , we have $\|\lambda_1 v_1 v_1^\top - A\|_F^2 \leq \|A - B\|_F^2$. We thus get that

$$\begin{aligned}
\|\lambda_1 v_1 v_1^\top - B\|_F^2 &= \|(\lambda_1 v_1 v_1^\top - A) + (A - B)\|_F^2 \\
&= \|\lambda_1 v_1 v_1^\top - A\|_F^2 + 2 \langle \lambda_1 v_1 v_1^\top - A, A - B \rangle + \|A - B\|_F^2 \\
&\leq 2 \left(\|A - B\|_F^2 + \langle \lambda_1 v_1 v_1^\top - A, A - B \rangle \right) \\
&= 2 \langle \lambda_1 v_1 v_1^\top - B, A - B \rangle,
\end{aligned}$$

which completes the proof. \square

Next, we prove that if $aa^\top - bb^\top$ is small, then the vector $\frac{a}{\|a\|}$ is close to $\pm \frac{b}{\|b\|}$.

Claim (Restatement of [Claim 5.4](#)). *Let $a, b \in \mathbb{R}^d$. Let $\beta > 0$. Suppose that $\|aa^\top - bb^\top\|_F \leq \beta$. Then, we have that, for $u = \frac{a}{\|a\|}$ and $v = \frac{b}{\|b\|}$,*

$$\min\{\|u + v\|^2, \|u - v\|^2\} \leq 8 \frac{\beta^2}{\|a\|^4}$$

Proof. Note that $\min\{\|u + v\|^2, \|u - v\|^2\} = 2 - 2|\langle u, v \rangle| \leq 2$ and thus if $\beta \geq \frac{\|a\|^2}{2}$, then the conclusion trivially holds. Thus, from now on, we assume that $\beta < \frac{\|a\|^2}{2}$.

Since both the assumption and the conclusion are invariant under replacing b by $-b$, so we can without loss of generality assume that $\langle a, b \rangle \geq 0$. Let $\theta \in [0, \frac{\pi}{2}]$ be the angle between a, b , or equivalently, $\langle a, b \rangle = \|a\| \|b\| \cos \theta$. Then observe that

$$\|aa^\top - bb^\top\|_F^2 = \|aa^\top\|_F^2 + \|bb^\top\|_F^2 - 2 \langle aa^\top, bb^\top \rangle$$

$$\begin{aligned}
&= \|a\|^4 + \|b\|^4 - 2\langle a, b \rangle^2 \tag{A.4} \\
&= \|a\|^4 + \|b\|^4 - 2\|a\|^2\|b\|^2 \cos^2 \theta \\
&= (\|a\|^2 - \|b\|^2)^2 + 2\|a\|^2\|b\|^2(1 - \cos^2 \theta) \\
&= (\|a\|^2 - \|b\|^2)^2 + 2\|a\|^2\|b\|^2 \sin^2 \theta
\end{aligned}$$

Both terms are non-negative. Since the left-hand side is at most β^2 , we get that $|\|a\|^2 - \|b\|^2| \leq \beta$ and $\sqrt{2} \cdot \|a\|\|b\| \sin \theta \leq \beta$. From the first one we obtain

$$\|a\|^2 - \|b\|^2 \leq \beta < \frac{\|a\|^2}{2} \implies \|b\| \geq \frac{\|a\|}{\sqrt{2}}.$$

Using this lower bound on the second term, we get

$$\|a\|^2 \sin \theta \leq \sqrt{2} \cdot \|a\|\|b\| \sin \theta \leq \beta \implies \sin \theta \leq \frac{\beta}{\|a\|^2}.$$

Combined with $\beta < \|a\|^2/2$, we have $\sin \theta < 1/2$, which implies $\theta < \frac{\pi}{6}$. Finally, note that $\min\{\|u+v\|^2, \|u-v\|^2\} = \|u-v\|^2$ since we assumed $\langle u, v \rangle \geq 0$. Thus, we get that

$$\min\{\|u+v\|^2, \|u-v\|^2\} = \|u-v\|^2 = 2 - 2\cos \theta = 4\sin^2\left(\frac{\theta}{2}\right) \stackrel{(i)}{\leq} 2\sin^2 \theta \leq 2\frac{\beta^2}{\|a\|^4}$$

where in inequality (i) we used that for $0 \leq \theta < \frac{\pi}{6}$, $\sin \frac{\theta}{2} \leq \frac{1}{\sqrt{2}} \sin \theta$. □

Finally, we show how to relate $aa^\top - bb^\top$ to $a - b$ and $a + b$.

Fact (Restatement of [Fact 5.12](#)). *For any $a, b \in \mathbb{R}^d$ we have that $\|aa^\top - bb^\top\|_F \leq \|a+b\| \cdot \|a-b\|$.*

Proof. Recall from [\(A.4\)](#) that

$$\|aa^\top - bb^\top\|_F^2 = \|a\|^4 + \|b\|^4 - 2\langle a, b \rangle^2.$$

The proof now follows by the following calculation:

$$\begin{aligned}
\|aa^\top - bb^\top\|_F^2 &= \|a\|^4 + \|b\|^4 - 4\langle a, b \rangle^2 + 2\langle a, b \rangle^2 \\
&\leq \|a\|^4 + \|b\|^4 - 4\langle a, b \rangle^2 + 2\|a\|^2\|b\|^2 \\
&= (\|a\|^2 + \|b\|^2 + 2\langle a, b \rangle) \cdot (\|a\|^2 + \|b\|^2 - 2\langle a, b \rangle) \\
&= \|a+b\|^2 \cdot \|a-b\|^2.
\end{aligned}$$

□

A.3 Concentration results and error analysis for sample moments

In this section, we prove two claims that we deferred from [Section 5](#) about the error due to the fact that we only have access to the *sample* moments.

Claim (Restatement of [Claim 5.7](#)). *Let v be the mean of $\mathcal{N}(\mu, \Sigma)_{|\langle w, x \rangle \leq \tau}$. Let v be a fixed unit vector. Let $\{x_i\}_{i=1}^n$ be i.i.d samples from $\mathcal{N}(\mu, \Sigma)_{|\langle w, x \rangle \leq \tau}$ and let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Then, we have the following with probability at least $1 - 4e^{-10d}$*

$$\left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - \bar{x} \rangle^3 - \langle v, x_i - v \rangle^3 \right| \lesssim \kappa_2(\gamma)^{-3/2} \cdot \left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{3/2} \right).$$

Proof. For every $i \in [n]$, we have that

$$\begin{aligned}\langle v, x_i - \bar{x} \rangle^3 &= (\langle v, x_i - v \rangle + \langle v, v - \bar{x} \rangle)^3 \\ &= \langle v, x_i - v \rangle^3 + 3 \langle v, x_i - v \rangle^2 \langle v, v - \bar{x} \rangle + 3 \langle v, x_i - v \rangle \langle v, v - \bar{x} \rangle^2 + \langle v, v - \bar{x} \rangle^3.\end{aligned}$$

Thus, summing over all $i \in [n]$, we get that

$$\left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - \bar{x} \rangle^3 - \langle v, x_i - v \rangle^3 \right| \leq 3 |\langle v, \bar{x} - v \rangle| \cdot \left| \frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^2 \right| + 4 |\langle v, \bar{x} - v \rangle|^3,$$

where we used that $\sum_{i=1}^n \langle v, x_i - v \rangle = \langle v, \bar{x} - v \rangle$. Thus, it remains to bound, with high probability, the two terms $|\langle v, \bar{x} - v \rangle|$ and $|\frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^2|$. We do so separately.

Bounding $|\langle v, \bar{x} - v \rangle|$. By [Corollary 4.4](#), for every $i \in [n]$, $\langle v, x_i - v \rangle$ is sub-Gaussian with parameter $\sigma = \|\Sigma^{1/2}\| = O(1/\sqrt{\kappa_2(\gamma)})$, where we used [Lemma 4.6](#) to bound $\|\Sigma\|$. Thus, for any fixed vector v , Hoeffding's inequality (cf. [\[Ver18, Theorem 2.7.3\]](#)) implies that for every $t > 0$ we have

$$\mathbb{P}[|\langle v, \bar{x} - v \rangle| \geq t] \leq 2 \exp(-\Omega(nt^2 \cdot \kappa_2(\gamma))).$$

In particular, taking $t^2 = O\left(\frac{d}{\kappa_2(\gamma)n}\right)$, we get that with probability at least $1 - 2e^{-10d}$ we have

$$|\langle v, \bar{x} - v \rangle| \lesssim \sqrt{\frac{d}{\kappa_2(\gamma)n}}.$$

Bounding $|\frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^2|$. For every $i \in [n]$, since $\langle v, x_i - v \rangle$ is $O(1/\kappa_2(\gamma)^{1/2})$ -sub-Gaussian, $\langle v, x_i - v \rangle^2$ is $O(1/\kappa_2(\gamma))$ -sub-exponential. Thus, for fixed v , the Bernstein inequality (cf. [\[Ver18, Theorem 2.9.1\]](#)) implies that for every $t > 0$, we have

$$\mathbb{P}\left[\left|\frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^2 - \mathbb{E}\langle v, x_1 - v \rangle^2\right| \geq t\right] \leq 2 \exp(-\Omega(\min\{nt^2 \cdot \kappa_2(\gamma)^2, nt \cdot \kappa_2(\gamma)\})).$$

Taking $t = O\left(\frac{1}{\kappa_2(\gamma)} \cdot \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\right)$, it then follows that, with probability at least $1 - 2e^{-10d}$, we have that

$$\left|\frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^2\right| \lesssim \frac{1}{\kappa_2(\gamma)} \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right) + \mathbb{E}\langle v, x_1 - v \rangle^2.$$

Since $\mathbb{E}\langle v, x_1 - \mu \rangle^2 \leq O\left(\frac{1}{\kappa_2(\gamma)}\right)$ because of sub-Gaussianity, we have with probability at least $1 - 2e^{-10d}$ that

$$\left|\frac{1}{n} \sum_{i=1}^n \langle v, x_i - v \rangle^2\right| \lesssim \frac{1}{\kappa_2(\gamma)} \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right) + \frac{1}{\kappa_2(\gamma)} \lesssim \frac{1}{\kappa_2(\gamma)} \left(1 + \frac{d}{n}\right).$$

Taking a union bound over these two events, we get that, with probability $1 - 4e^{-10d}$,

$$\left|\frac{1}{n} \sum_{i=1}^n \langle v, x_i - \bar{x} \rangle^3 - \langle v, x_i - v \rangle^3\right| \lesssim \kappa_2(\gamma)^{-3/2} \cdot \left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{3/2}\right),$$

which completes the proof. \square

We now prove [Lemma 5.8](#). Recall that in [Section 5](#) we already proved the statement for fixed v . So, it remains to bound the supremum over *all* unit vectors v .

Lemma (Restatement of [Lemma 5.8](#)). *Let $\{x_i\}_{i=1}^n$ be i.i.d. samples from $\mathcal{N}(\mu, \Sigma)_{|\langle v, x \rangle| \leq \tau}$. Let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Let $\widehat{M}_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^{\otimes 3}$. Then we have with probability at least $1 - 6e^{-7d}$ that the following holds*

$$\sup_{\|v\|=1} \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \lesssim \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right).$$

Proof. In [Section 5](#), we showed that for every fixed $v \in \mathcal{S}^{d-1}$ we have

$$\left| \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \right| \leq K \cdot \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right)$$

with probability $1 - 6e^{-10d}$, where $K > 0$ is a universal constant (independent of v). Let

$$T := \widehat{M}_3 - M_3.$$

Since both \widehat{M}_3 and M_3 are symmetric order-3 tensors, T is symmetric as well. Fix a δ -net $\mathcal{N}_\delta \subseteq \mathcal{S}^{d-1}$ of the unit sphere (for some δ to be chosen later). We will show that for any $v \in \mathbb{R}^d$, $|\langle v^{\otimes 3}, T \rangle|$ is bounded by $\sup_{u \in \mathcal{N}_\delta} |\langle u^{\otimes 3}, T \rangle|$ and finish the proof by a union-bound argument over \mathcal{N}_δ .

For any $v \in \mathcal{S}^{d-1}$, choose $v' \in \mathcal{N}_\delta$ such that $\|v - v'\| \leq \delta$ and write $\Delta := v - v'$. By definition, we have $v = \Delta + v'$, and expanding $(\Delta + v')^{\otimes 3}$ gives

$$(\Delta + v')^{\otimes 3} = \Delta^{\otimes 3} + \text{Sym}(\Delta^{\otimes 2} \otimes v') + \text{Sym}(\Delta \otimes (v')^{\otimes 2}) + (v')^{\otimes 3},$$

where we use Sym to denote all permutations of a tensor (so, for example, $\text{Sym}(\Delta^{\otimes 2} \otimes v')$ is equal to $\Delta \otimes \Delta \otimes v' + \Delta \otimes v' \otimes \Delta + v' \otimes \Delta \otimes \Delta$). Taking inner product with T , and using the symmetry of T , we get

$$\begin{aligned} \langle v^{\otimes 3}, T \rangle &= \langle (\Delta + v')^{\otimes 3}, T \rangle \\ &= \langle \Delta^{\otimes 3}, T \rangle + 3 \langle \Delta^{\otimes 2} \otimes v', T \rangle + 3 \langle \Delta \otimes (v')^{\otimes 2}, T \rangle + \langle (v')^{\otimes 3}, T \rangle. \end{aligned}$$

Now, we bound the first three terms in terms of the injective tensor norm. Since T is symmetric,

$$\sup_{\|a\|=\|b\|=\|c\|=1} \langle a \otimes b \otimes c, T \rangle = \sup_{\|u\|=1} \langle u^{\otimes 3}, T \rangle.$$

Therefore, for any $a, b, c \in \mathbb{R}^d$, we have

$$\langle a \otimes b \otimes c, T \rangle \leq \|a\| \|b\| \|c\| \sup_{\|u\|=1} \langle u^{\otimes 3}, T \rangle.$$

Applying this with $(a, b, c) = (\Delta, \Delta, \Delta)$, (Δ, Δ, v') , and (Δ, v', v') , we obtain

$$\begin{aligned} \langle v^{\otimes 3}, T \rangle &\leq (\|\Delta\|^3 + 3\|\Delta\|^2\|v'\| + 3\|\Delta\|\|v'\|^2) \cdot \sup_{\|u\|=1} \langle u^{\otimes 3}, T \rangle + \langle (v')^{\otimes 3}, T \rangle \\ &\leq (\delta^3 + 3\delta^2 + 3\delta) \cdot \sup_{\|u\|=1} \langle u^{\otimes 3}, T \rangle + \langle (v')^{\otimes 3}, T \rangle \\ &\leq (\delta^3 + 3\delta^2 + 3\delta) \cdot \sup_{\|u\|=1} \langle u^{\otimes 3}, T \rangle + \sup_{u \in \mathcal{N}_\delta} \langle u^{\otimes 3}, T \rangle, \end{aligned}$$

where in the second step we used $\|v'\| = 1$ (since $v' \in \mathcal{N}_\delta \subseteq \mathcal{S}^{d-1}$) and $\|\Delta\| = \|v - v'\| \leq \delta$. This holds for every $v \in \mathcal{S}^{d-1}$. Taking the supremum over $v \in \mathcal{S}^{d-1}$ yields

$$\sup_{\|v\|=1} \langle v^{\otimes 3}, T \rangle \leq (\delta^3 + 3\delta^2 + 3\delta) \cdot \sup_{\|v\|=1} \langle v^{\otimes 3}, T \rangle + \sup_{u \in \mathcal{N}_\delta} \langle u^{\otimes 3}, T \rangle.$$

Rearranging, we get that

$$(1 - 3\delta - 3\delta^2 - \delta^3) \cdot \sup_{\|v\|=1} \langle v^{\otimes 3}, T \rangle \leq \sup_{u \in \mathcal{N}_\delta} \langle u^{\otimes 3}, T \rangle.$$

Now, we set $\delta = \frac{1}{4}$. Then, we have that

$$1 - 3\delta - 3\delta^2 - \delta^3 = 1 - \frac{3}{4} - \frac{3}{16} - \frac{1}{64} = \frac{3}{64},$$

and thus

$$\sup_{\|v\|=1} \langle v^{\otimes 3}, T \rangle \leq \frac{64}{3} \sup_{u \in \mathcal{N}_{1/4}} \langle u^{\otimes 3}, T \rangle.$$

Moreover, there exists a $1/4$ -net $\mathcal{N}_{1/4} \subseteq \mathcal{S}^{d-1}$ with cardinality $|\mathcal{N}_{1/4}| \leq 9^d$ [Ver18, Corollary 4.2.11]. Applying our fixed direction bound to each fixed $u \in \mathcal{N}_{1/4}$, and then taking a union bound, we get

$$\begin{aligned} \mathbb{P} \left[\sup_{\|v\|=1} \langle v^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \geq \frac{64K}{3} \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right) \right] \\ \leq \mathbb{P} \left[\sup_{u \in \mathcal{N}_{1/4}} \langle u^{\otimes 3}, \widehat{M}_3 - M_3 \rangle \geq K \kappa_2(\gamma)^{-3/2} \left(\sqrt{\frac{d}{n}} + \frac{d^{3/2}}{n} \right) \right] \\ \leq |\mathcal{N}_{1/4}| \cdot 6e^{-10d} \leq 9^d \cdot 6e^{-10d} \leq 6e^{-7d}. \end{aligned}$$

This proves the lemma. □

B Computations

In this appendix, we prove some of the results used for the proof of [Lemma 3.17](#). Furthermore, we will give a proof of [Claims 3.19 to 3.21](#). Finally, we provide a proof for a fact used in [Appendix A](#). All of the proofs here are (partially) computer-assisted in that we provide Mathematica code to verify some properties of certain functions.

Recall that for [Lemma 3.17](#) we there used the function

$$p_t(x) = 2t^3 - 6t + (t^4 + 14t^2 - 5)x + (2t^3 + 20t)x^2 + (t^2 + 8)x^3.$$

with derivative

$$\left(\frac{\partial}{\partial x} p_t \right) (x) = t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x + 3(t^2 + 8) \cdot x^2.$$

Fact B.1. Define $x_\ell(t) = \frac{e^{-t^2/2}}{\sqrt{\pi/2+t}}$. For any $t \geq 0$, we have that

$$\left(\frac{\partial}{\partial x} p_t \right) (x_\ell(t)) > 0$$

and

$$p_t(x_\ell(t)) > 0.$$

Proof. This can be verified with the following Mathematica code:

```

1 lb[t_] = Exp[-t^2/2]/(Sqrt[Pi/2] + t);
2 Reduce[t^4 + 14 t^2 - 5 + 4 (t^3 + 10 t) lb[t] + 3 (t^2 + 8) lb[t]^2 > 0]
3 Reduce[2 t^3 - 6 t + (t^4 + 14 t^2 - 5) lb[t] + (2 t^3 + 20 t) lb[t]^2 + (t^2 + 8) lb[
t]^3 > 0]

```

This shows that the two statements are true for $t \neq -\sqrt{\frac{\pi}{2}}$ and $t > -\sqrt{\frac{\pi}{2}}$ respectively and thus in particular for $t \geq 0$. \square

Fact B.2. *Define*

$$x_\ell(t) = \frac{\sqrt{(\pi - 2)^2 t^2 + 2\pi} - 2t}{\pi} \quad \text{and} \quad x_u(t) = \frac{\sqrt{t^2 + 2\pi} - (\pi - 1)t}{\pi}.$$

For $-\frac{4}{10} < t \leq 0$, we have

$$t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x_\ell(t)^2 > 0$$

and

$$p_t(x_\ell(t)) > 0.$$

Proof. This can be verified with the following Mathematica code:

```

1 x1[t_] = (Sqrt[(Pi - 2)^2 t^2 + 2 Pi] - 2 t)/Pi;
2 xu[t_] = (Sqrt[t^2 + 2 Pi] - (Pi - 1) t)/Pi;
3 Reduce[t^4 + 14 t^2 - 5 + 4 (t^3 + 10 t) xu[t] + 3 (t^2 + 8) x1[t]^2 > 0]
4 N[%]
5 Reduce[2 t^3 - 6 t + (t^4 + 14 t^2 - 5) x1[t] + (2 t^3 + 20 t) x1[t]^2 + (t^2 + 8) x1[
t]^3 > 0]
6 N[%]

```

This shows that the statements are true for $-1.83192 < t < 4.5716$ and $-0.419026 < t < 0.819969$ respectively and thus in particular for $-\frac{4}{10} < t \leq 0$. \square

Fact B.3. *Define*

$$r_{k+1}(x) := x + \frac{1}{x + \frac{2}{x + \frac{3}{\ddots x + \frac{k}{x}}}}.$$

Let $x_\ell(t) = r_{81}(-t)$ and $x_u(t) = r_{20}(-t)$. For $t \leq -\frac{4}{10}$, we have $x_\ell(t) > 0$,

$$t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x_\ell(t)^2 > 0 \quad (\text{B.1})$$

and

$$p_t(x_\ell(t)) > 0. \quad (\text{B.2})$$

Proof. To prove this fact, we additionally define the following function $x'_\ell(t) = r_{27}(-t)$. We claim that $x_\ell(t) \geq x'_\ell(t) > 0$ (which will also show $x_\ell(t) > 0$) and thus, since $3(t^2 + 8) > 0$, it is sufficient to show

$$t^4 + 14t^2 - 5 + 4(t^3 + 10t) \cdot x_u(t) + 3(t^2 + 8) \cdot x'_\ell(t)^2. \quad (\text{B.3})$$

in order to show (B.1). We can verify $x_\ell(t) \geq x'_\ell(t) > 0$, (B.3) and (B.2) with the following Mathematica code:

```

1 r[k_Integer, x_] := Module[{result = x, i}, For[i = k - 1, i >= 1, i--, result = x + i
  /result]; result]
2 Reduce[r[81, -t] >= r[27, -t] > 0]
3 Reduce[t^4 + 14 t^2 - 5 + 4 (t^3 + 10 t) r[20, -t] + 3 (t^2 + 8) r[27, -t]^2 > 0]
4 N[%]
5 Reduce[2 t^3 - 6 t + (t^4 + 14 t^2 - 5) r[81, -t] + (2 t^3 + 20 t) r[81, -t]^2 + (t^2
  + 8) r[81, -t]^3 > 0]
6 N[%]

```

This shows that the three statements we need to verify are true for $t < 0$, $t < -0.271834$ and $t < -0.370331$ respectively and thus in particular for $t < -\frac{4}{10}$. \square

Fact B.4. Define $\kappa_3(t) = \frac{\varphi(t)}{\Phi(t)} \cdot \left(1 - t^2 - 3t \cdot \frac{\varphi(t)}{\Phi(t)} - 2 \cdot \left(\frac{\varphi(t)}{\Phi(t)}\right)^2\right)$.

Then for $-2 < t < 2$ we have $|\kappa_3(t)| \geq \frac{1}{20}$.

Proof. This can be verified with the following Mathematica code:

```

1 phi[t_] = PDF[NormalDistribution[0, 1], t];
2 pphi[t_] = CDF[NormalDistribution[0, 1], t];
3 kappa3[t_] =
4 phi[t]/pphi[t]*(1 - t^2 - 3 t*phi[t]/pphi[t] - 2 (phi[t]/pphi[t])^2);
5 Minimize[{-kappa3[t], -2 <= t <= 2}, t];
6 Reduce[First[%] > 1/20]

```

This shows that the statement is true for $-2 < t < 2$. \square

Now, we give a proof of [Claims 3.18](#) to [3.21](#) that we restate below.

Claim (Restatement of [Claim 3.18](#)). We have $\lim_{t \rightarrow -\infty} \text{skew}(t) = -2$ and $\lim_{t \rightarrow +\infty} \text{skew}(t) = 0$. Thus, for all $t \in \mathbb{R}$ we have $\text{skew}(t) \in [-2, 0]$ and in particular $|\text{skew}(t)| \leq 2$.

Proof. We can compute these limits with the following Mathematica code:

```

1 phi[t_] = PDF[NormalDistribution[0, 1], t];
2 Phi[t_] = CDF[NormalDistribution[0, 1], t];
3 kappa2[t_] = 1 - t phi[t]/Phi[t] - (phi[t]/Phi[t])^2;
4 kappa3[t_] =
5 phi[t]/Phi[t] (1 - t^2 - 3 t phi[t]/Phi[t] - 2 (phi[t]/Phi[t])^2);
6 skew[t_] = kappa3[t]/kappa2[t]^(3/2);
7 Limit[skew[t], t -> -\[Infinity]]
8 Limit[skew[t], t -> \[Infinity]]

```

This shows that $\lim_{t \rightarrow -\infty} \text{skew}(t) = -2$ and $\lim_{t \rightarrow +\infty} \text{skew}(t) = 0$. Together with the monotonicity of skew (cf. [Lemma 3.17](#)), this implies that $\text{skew}(t) \in [-2, 0]$ and in particular $|\text{skew}(t)| \leq 2$ for all $t \in \mathbb{R}$. \square

Claim (Restatement of [Claim 3.19](#)). There exists a constant $C > 0$ such that for all $t \in \mathbb{R}$ we have

$$\frac{\frac{d}{dt} \text{skew}(t)}{\left| \frac{d}{dt} \kappa_2(t) \right|} \geq C.$$

Moreover, $|\kappa_2(a) - \kappa_2(b)| \leq \frac{1}{C} \cdot |\text{skew}(a) - \text{skew}(b)|$ for every $a, b \in \mathbb{R}$.

Proof. We define the function $g(t) := \frac{\frac{d}{dt} \text{skew}(t)}{\left| \frac{d}{dt} \kappa_2(t) \right|}$. Note that by [Lemma 3.17](#), we have $g(t) > 0$ pointwise for every $t \in \mathbb{R}$. Our goal is to show that also $\inf_{t \in \mathbb{R}} g(t) > 0$. The Mathematica code below can be used to compute the two limits $\lim_{t \rightarrow \infty} g(t)$ and $\lim_{t \rightarrow -\infty} g(t)$.

```

1 phi[t_] = PDF[NormalDistribution[0, 1], t];
2 Phi[t_] = CDF[NormalDistribution[0, 1], t];
3 kappa2[t_] = 1 - t phi[t]/Phi[t] - (phi[t]/Phi[t])^2;
4 kappa3[t_] = phi[t]/Phi[t] (1 - t^2 - 3 t phi[t]/Phi[t] - 2 (phi[t]/Phi[t])^2);
5 skew[t_] = kappa3[t]/kappa2[t]^(3/2);
6 dskew[t_] = D[skew[t], t];
7 dkappa2[t_] = D[kappa2[t], t];
8 Limit[dskew[t]/Abs[dkappa2[t]], t -> -\[Infinity]]
9 Limit[dskew[t]/Abs[dkappa2[t]], t -> \[Infinity]]

```

This shows that $\lim_{t \rightarrow \infty} g(t) = 6$ and $\lim_{t \rightarrow -\infty} g(t) = \infty$. Thus, there exist t_- and t_+ such that for all $t \notin [t_-, t_+]$ we have $g(t) \geq 5$. Finally, since the function is continuous, the minimum of g over the interval $[t_-, t_+]$ is achieved and thus there exists a constant $C' > 0$ such that also $g(t) \geq C'$ for all $t \in [t_-, t_+]$. Hence, we have that $\inf_{t \in \mathbb{R}} g(t) \geq \min(C', 5)$. Choosing $C = \min(C', 5)$ gives the first part of the claim. For the second part, note that for any $a, b \in \mathbb{R}$ we have

$$\begin{aligned}
|\kappa_2(a) - \kappa_2(b)| &= \left| \int_{\min(a,b)}^{\max(a,b)} \frac{d}{dt} \kappa_2(s) ds \right| \\
&\leq \int_{\min(a,b)}^{\max(a,b)} \left| \frac{d}{dt} \kappa_2(s) \right| ds \\
&\leq \frac{1}{C} \cdot \int_{\min(a,b)}^{\max(a,b)} \frac{d}{dt} \text{skew}(s) ds \\
&= \frac{1}{C} \cdot \text{skew}(\max(a, b)) - \text{skew}(\min(a, b)) \\
&= \frac{1}{C} \cdot |\text{skew}(a) - \text{skew}(b)|.
\end{aligned}$$

In the third step, we used the first part of the claim. And in the final step, we used that skew is monotonic by [Lemma 3.17](#). This shows the second part of the claim. \square

Claim (Restatement of [Claim 3.20](#)). Define $\psi(t) := \frac{\kappa_1(t)}{\kappa_2(t)^{1/2}}$. Then, there exists a constant $C > 0$ such that for all $t \in \mathbb{R}$ we have

$$\frac{\frac{d}{dt} \text{skew}(t)}{\left| \frac{d}{dt} \psi(t) \right|} \geq C \cdot \kappa_2(t)^2.$$

Moreover, $|\psi(a) - \psi(b)| \leq \frac{1}{C} \cdot |\text{skew}(a) - \text{skew}(b)| \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\}$ for every $a, b \in \mathbb{R}$.

Proof. We define the function $g(t) = \frac{\frac{d}{dt} \text{skew}(t)}{\left| \frac{d}{dt} \psi(t) \right| \cdot \kappa_2(t)^2}$. As in the proof of [Claim 3.19](#), we have by [Lemma 3.17](#) that $g(t) > 0$ point-wise and we want to show that also the limits $\lim_{t \rightarrow \infty} g(t)$ and $\lim_{t \rightarrow -\infty} g(t)$ are strictly positive. The following Mathematica code shows that $\lim_{t \rightarrow \infty} g(t) = \infty$ and $\lim_{t \rightarrow -\infty} g(t) = 6$.

```

1 phi[t_] = PDF[NormalDistribution[0, 1], t];
2 Phi[t_] = CDF[NormalDistribution[0, 1], t];
3 kappa1[t_] = -(phi[t]/Phi[t]);
4 kappa2[t_] = 1 - t phi[t]/Phi[t] - (phi[t]/Phi[t])^2;
5 kappa3[t_] = 
6 phi[t]/Phi[t] (1 - t^2 - 3 t phi[t]/Phi[t] - 2 (phi[t]/Phi[t])^2);
7 skew[t_] = kappa3[t]/kappa2[t]^(3/2);
8 dskew[t_] = D[skew[t], t];
9 psi[t_] = kappa1[t]/kappa2[t]^(1/2);

```

```

10 dpsi[t_] = D[psi[t], t];
11 Limit[dskew[t]/Abs[dpsi[t]*kappa2[t]^2], t -> -\[Infinity]]
12 Limit[dskew[t]/Abs[dpsi[t]*kappa2[t]^2], t -> \[Infinity]]

```

Thus, as in the proof of [Claim 3.19](#), since g is continuous, we get that there exists a constant $C > 0$ such that $g(t) \geq C$ for all $t \in \mathbb{R}$. For the second part, we argue analogous to the proof of [Claim 3.19](#). Since κ_2 is monotonically increasing by [Claim 3.8](#), we have $\kappa_2(s)^{-2} \leq \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\}$ for every s between a and b . Thus, we get that

$$\begin{aligned}
|\psi(a) - \psi(b)| &\leq \int_{\min(a,b)}^{\max(a,b)} \left| \frac{d}{dt} \psi(s) \right| ds \\
&\leq \frac{1}{C} \cdot \int_{\min(a,b)}^{\max(a,b)} \frac{d}{dt} \text{skew}(s) \cdot \kappa_2(s)^{-2} ds \\
&\leq \frac{1}{C} \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\} \int_{\min(a,b)}^{\max(a,b)} \frac{d}{dt} \text{skew}(s) \cdot ds \\
&\leq \frac{1}{C} \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\} \cdot |\text{skew}(a) - \text{skew}(b)|,
\end{aligned}$$

which completes the proof. □

Claim (Restatement of [Claim 3.21](#)). *There exists a constant $C > 0$ such that for all $t \in \mathbb{R}$ we have*

$$\frac{\frac{d}{dt} \kappa_2(t)}{\left| \frac{d}{dt} \kappa_1(t) \right|} \geq C \cdot \kappa_2(t)^2.$$

Moreover, $|\kappa_1(a) - \kappa_1(b)| \leq \frac{1}{C} \cdot |\kappa_2(a) - \kappa_2(b)| \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\}$ for every $a, b \in \mathbb{R}$.

Proof. We define the function $g(t) = \frac{\frac{d}{dt} \kappa_2(t)}{\left| \frac{d}{dt} \kappa_1(t) \right| \cdot \kappa_2(t)^2}$. Analogous as for [Claims 3.19](#) and [3.20](#), we have by [Claim 3.8](#) that $g(t) > 0$ point-wise and we want to show that also the two limits $\lim_{t \rightarrow \infty} g(t)$ and $\lim_{t \rightarrow -\infty} g(t)$ are strictly positive. With the following Mathematica code we get that $\lim_{t \rightarrow \pm\infty} g(t) = \infty$.

```

1 phi[t_] = PDF[NormalDistribution[0, 1], t];
2 Phi[t_] = CDF[NormalDistribution[0, 1], t];
3 kappa1[t_] = -(phi[t]/Phi[t]);
4 kappa2[t_] = 1 - t phi[t]/Phi[t] - (phi[t]/Phi[t])^2;
5 dkappa1[t_] = D[kappa1[t], t];
6 dkappa2[t_] = D[kappa2[t], t];
7 Limit[dkappa2[t]/Abs[dkappa1[t]*kappa2[t]^2], t -> -\[Infinity]]
8 Limit[dkappa2[t]/Abs[dkappa1[t]*kappa2[t]^2], t -> \[Infinity]]

```

As for the proof of [Claim 3.20](#), this implies that $g(t) \geq C$ for all $t \in \mathbb{R}$ for some constant $C > 0$. Furthermore, also as in [Claim 3.20](#), we get that

$$\begin{aligned}
|\kappa_1(a) - \kappa_1(b)| &\leq \frac{1}{C} \int_{\min(a,b)}^{\max(a,b)} \frac{d}{dt} \kappa_2(s) \cdot \kappa_2(s)^{-2} ds \\
&\leq \frac{1}{C} \cdot \max\{\kappa_2(a)^{-2}, \kappa_2(b)^{-2}\} \cdot |\kappa_2(a) - \kappa_2(b)|,
\end{aligned}$$

which completes the proof. □

Finally, we prove the following fact that we used in [Appendix A](#).

Fact B.5. *The function*

$$g(t) = t^2 - 1 + \frac{\exp(-t^2/2)}{\sqrt{\pi/2 + t}} \cdot \left(3t + 2 \frac{\exp(-t^2/2)}{\sqrt{\pi/2 + t}} \right)$$

is monotonically increasing on the interval $(0, 1)$.

Proof. This can be verified by the following Mathematica code, which confirms that $\frac{d}{dt}g(t) > 0$ for all $t \in (0, 1)$.

```

1 g[t_] = t^2 - 1 +
2 Exp[-t^2/2]/(Sqrt[Pi/2] + t)*(3 t + 2 Exp[-t^2/2]/(Sqrt[Pi/2] + t));
3 dg[t_] = D[g[t], t];
4 Reduce[{dg[t] > 0, 0 < t < 1}]

```

□

C Determining w

In this appendix, we show that for a vector $u \propto \Sigma w$, we can compute a vector proportional to w by computing $M_2^{-1}u$. This is not needed for our algorithm, but we include it for completeness (instead, this would be needed if one wants to estimate the survival set as they do in [LMZ24]). Note that if Σ is *isotropic* (or more generally, w is an *eigenvector* of Σ), then we can learn the direction of w from the third central moment. For arbitrary w and Σ , we get the following (where we assume $\Sigma > 0$; otherwise we cannot invert M_2).

Claim C.1. *Assume $\Sigma > 0$. Then the second central moment M_2 of the truncated distribution is invertible and for $u \propto \Sigma w$ we have $\tilde{w} := M_2^{-1}u \propto w$.*

Proof. From Lemma 4.2 we know

$$M_2 = \Sigma - (1 - \kappa_2(\gamma)) \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma w}{\|\Sigma^{1/2}w\|} \right)^\top,$$

so we get

$$M_2 w = \Sigma w - (1 - \kappa_2(\gamma)) \Sigma w = \kappa_2(\gamma) \Sigma w,$$

where we used that $w^\top \Sigma w = \|\Sigma^{1/2}w\|^2$.

We now show that M_2 is invertible. Let $\{v_i\}_{i=1}^{d-1}$ be an orthonormal basis for the orthogonal complement of $\Sigma^{1/2}w$ with respect to \mathbb{R}^d . Recall from Claim 3.7 that we have $\kappa_2(t) \in (0, 1]$. It follows that

$$\begin{aligned} & I_d - (1 - \kappa_2(\gamma)) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top \\ &= \sum_{j=1}^{d-1} v_j v_j^\top + \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top - (1 - \kappa_2(\gamma)) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top \\ &= \kappa_2(\gamma) \cdot \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top + \sum_{j=1}^{d-1} v_j v_j^\top \end{aligned}$$

$$\geq \kappa_2(\gamma) \left(\left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top + \sum_{j=1}^{d-1} v_j v_j^\top \right) = \kappa_2(\gamma) \cdot I_d$$

where the last inequality follows since $\kappa_2(\gamma) \leq 1$. This implies that, using $\kappa_2(\gamma) > 0$,

$$\begin{aligned} M_2 &= \Sigma^{1/2} \left(I_d - (1 - \kappa_2(\gamma)) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right) \left(\frac{\Sigma^{1/2}w}{\|\Sigma^{1/2}w\|} \right)^\top \right) \Sigma^{1/2} \\ &\geq \Sigma^{1/2} \kappa_2(\gamma) I_d \Sigma^{1/2} = \kappa_2(\gamma) \Sigma > 0. \end{aligned}$$

Thus M_2 is invertible and we obtain

$$w = \kappa_2(\gamma) M_2^{-1} \Sigma w.$$

Hence, if $u \propto \Sigma w$, then we get

$$\tilde{w} = M_2^{-1} u \propto M_2^{-1} \Sigma w \propto w,$$

which completes the proof. \square

References

- [AAL23] Jamil Arbas, Hassan Ashtiani, and Christopher Liaw, *Polynomial time and private learning of unbounded gaussian mixture models*, Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 1018–1040. [44](#)
- [ADVW⁺25] Josh Alman, Ran Duan, Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou, *More asymmetry yields faster matrix multiplication*, Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, Philadelphia, PA, 2025, pp. 2005–2039. MR 4863478 [2](#)
- [BCV14] Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan, *Uniqueness of tensor decompositions with applications to polynomial identifiability*, Conference on Learning Theory, PMLR, 2014, pp. 742–778. [5](#)
- [Ber60] Daniel Bernoulli, *Essai d’une nouvelle analyse de la mortalite causee par la petite verole, et des avantages de l’inoculation pour la prevenir*, Histoire de l’Acad., Roy. Sci. (Paris) avec Mem (1760), 1–45. [1](#)
- [Bha97] Rajendra Bhatia, *Matrix analysis*, Graduate Texts in Mathematics, vol. 169, Springer-Verlag, New York, 1997. [59](#)
- [Bir42] Z. W. Birnbaum, *An inequality for Mill’s ratio*, Ann. Math. Statistics **13** (1942), 245–246. MR 6640 [14](#)
- [BMA26] Mathias Barreto, Olivier Marchal, and Julyan Arbel, *Optimal sub-Gaussian variance proxy for truncated Gaussian and exponential random variables*, Statist. Probab. Lett. **228** (2026), Paper No. 110555, 9. MR 4963412 [23](#)
- [Boy59] A. V. Boyd, *Inequalities for Mills’ ratio*, Rep. Statist. Appl. Res. Un. Japan. Sci. Engrs. **6** (1959), 44–46. MR 118856 [14](#)

- [BRV13] Mikhail Belkin, Luis Rademacher, and James Voss, *Blind signal separation in the presence of gaussian noise*, Conference on Learning Theory, PMLR, 2013, pp. 270–287. [5](#)
- [BS37] Chester Ittner Bliss and W. L. Stevens, *The calculation of the time-mortality curve*, Annals of Applied Biology **24** (1937), no. 4, 815–852. [1](#)
- [BS16] Boaz Barak and David Steurer, *Proofs, beliefs, and algorithms through the lens of sum-of-squares*, Course notes: <http://www.sumofsquares.org/public/index.html> **1** (2016), 3. [5](#)
- [CGWY25] Clement Louis Canonne, Themis Gouleakis, Yuhao Wang, and Qiping Yang, *Gaussian mean testing under truncation*, Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, eds.), Proceedings of Machine Learning Research, vol. 258, PMLR, 03–05 May 2025, pp. 4879–4887. [5](#)
- [CHLZ12] Bilian Chen, Simai He, Zhening Li, and Shuzhong Zhang, *Maximum block improvement and polynomial optimization*, SIAM J. Optim. **22** (2012), no. 1, 87–107. MR 2902686 [12](#)
- [Coh91] A. Clifford Cohen, *Truncated and censored samples*, Statistics: Textbooks and Monographs, vol. 119, Marcel Dekker, Inc., New York, 1991, Theory and applications. MR 1153408 [1](#)
- [CZZ10] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou, *Optimal rates of convergence for covariance matrix estimation*, Ann. Statist. **38** (2010), no. 4, 2118–2144. MR 2676885 [1](#), [2](#)
- [DGTZ18] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis, *Efficient statistics, in high dimensions, from truncated samples*, 59th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2018, IEEE Computer Soc., Los Alamitos, CA, 2018, pp. 639–649. MR 3899629 [1](#), [3](#), [10](#)
- [DHPT25] Ilias Diakonikolas, Samuel B. Hopkins, Ankit Pensia, and Stefan Tiegel, *SoS certifiability of subgaussian distributions and its algorithmic applications*, STOC’25—Proceedings of the 57th Annual ACM Symposium on Theory of Computing, ACM, New York, 2025, pp. 1689–1700. MR 4928554 [50](#), [51](#)
- [DK70] Chandler Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. III*, SIAM J. Numer. Anal. **7** (1970), 1–46. MR 264450 [59](#)
- [DK23] Ilias Diakonikolas and Daniel M Kane, *Algorithmic high-dimensional robust statistics*, Cambridge university press, 2023. [4](#), [5](#)
- [DKK⁺19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart, *Robust estimators in high-dimensions without the computational intractability*, SIAM J. Comput. **48** (2019), no. 2, 742–864. MR 3945261 [5](#), [6](#)
- [DKPZ24] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis, *Statistical query lower bounds for learning truncated gaussians*, The Thirty Seventh Annual Conference on Learning Theory, PMLR, 2024, pp. 1336–1363. [1](#)
- [DLNS24] Anindya De, Huan Li, Shivam Nadimpalli, and Rocco A Servedio, *Detecting low-degree truncation*, Proceedings of the 56th Annual ACM Symposium on Theory of Computing, 2024, pp. 1027–1038. [5](#)

- [DNS23] Anindya De, Shivam Nadimpalli, and Rocco A Servedio, *Testing convex truncation*, Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, 2023, pp. 4050–4082. [5](#)
- [Fis31] Ronald Aylmer Fisher, *Properties and applications of Hh functions*, Introduction to British A.A.S. Math. Tables **1** (1931), xxvi–xxxv. [1](#)
- [Fri13] Shmuel Friedland, *Best rank one approximation of real symmetric tensors can be chosen symmetric*, Front. Math. China **8** (2013), no. 1, 19–40. MR 3010470 [12](#)
- [Gal98] Francis Galton, *An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data*, Proceedings of the Royal Society of London **62** (1898), no. 379-387, 310–315. [1](#)
- [GU14] Armengol Gasull and Frederic Utzet, *Approximating Mills ratio*, J. Math. Anal. Appl. **420** (2014), no. 2, 1832–1853. MR 3240110 [14](#)
- [Har70] Richard A Harshman, *Foundations of the parafac procedure: Models and conditions for an “explanatory” multi-modal factor analysis*, UCLA working papers in phonetics **16** (1970), no. 1, 84. [11](#)
- [HL13] Christopher J. Hillar and Lek-Heng Lim, *Most tensor problems are NP-hard*, J. ACM **60** (2013), no. 6, Art. 45, 39. MR 3144915 [28](#)
- [HRVF04] Philippe Huber, Elvezio Ronchetti, and Maria-Pia Victoria-Feser, *Estimation of generalized linear latent variable models*, Journal of the Royal Statistical Society Series B: Statistical Methodology **66** (2004), no. 4, 893–908. [5](#)
- [Hub92] Peter J Huber, *Robust estimation of a location parameter*, Breakthroughs in statistics: Methodology and distribution, Springer, 1992, pp. 492–518. [5](#)
- [KC22] Arun Kumar Kuchibhotla and Abhishek Chakraborty, *Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression*, Inf. Inference **11** (2022), no. 4, 1389–1456. MR 4526326 [20](#)
- [KKT25] Konstantinos Karatapanis, Vasilis Kontonis, and Christos Tzamos, *Oracle efficient truncated statistics*, The Thirteenth International Conference on Learning Representations, 2025. [2](#), [3](#)
- [KSS18] Pravesh K. Kothari, Jacob Steinhardt, and David Steurer, *Robust moment estimation and improved clustering via sum of squares*, STOC’18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2018, pp. 1035–1046. MR 3826315 [4](#), [5](#), [50](#), [51](#), [54](#)
- [KTZ19] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis, *Efficient truncated statistics with unknown truncation*, 2019 IEEE 60th Annual Symposium on Foundations of Computer Science, IEEE Comput. Soc. Press, Los Alamitos, CA, 2019, pp. 1578–1595. MR 4228243 [1](#), [2](#), [3](#)
- [Lee14] Alice Lee, *Table of the gaussian “tail” functions; when the “tail” is larger than the body*, Biometrika **10** (1914), no. 2/3, 208–214. [1](#)

- [LL22] Allen Liu and Jerry Li, *Clustering mixtures with almost optimal separation in polynomial time*, Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022, pp. 1248–1261. [5](#)
- [LMZ24] Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis, *Efficient statistics with unknown truncation, polynomial time algorithms, beyond Gaussians*, 2024 IEEE 65th Annual Symposium on Foundations of Computer Science—FOCS 2024, IEEE Computer Soc., Los Alamitos, CA, 2024, pp. 988–1006. MR 4849262 [1](#), [3](#), [4](#), [5](#), [10](#), [12](#), [13](#), [84](#)
- [LMZ26] ———, *Smoothed analysis of learning from positive samples*, Proceedings of the 58th Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '26, Association for Computing Machinery, 2026, p. 42–53. [5](#)
- [LRA93] S. E. Leurgans, R. T. Ross, and R. B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 1064–1083. MR 1238921 [11](#)
- [LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala, *Agnostic estimation of mean and covariance*, 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2016, pp. 665–674. [5](#)
- [MV10] Ankur Moitra and Gregory Valiant, *Settling the polynomial learnability of mixtures of gaussians*, 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, IEEE, 2010, pp. 93–102. [5](#)
- [Pea94] Karl Pearson, *Contributions to the mathematical theory of evolution*, Philosophical Transactions of the Royal Society of London. A **185** (1894), 71–110. [5](#)
- [Pea02] Karl Pearson, *On the systematic fitting of curves to observations and measurements*, Biometrika **1** (1902), no. 3, 265–303. [1](#)
- [PL08] Karl Pearson and Alice Lee, *On the generalised probable error in multiple normal correlation*, Biometrika **6** (1908), no. 1, 59–68. [1](#)
- [Rig15] Philippe Rigollet, *18.s997: High dimensional statistics*, (Lecture Notes), Cambridge, MA, USA: MIT Open-CourseWare (2015). [44](#)
- [Sam53] M. R. Sampford, *Some inequalities on Mill’s ratio and related functions*, Ann. Math. Statistics **24** (1953), 130–132. MR 54890 [14](#)
- [Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018. [19](#), [44](#), [77](#), [79](#)
- [vH16] Ramon van Handel, *Probability in high dimension*, Lecture notes for APC 550, Princeton University, December 2016, Version: December 21, 2016. [19](#)
- [Wey12] Hermann Weyl, *Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)*, Math. Ann. **71** (1912), no. 4, 441–479. MR 1511670 [59](#)
- [ZZ25] Manolis Zampetakis and Felix Zhou, *Private statistical estimation via truncation*, Advances in Neural Information Processing Systems (D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, eds.), vol. 38, Curran Associates, Inc., 2025, pp. 90015–90049. [5](#)