Proceedings of Machine Learning Research vol 291:1–28, 2025 38th Annual Conference on Learning Theory

Lower Bounds for Private Estimation of Gaussian Covariance Matrices under All Reasonable Parameter Regimes

Victor S. Portella VICTORSP@IME.USP.BR
University of São Paulo, Brazil

Nicholas J. A. Harvey NICKHAR@CS.UBC.CA
University of British Columbia, Canada

Editors: Nika Haghtalab and Ankur Moitra

Abstract

One of the most basic problems in statistics is estimating the covariance matrix of a Gaussian distribution. Over the past decade, researchers have studied the efficiency of covariance estimation in the setting of differential privacy. The goal is to minimize the number of samples needed to achieve the desired accuracy and privacy guarantees.

We prove lower bounds on the number of samples needed to privately estimate the covariance matrix of a Gaussian distribution. Our bounds match existing upper bounds in the widest known setting of parameters. Our analysis can be seen as a fingerprinting argument, one of the main techniques used to prove lower bounds in differential privacy. Most fingerprinting arguments rely on results analogous to the celebrated Stein's identity from probability theory. We use a matrix extension of this identity known as the Stein-Haff identity.

1. Introduction

Differential Privacy (DP) is a widely adopted framework to perform data analysis while avoiding leakage of sensitive information (Dwork et al., 2006). A major thrust of research in DP is developing privacy preserving algorithms for a variety of fundamental problems in computer science and statistics. In the past few years, a direction of particular interest has been parameter estimation of probability distributions.

Multivariate Gaussians are perhaps the canonical distribution for which to study parameter estimation. Indeed, there has been considerable work on algorithms to estimate the mean and covariance of a Gaussian distribution under both pure and approximate differential privacy. (A brief review of the literature is in Section 1.2.) To understand whether these algorithms are optimal, we require lower bounds for the error under DP. For *mean* estimation with a known covariance matrix there are lower bounds that match existing sample complexity upper bounds (Kamath et al., 2019; Aden-Ali et al., 2021). However, *covariance matrix* estimation under (ε, δ)-DP is a problem that, prior to our work, was not completely understood in all parameter regimes.

The current best (ε, δ)-DP sample complexity bound for estimating a $d \times d$ covariance matrix up to α error in Frobenius norm is $n = \tilde{O}(d^2/\alpha^2 + d^2/\alpha\varepsilon + \log(1/\delta)/\varepsilon)$ samples, due to Aden-Ali et al. (2021). (For convenience throughout this section we restrict attention to covariance matrices with all eigenvalues in $\Theta(1)$.) Regarding lower bounds, $\Omega(d^2/\alpha^2)$ samples are needed even without privacy, and at least $\Omega(\log(1/\delta)/\varepsilon)$ samples are necessary with (ε, δ)-DP (Karwa and Vadhan, 2018). Finally, Kamath et al. (2022a) and Narayanan (2023, 2025) have shown $n = \tilde{\Omega}(d^2/\alpha\varepsilon)$ lower bounds for some regimes of α and δ.

- *High-accuracy regime:* the error in Frobenius norm is $\alpha = O(1)$. In this regime, Kamath et al. (2022a) shows that if[1] $\delta = \tilde{O}(1/n)$, then $n = \Omega(d^2/\varepsilon\alpha)$;
- *Low-accuracy regime:* the error in Frobenius norm is $\alpha = O(\sqrt{d})$. (For larger α the problem is trivial.) In this regime Narayanan (2023) shows that, if $\delta = O(1/d^2)$, then $n = \tilde{\Omega}(d^2/\varepsilon\alpha)$. This result has a less restrictive hypothesis on α, but a more restrictive hypothesis on δ when $\alpha = \omega(1)$ (see Figure 1).

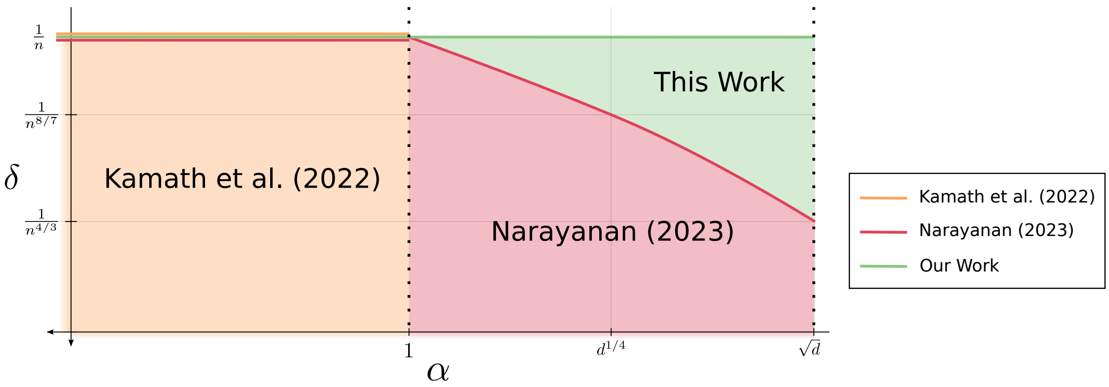

Figure 1: The largest value of δ (capped at $\tilde{O}(1/n)$) for which the lower bounds on (ε, δ)-DP estimation of Gaussian covariance matrix match the sample complexity of Aden-Ali et al. (2021) with accuracy α in Frobenius norm. Here we focus on the regime $n = \Theta(d^2/\alpha\varepsilon)$ and $\varepsilon = \Theta(1)$. Both axes are in log-scale, ignoring constants and log-factors. Although the results of Kamath et al. (2022a) may allow for larger values of δ (as well as the ones of Narayanan (2023) in the high-accuracy regime), our work recovers the results of both works for $\delta = \tilde{O}(1/n)$.

Our work completes the picture, showing $n = \Omega(d^2/\alpha\varepsilon)$ under both accuracy regimes, and with $\delta = O(1/n \ln n)$, which is near the largest meaningful value of δ (see Kamath et al. 2022a, Remark 3.7 or Vadhan 2017, §1.6). Moreover, our bound on n has no extraneous logarithmic factors and requires no regularity conditions on the mechanism. Finally, we believe our analysis technique suggests a more general strategy to obtain lower bounds for approximate DP mechanisms.

1.1. Our contributions

Our lower bounds are based on the mechanism's error α estimating a random[2] covariance matrix Σ that is "well-conditioned", by which we mean that all eigenvalues are in $[0.09, 10]$.

Theorem 1.1 (Main Theorem) *There is a random positive definite matrix Σ in $\mathbb{R}^{d \times d}$ with eigenvalues in $[0.09, 10]$ such that the following holds. Let $\mathcal{M} \colon (\mathbb{R}^d)^n \to \mathbb{R}^{d \times d}$ be (ε, δ)-DP with $\varepsilon \in (0,1)$ and*

$$\delta \leq \frac{1}{3n \ln(en)}. \tag{1}$$

Let α^2 denote the expected squared Frobenius norm error of $\mathcal{M}$ in estimating Σ with n samples from the normal distribution $\mathcal{N}(0, \Sigma)$. If $\alpha \in [2^{-d}, \frac{\sqrt{d}}{15}]$, then

$$n \;=\; \Omega\Big(\frac{d^2}{\varepsilon\alpha}\Big).$$

In our formal result (Section 2.3), Σ has support on all positive semidefinite matrices, but α is the error conditioned on Σ being well-conditioned (see (5)); the mechanism's performance may be arbitrarily poor otherwise. The statement above follows by restricting the random matrix Σ to the well-conditioned region.

[1] Their results hold for $\delta = \tilde{O}(d^2/n)$, but we often focus attention on $\delta = \tilde{O}(1/n)$ since (ε, δ)-DP loses meaning for larger δ.

[2] Although mechanisms for this problem are designed to work with inputs drawn from a distribution with a *fixed* covariance matrix Σ, we will choose Σ randomly so that it is unknown to the mechanism. Randomizing the parameters the mechanism aims to estimate is usually a fundamental part of fingerprinting arguments. The distribution on Σ is specified in Section 2.1.

The above lower bounds are quantitatively stronger than previous work (Kamath et al., 2022a; Narayanan, 2023), have no polylog factors, and do not require any regularity conditions on the mechanism. The proof relies on a fingerprinting lemma argument, a powerful strategy to obtain lower bounds in differential privacy (see Section 1.2 for an overview of the origins of the strategy). Previous fingerprinting arguments often rely on the celebrated Stein's identity (Stein, 1971) or analogous results for non-Gaussian distributions. This ultimately requires the distribution on the parameters being estimated (the entries of Σ in our case) to be independent, which restricts the distributions we can choose, an issue also discussed by Kamath et al. (2022a). We circumvent this limitation with a mathematically clean argument involving a generalization of Stein's identity, known as the Stein-Haff identity (Haff, 1979). We believe these techniques suggest a general strategy to prove fingerprinting arguments via Stokes' theorem, avoiding the need for independence.

Follow-up Work. Since a preliminary version of this work was made available online, Lyu and Talwar (2024) extended our fingerprinting arguments to a broad class of problems. One of our key contributions is using the Stein-Haff identity—a consequence of Stokes' theorem—to enable the use of distributions on Σ with non-independent entries in our fingerprinting argument. Lyu and Talwar (2024) later developed a general framework for fingerprinting arguments for a wide class of DP problems. As in our work, their approach allows for non-independent randomization of the parameters being estimated, using Stokes' theorem to overcome limitations of traditional fingerprinting arguments.

1.2. Related Work

Covariance Matrix Estimation. To learn a Gaussian it is sufficient (Ashtiani and Liaw, 2022, §2.1) and (in some sense) necessary (Arbas et al., 2023, Thm. 1.8) to estimate the mean and covariance matrix. For learning Gaussians under approximate differential privacy, Karwa and Vadhan (2018) showed polynomial-time algorithms to learn unbounded *one dimensional* Gaussians. Following their work, there were a series of works on Gaussian covariance estimation under approximate DP (Aden-Ali et al., 2021; Tsfadia et al., 2022; Liu et al., 2022; Ashtiani and Liaw, 2022; Kamath et al., 2022b; Hopkins et al., 2023), concentrated DP (Kamath et al., 2019), and pure DP (Bun et al., 2021; Hopkins et al., 2023). See Hopkins et al. (2023, Table 1) for a summary of the results on Gaussian covariance estimation. The best sample complexity known to approximate a Gaussian covariance estimation up to α error in Mahalanobis norm under (ε, δ)-DP is $n = \tilde{O}(d^2/\alpha^2 + d^2/\alpha\varepsilon + \log(1/\delta)/\varepsilon)$ due to Aden-Ali et al. (2021), with a polynomial time algorithm recently proposed by Hopkins et al. (2023). Karwa and Vadhan (2018, Thm. 1.4) shows an $\Omega(\log(1/\delta)/\varepsilon)$ lower bound for learning one dimensional Gaussians, and Kamath et al. (2019, Theorem 56) shows an $\Omega(d/\alpha\varepsilon)$ lower bound for learning spherical Gaussians. Thus, it only remains to show a $\Omega(d^2/\alpha\varepsilon)$ lower bound to conclude that the currently best-known sample complexities are not improvable (up to poly-logarithmic factors). A related problem is estimating the empirical covariance matrix out of worst-case (bounded) data (e.g., Dwork et al., 2014; Dong et al., 2022, see Narayanan 2023, §6 for a discussion on the connections with Gaussian covariance estimation).

Lower Bounds in DP. Even before the inception of DP, researchers had devised lower bounds on the accuracy of algorithms that avoid data re-identification (Dinur and Nissim, 2003). Since then there has been a long line of work on lower bounds for DP algorithms, such as packing arguments for pure (i.e., $(\varepsilon, 0)$-) DP (Hardt and Ullman, 2014), reconstruction arguments using discrepancy theory (Muthukrishnan and Nikolov, 2012), information theoretical tools (Acharya et al., 2021), or fingerprinting techniques.

Fingerprinting Techniques. Fingerprinting codes (Boneh and Shaw, 1998; Tardos, 2008) were first used in DP by Bun et al. (2018) to prove lower bounds for answering counting queries on data from the hypercube. Several works then built upon these ideas to obtain lower bounds for approximate DP algorithms for a variety of problems: statistical queries (Hardt and Ullman, 2014; Steinke and Ullman, 2015), private subspace estimation (Dwork et al., 2014), mean of vectors with ±1 entries (Steinke and Ullman, 2016), and other problems under very weak accuracy guarantees (Peter et al., 2024). A problem with fingerprinting codes is that they usually have a non-trivial construction and, because of that, are used nearly as black-boxes. Dwork et al. (2015) was one of the first works to boil down the main techniques to simpler calculations on the expected value of some "correlation statistics". This general strategy was later called the "Fingerprinting Lemma" by Bun et al. (2017). Peter et al. (2024) thoroughly discusses the main differences between fingerprinting codes and fingerprinting lemmas.

Beyond i.i.d. Priors in Fingerprinting Lemmas. Many of the previous fingerprinting proofs required randomizing the parameters *independently* of each other, which is not well suited for more structured problems such as covariance estimation. For Gaussian covariance estimation, Kamath et al. (2019) shows lower bound under *pure* DP with techniques similar to fingerprinting arguments. For the proof, they choose the entries of the *inverse* covariance matrix randomly and independently, restricting their results to the high-accuracy regime defined earlier. Kamath et al. (2022a) generalizes this argument: they propose a generalized fingerprinting lemma for parameter estimation of exponential family distributions under *approximate* DP. They use these general tools to obtain the first tight lower bounds for (ε, δ)-DP Gaussian covariance estimation, and carefully handle δ, ultimately allowing $\delta = \tilde{O}(d^2/n)$ in their results. Their argument still requires a distribution with independent entries (and bounded support) over the *natural parameters* of the exponential family parameterization, which restrict their results for covariance estimation to the high-accuracy regime.

Our lower bounds for covariance estimation use a fingerprinting argument with a statistic proposed by Cai et al. (2023) but with a different analysis strategy.[3] The distribution we use over covariance matrices is a Wishart distribution, which has *unbounded support* and *non-independent* entries. Key to our analysis is a generalization of Stein's identity known as the Stein-Haff identity (Haff, 1979). This relates to the fact that many of previous fingerprinting lemmas used results analogous to the Stein identity for specific (sometimes discrete) one-dimensional distributions (e.g., Dwork et al., 2015, Lemma 14, Kamath et al., 2019, proof of Lemma 6.8, and Cai et al., 2023, Sec. 2.2.2). Moreover, our analysis technique suggests a general strategy to handle general prior distributions over high-dimensional parameters using Stokes' Theorem. Indeed, as discussed in Section 1.1 above, this strategy was subsequently pursued by Lyu and Talwar (2024). These connections are discussed further in Section 3.

During the development of our work, Narayanan (2023) proved an $n = \tilde{\Omega}(d^2/\varepsilon\alpha)$ lower bound (losing some polylog factors due to reductions employed in the proof) in the low-accuracy ($\alpha = O(\sqrt{d})$) regime when $\delta = O(1/d^2)$. They elegantly use a Bayesian argument, relying on the fact that the *inverse* Wishart distribution is the conjugate prior of the Wishart distribution. While this regime of δ allows for $\delta = \tilde{O}(1/n)$ in the high-accuracy regime, it restricts δ to be strictly smaller than $1/n$ when $\alpha = \omega(1)$, as discussed in the journal version (Narayanan, 2025, §1.B) and illustrated in Figure 1. Our result allows δ as large as $O(1/n \ln n)$ even in the low-accuracy regime. Although both our work and Narayanan (2023) use fingerprinting-type arguments, the technical approaches are unrelated. The Bayesian viewpoint of Narayanan (2023) on fingerprinting arguments is of independent interest and opens new directions for deriving lower bounds in DP.

1.3. Notation

We write $a \lesssim b$ if there is some universal constant $C > 0$ such that $a \leq Cb$. We denote by $\mathbb{S}^d$ and $\mathbb{S}^d_+$ the set of symmetric and the set of positive semi-definite $d \times d$ matrices, respectively, and the Frobenius norm by $\|\cdot\|_F$. We let $\mathrm{E}\left[\,\cdot\,\right]$ denote an unconditional expectation, and $\mathrm{E}_{|\Sigma}\left[\,\cdot\,\right]$ denote an expectation conditioned on Σ. The notation $\Pr\left[\,\cdot\,\right]$ and $\Pr_{|\Sigma}[\cdot]$ are defined analogously. Finally, let $\mathcal{M} \colon (\mathbb{R}^d)^n \to \mathbb{S}^d_+$ be a measurable function throughout.

[3] Specifically, Cai et al. (2023, Prop. 2.2) only holds when the parameters are independent (although not explicitly stated).

2. Lower Bounds on DP Mechanisms via Fingerprinting

Let us use the setting of Gaussian covariance estimation to discuss more in depth fingerprinting arguments in general and the details of our approach. Throughout the paper, let $x_1, \ldots, x_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with distribution $\mathcal{N}(0, \Sigma)$ and density $p(\cdot \,|\, \Sigma)$, where $\Sigma \in \mathbb{S}^d_+$. Let $\mathcal{M}$ be a mechanism that estimates Σ when given as input the matrix $X \in \mathbb{R}^{d \times n}$ whose columns are $x_1, \ldots, x_n$.

The main intuition behind fingerprinting arguments is that, if $\mathcal{M}(X)$ predicts Σ somewhat accurately, it should have some correlation with $x_1, \ldots, x_n$. This intuition is not true in general since the mechanism that always outputs Σ is perfectly accurate and completely independent of its input. Yet, if Σ is unknown to the mechanism (e.g., if it is chosen randomly in the right way), this intuition can often be formalized.

The argument uses the quantity $\mathcal{A}(z, \mathcal{M}(X))$, a "correlation statistic" of $z \in \mathbb{R}^d$ with $\mathcal{M}(X)$. It should have the property that, for random Σ,

(i) $\mathrm{E}\,[\,|\mathcal{A}(z, \mathcal{M}(X))|\,]$ is small if z is independent of X and $\mathcal{M}$;

(ii) $\mathrm{E}\,[\,\mathcal{A}(z, \mathcal{M}(X))\,]$ is large if $z = x_i$ for a uniformly random $i \in [n]$;

Property (i) is usually guaranteed by the design of $\mathcal{A}$. Property (ii) often requires a more careful analysis and choice of the distribution of Σ, and this is the property usually called "Fingerprinting Lemma" in earlier works. None of the above properties depend on differential privacy: DP comes into play to show that, in fact, the expected statistics in both cases above are close to each other. Intuitively, if $\mathcal{M}$ is differentially private, it cannot be too correlated with x_i for any $i \in [n]$. In this paper we will barely even use the definition of (ϵ, δ)-DP because it is only used inside Theorem 2.1, which we use as a black box.

We use a statistic proposed in the "score attack" framework of Cai et al. (2023) defined by

$$\mathcal{A}(z, \mathcal{M}(X)) \coloneqq \langle \mathcal{M}(X) - \Sigma, \nabla_\Sigma \ln p(z \,|\, \Sigma) \rangle, \tag{2}$$

where[4] $\nabla_\Sigma \ln p(z \,|\, \Sigma) \in \mathbb{S}^d_+$ is known as the score function in the statistics literature. It is worth noting that the score attack statistic for Gaussian *mean* estimation nearly matches the statistic used in other works (e.g., see Dwork et al., 2015; Bun et al., 2017 for exact matches or Kamath et al., 2019, Lemma 6.8 for a near match). The next theorem summarizes properties of the above statistic shown by Cai et al. (2023). Note that in the next theorem, T is a parameter that can be optimized to improve the upper bound.

Theorem 2.1 (Properties of the score attack statistic, Cai et al., 2023) *For all random $\Sigma \in \mathbb{S}^d_+$,*

$$\sum_{i=1}^n \mathrm{E}_{|\Sigma}\,[\,\mathcal{A}(x_i, \mathcal{M}(X))\,] \;=\; \sum_{i,j \in [d] \colon i \geq j} \frac{\partial}{\partial \Sigma_{ij}} g(\Sigma)_{ij}, \qquad \text{where } g \coloneqq \mathrm{E}_{|\Sigma}\,[\,\mathcal{M}(X)\,]. \tag{3}$$

Moreover, if $\mathcal{M}$ is (ε, δ)-DP with $\varepsilon \in (0,1)$ and $\delta > 0$, then for all T and

$$\sum_{i=1}^n \mathrm{E}_{|\Sigma}\,[\,\mathcal{A}(x_i, \mathcal{M}(X))\,] \;\leq\; \sum_{i=1}^n \Big(2\varepsilon \alpha_\Sigma \sqrt{\lambda_{\max}(\mathcal{I}(\Sigma))} + 2\delta T + \int_T^\infty \Pr_{|\Sigma}[|\mathcal{A}(x_i, \mathcal{M}(X))| \geq t]\,\mathrm{d}t \Big),$$

where $\alpha_\Sigma \coloneqq \mathrm{E}_{|\Sigma}[\,\|\mathcal{M}(X) - \Sigma\|_F^2\,]^{1/2}$ and $\mathcal{I}(\Sigma)$ is the Fisher information matrix of $p(\cdot \,|\, \Sigma)$.

The inequality above is useful to show that score attack statistics in (i) are roughly upper bounded by $\alpha_\Sigma \sqrt{\lambda_{\max}(\mathcal{I}(\Sigma))}$ and uses that scenarios (i) and (ii) are not too far apart. Identity (3) roughly shows that (ii) will be $\Omega(d^2)$ if $g(\Sigma) = \mathrm{E}_{|\Sigma}\,[\,\mathcal{M}(X)\,] \approx \Sigma$ (the right-hand side of (3) is exactly $d(d+1)/2$ if $g(\Sigma) = \Sigma$). Yet, one needs to carefully pick the distribution of Σ to formalize this intuition.

[4] The score function in this case should be the gradient of p as a function of only the lower triangular entries of Σ to account for symmetry. Yet, we can in this case use the matrix gradient. These details are important as well when discussing the Fisher information. We defer the details to Appendix C.

2.1. Proof Overview

Let us overview how the framework above can be used to prove Theorem 1.1. Note that Theorem 2.1 does not require Σ to be random. The challenge is to randomize Σ in a way that we can meaningfully lower bound the right-hand side of (3). We will choose Σ to have a Wishart distribution with appropriate parameters, then show both an upper and a lower bound on the expected statistics $\sum_{i=1}^{n} \mathrm{E}\left[\, \mathcal{A}(x_i, \mathcal{M}(X)) \,\right]$.

- *A lower bound* of $\Omega(d^2)$ is proven in Lemma 2.2.
- *An upper bound* of roughly $O(n\epsilon\alpha)$ is proven in Lemma 2.4.

The complete proof of Theorem 1.1, given in Section 2.3, straightforwardly combines these bounds.

2.2. Distribution on Σ

As discussed as the beginning of this section, for the lower bound in Theorem 2.1 (or any fingerprinting argument) to be non-trivial we need to carefully select a distribution on the covariance matrix Σ. We will use one of the most natural distributions over $\mathbb{S}^d_+$, the (normalized) Wishart distribution.

The normalized Wishart Distribution. Let G be a $d \times D$ random standard Gaussian matrix, and let

$$\Sigma \coloneqq \frac{1}{D} G G^{\mathsf{T}} \quad \text{with} \quad D \;=\; 2d. \tag{4}$$

The distribution of GG^{T} is known as the *Wishart distribution* (of dimension d) with D degrees of freedom. We refer to the distribution of Σ as above as the *normalized Wishart distribution* with D degrees of freedom. The choice $D = 2d$ is to ensure that Σ has constant condition number with high probability.

Although natural, this distribution was not used in previous fingerprinting arguments. Kamath et al. (2022a) proposes the Generalized Fingerprinting Lemma (for exponential families). As stated, it requires the distribution of each of the coordinates of Σ^{-1} to be independent and uniform over a bounded interval, which already rules out a Wishart distribution, even if truncated to be bounded. This also forces the distribution to be such that the diameter $\mathrm{E}[\, \|\Sigma - \mathrm{E}\left[\, \Sigma \,\right]\|_F^2 \,]$ is $O(1)$, which makes their bounds only hold on the high-accuracy regime. Narayanan (2023) uses a Wishart distribution for the *inverse covariance matrix* Σ^{-1}, which has diameter $\Theta(d)$. However, their analysis requires $\delta = O(1/d^2)$ to get a tight upper bound on his correlation statistics (which are not the statistics $\mathcal{A}$ defined in (2)). In our case, we also want a distribution with diameter $\Theta(d)$ for which, simultaneously, we can meaningfully lower bound the expected value of (3). As we shall see, the choice of the Wishart distribution leads to an elegant analysis.

Error of the Mechanism. Let $\mathcal{W} \coloneqq \left\{\, A \in \mathbb{S}^d_+ \,:\, 0.09 I \preceq A \preceq 10 I \,\right\}$ be the set of what we shall call *well-conditioned matrices*. Define the expected error α^2 of $\mathcal{M}$ by

$$\alpha^2 \coloneqq \mathrm{E}\left[\, \alpha_\Sigma^2 \;\middle|\; \Sigma \in \mathcal{W} \,\right] \qquad \text{where} \qquad \alpha_\Sigma^2 \coloneqq \mathrm{E}_{|\Sigma}\left[\, \|\mathcal{M}(X) - \Sigma\|_F^2 \,\right]. \tag{5}$$

Readers familiar with the Mahalanobis norm will note that, under the event $\Sigma \in \mathcal{W}$, the error in Mahalanobis norm is the same as $\|\mathcal{M}(X) - \Sigma\|_F$ up to constants. Thus, lower bounds using α imply lower bounds on mechanisms with guarantees under the Mahalanobis norm.

To use the lower bound from the framework of Theorem 2.1, the next lemma (the "Fingerprinting Lemma" of our argument) will show non-trivial lower bounds for Σ as described in (4). We note that the boundedness of $\mathcal{M}(X)$ is a technicality discussed in Appendix A and, for our purporses, can be assumed without loss of generality (see the proof of Theorem 1.1 in Section 2.3).

Lemma 2.2 (Main Lower Bound) *Assume $\mathcal{M}(X) \preceq \beta_u I$ for $\beta_u > 0$. If $\alpha \leq \sqrt{d}/15$ and $d \geq 20$, then*

$$\sum_{i=1}^{n} \mathrm{E}\left[\, \mathcal{A}(x_i, \mathcal{M}(X)) \,\right] \;\geq\; \frac{d^2}{4}.$$

The main technical step in applying the upper bound from Theorem 2.1, whose proof we defer to Section 4, yields the following lemma.

Lemma 2.3 (Main Upper Bound, Fixed Σ) *Let $\beta_u > 0$ be an arbitrary constant. Assume that δ satisfies (1), that $\mathcal{M}(X)$ is (ε, δ)-DP, and that $\mathcal{M}(X) \preceq \beta_u I$. Then, for sufficiently large n,*

$$\sum_{i=1}^{n} \mathrm{E}_{|\Sigma}\left[\mathcal{A}(x_i, \mathcal{M}(X))\right] \;\le\; 2n\varepsilon \frac{\alpha_\Sigma}{\lambda_{\min}(\Sigma)} + 36\left(\frac{\beta_u}{\lambda_{\min}(\Sigma)} + 1\right) d^{3/2}.$$

2.3. Completing the proof of Theorem 1.1

Our upper bound (Lemma 2.3) holds for all Σ. To obtain a bound depending only on the problem parameters (n, d, etc.) we will let Σ follow a normalized Wishart distribution as described above.

Lemma 2.4 (Upper Bound with Random Σ) *Assume $\mathcal{M}(X)$ is (ε, δ)-DP with $\delta \le 1/3n\ln(en)$ and that $\mathcal{M}(X) \preceq 10I$. Then there is a constant $C > 0$ such that, for large enough n,*

$$\sum_{i=1}^{n} \mathrm{E}\left[\mathcal{A}(x_i, \mathcal{M}(X))\right] \;\le\; Cn\varepsilon(\alpha + 2^{-d}) + 2400 \cdot d^{3/2}.$$

We give a sketch below, and a complete proof in Appendix B.1.

Proof [Sketch] Taking the expectation with respect to Σ on the inequality of Lemma 2.3 yields[5]

$$\sum_{i=1}^{n} \mathrm{E}\left[\mathcal{A}(x_i, \mathcal{M}(X))\right] \le 2n\varepsilon \underbrace{\mathrm{E}\left[\frac{\alpha_\Sigma}{\lambda_{\min}(\Sigma)}\right]}_{O(\alpha+2^{-d})} + \underbrace{36\left(10\,\mathrm{E}\left[\frac{1}{\lambda_{\min}(\Sigma)}\right] + 1\right)}_{O(1)} d^{3/2}.$$

For the second term, a standard bound for Wishart matrices is $\mathrm{E}\left[\lambda_{\min}^{-1}(\Sigma)\right] \le 6.5$ (see Lemma D.8).

The first term requires care because α_Σ is random since it depends on Σ. Most of the contribution of this term is on the event $\mathcal{E} := \{\Sigma \in \mathcal{W}\}$ since $\Pr[\Sigma \notin \mathcal{W}]$ is exponentially small. On the event $\mathcal{E}$ we have $1/\lambda_{\min}(\Sigma) = O(1)$, so the contribution is $O(1) \cdot \mathrm{E}\left[\mathbb{1}_{\mathcal{E}} \cdot \alpha_\Sigma\right] \le O(1) \cdot \sqrt{\mathrm{E}\left[\mathbb{1}_{\mathcal{E}} \cdot \alpha_\Sigma^2\right]} = O(\alpha)$. ■

Combining our main lower bound (Lemma 2.2) and main upper bound (Lemma 2.4), it is straightforward to obtain our main theorem.

Theorem 1.1. Let $\mathcal{M} \colon (\mathbb{R}^d)^n \to \mathbb{S}_+^d$ be (ε, δ)-DP where $\varepsilon \in (0, 1)$ and $\delta \le 1/3n\ln(en)$. Suppose that $\alpha \in [2^{-d}, \frac{\sqrt{d}}{15}]$. Then

$$n \;=\; \Omega\left(\frac{d^2}{\varepsilon\alpha}\right).$$

Proof Without loss of generality, we may assume that the output of $\mathcal{M}$ lies in $\mathcal{W}$ by projecting the output $\mathcal{M}(X)$ onto $\mathcal{W}$. Doing so does not increase α_Σ for any $\Sigma \in \mathcal{W}$ (and hence does not increase α) since projection onto convex sets does not increase the Euclidean distance, and the Frobenius norm is a Euclidean norm.

Let Σ have the Wishart distribution described in Section 2.2. Then Lemmas 2.2 and 2.4 imply that

$$\frac{d^2}{5} \;\le\; \sum_{i=1}^{n} \mathrm{E}\left[\mathcal{A}(x_i, \mathcal{M}(X))\right] \;\le\; Cn\varepsilon(\alpha + 2^{-d}) + 2400d^{3/2} \;\le\; 2Cn\varepsilon\alpha + 2400d^{3/2},$$

since $\alpha \ge 2^{-d}$. Rearranging, we obtain $n \ge (d^2/5 - 2400d^{3/2})/2C\epsilon\alpha$, which is $\Omega(d^2/\varepsilon\alpha)$ as required. ■

[5] Here we are implicitly using the tower property of conditional expectations, that $\mathrm{E}\left[\mathrm{E}_{|\Sigma}\left[\,\cdot\,\right]\right] = \mathrm{E}\left[\,\cdot\,\right]$.

3. Lower Bound on the Correlation Statistics via the Stein-Haff Identity

In this section we shall prove our main lower bound on the correlation statistics, formally stated in Lemma 2.2. The main idea is to lower bound (3) using a result known as the Stein-Haff identity, which extends Stein's identity for Gaussian random variables to Wishart matrices. As we shall see, this suggests a strategy for fingerprinting arguments with non-independent parameters via the use of Stokes' Theorem.

First, we require some notation. Define the $d \times d$ matrix of differential operators $\mathbb{D}_\Sigma$ by

$$\mathbb{D}_\Sigma(i,j) = \frac{(1 + \mathbf{1}[i=j])}{2} \cdot \frac{\partial}{\partial \Sigma_{ij}}. \qquad \forall i,j \in [d],$$

where we set $\mathbf{1}[P]$ to be 1 if the predicate P is true, and 0 otherwise. Crucially, we identify Σ_{ij} and Σ_{ji} when differentiating. In other words, we see any function of a symmetric matrix Σ as a function of its lower triangular entries. This operator is the one that leads to the correct definition of a gradient over $\mathbb{S}^d_+$ (see Srinivasan and Panda 2023 or Appendix C.1). Surprisingly, even in prominent parts of the literature there is disagreement about the proper notion of a gradient that takes into account matrix symmetry. In this paper we treat these details carefully with the hope that it is instructive to the reader.[6] For this section, it helps to note that if $g \colon \mathbb{S}^d_+ \to \mathbb{S}^d_+$ is differentiable, then $\langle \mathbb{D}_\Sigma, g(\Sigma) \rangle \sum_{i \geq j} \partial_{\Sigma_{ij}} g(\Sigma)_{ij}$ is the *divergence* of g as a function of the lower triangular entries of Σ.

The next theorem is an extension of the classical Stein's identity (Stein, 1971; Stein et al., 2004) from normal random variables to Wishart random variables, and we shall state it in terms of general Wishart distributions. That is, we say that $\Sigma \sim \mathcal{W}_d(D; V)$ for non-singular $V \in \mathbb{S}^d_+$ if $\Sigma = GG^\mathsf{T}$ where G is a $\mathbb{R}^{d \times D}$ matrix whose columns are i.i.d. vectors each with distribution $\mathcal{N}(0, V)$.

Theorem 3.1 (Stein-Haff Identity, Haff, 1979, Thm. 2.1) *Assume $g \colon \mathbb{S}^d_+ \to \mathbb{S}^d_+$ satisfy some mild regularity conditions (see Appendix A), and let $\Sigma \sim \mathcal{W}_d(D; V)$ for some non-singular $V \in \mathbb{S}^d_+$. Then*

$$\mathrm{E}\left[\, \langle \mathbb{D}_\Sigma, g(\Sigma) \rangle \,\right] = \frac{1}{2} \mathrm{E}\left[\, \langle V^{-1} - (D - d - 1)\Sigma^{-1}, g(\Sigma) \rangle \,\right].$$

The original proof of this identity uses Stokes' Theorem and using the fact that

$$\mathbb{D}_\Sigma \cdot p_{\mathcal{W}}(\Sigma) \;=\; \tfrac{1}{2}(V^{-1} - (D - d - 1)\Sigma^{-1}) p_{\mathcal{W}}(\Sigma) \tag{6}$$

where $p_{\mathcal{W}}$ is the density of the Wishart distribution. The high level idea is to handle the left-hand side with integration by parts, moving the differential operator from g to the density of the Wishart distribution. This can be seen as a direct generalization of the integration by parts proof of the classical Stein's identity. A proof can be found in Haff (1979) and, using modern notation and tools, in Tsukuma and Kubokawa (2020, §5). Furthermore, this suggests a general avenue to prove lower bounds on the score statistics even when the parameters are not independent: use Stokes' theorem to write the expected divergence into an expression the depends on the gradient of the density, and then manipulate this expression to connect $g(\Sigma)$ to the accuracy of $\mathcal{M}$. That is exact what we do in Lemma 3.2. Following a preliminary version of this work available online, Lyu and Talwar (2024) extend the idea of using Stokes'theorem to lower bound the score statistics to other problems in differential privacy.

Lemma 3.2 *Let $g \colon \mathbb{S}^d_+ \to \mathbb{S}^d_+$ be continuously differentiable and bounded, and let $\Sigma \sim \mathcal{W}_d(D; V)$ for a non-singular $V \in \mathbb{S}^d_+$. Then*

$$\mathrm{E}\left[\, \langle \mathbb{D}_\Sigma, g(\Sigma) \rangle \,\right] \geq \frac{d(d+1)}{2} - \frac{1}{2}\sqrt{\mathrm{E}\left[\, \|\Sigma - g(\Sigma)\|_F^2 \,\right]} \cdot \sqrt{\mathrm{E}\left[\, \|(D - d - 1)\Sigma^{-1} - V^{-1}\|_F^2 \,\right]}.$$

[6] Care with symmetry is not a contribution of our work. Kamath et al. (2022a) handle symmetry similarly and our approach closely follows Srinivasan and Panda (2023); Magnus and Neudecker (1980). We defer a detailed discussion to Appendix C.

In particular, if $V = \frac{1}{D}I$ and $D = 2d$ we have

$$\mathrm{E}\left[\, \langle \mathbb{D}_\Sigma, g(\Sigma) \rangle \,\right] \geq \frac{d(d+1)}{2} - 2d^{1.5}\sqrt{\mathrm{E}\left[\, \|\Sigma - g(\Sigma)\|_F^2 \,\right]}.$$

Proof By the Stein-Haff identity, we have

$$\begin{aligned}\mathrm{E}\left[\, \langle \mathbb{D}_\Sigma, g(\Sigma) \rangle \,\right] &= \frac{1}{2}\,\mathrm{E}\left[\, \langle V^{-1} - (D-d-1)\Sigma^{-1}, g(\Sigma) \rangle \,\right] \\ &= \frac{1}{2}\,\mathrm{E}\left[\, \langle V^{-1} - (D-d-1)\Sigma^{-1}, g(\Sigma) - \Sigma \rangle + \langle V^{-1} - (D-d-1)\Sigma^{-1}, \Sigma \rangle \,\right].\end{aligned}$$

Using the fact that $\mathrm{E}\left[\, \Sigma \,\right] = DV$ we have

$$\begin{aligned}\mathrm{E}\left[\, \langle V^{-1} - (D-d-1)\Sigma^{-1}, \Sigma \rangle \,\right] &= \mathrm{E}\left[\, \mathrm{Tr}(\Sigma V^{-1} - (D-d-1)I) \,\right] \\ &= \mathrm{Tr}(DI - (D-d-1)I) = d(d+1).\end{aligned}$$

Finally, the desired inequality follows since, by Cauchy-Schwartz,

$$\begin{aligned}&\mathrm{E}\left[\, \langle V^{-1} - (D-d-1)\Sigma^{-1}, g(\Sigma) - \Sigma \rangle \,\right] \\ &\geq -\sqrt{\mathrm{E}\left[\, \|\Sigma - g(\Sigma)\|_F^2 \,\right]}\sqrt{\mathrm{E}\left[\, \|(D-d-1)\Sigma^{-1} - V^{-1}\|_F^2 \,\right]}.\end{aligned}$$

Moreover, since Σ follows a Wishart distribution, many properties of Σ^{-1}, such as the expectation and variance of its entries, are well known. (See Lemma D.3.) Specifically, $\mathrm{E}\left[\, \Sigma^{-1} \,\right] = \frac{1}{D-d-1}V^{-1}$ and, thus,

$$\begin{aligned}\mathrm{E}\left[\, \|(D-d-1)\Sigma^{-1} - V^{-1}\|_F^2 \,\right] &= (D-d-1)^2 \sum_{i,j\in[d]} \mathrm{E}\left[\, (\Sigma_{ij}^{-1} - \tfrac{1}{D-d-1}V_{ij}^{-1})^2 \,\right] \\ &= (D-d-1)^2 \sum_{i,j\in[d]} \mathrm{Var}\left[\, \Sigma_{ij}^{-1} \,\right].\end{aligned}$$

Now consider the case that $V = \frac{1}{D}I$ with $D = 2d$. Then $V_{ij}^{-1} = \mathbf{1}[i=j]D$ for all $i, j \in [d]$. Combined with the variance formulas from D.3, we get

$$\begin{aligned}(D-d-1)^2 \sum_{i,j\in[d]} \mathrm{Var}\left[\, \Sigma_{ij}^{-1} \,\right] &= d \cdot \frac{2D^2}{(D-d-3)} + d(d-1) \cdot \frac{(D-d-1)D^2}{(D-d-3)(D-d)} \\ &= d \cdot \frac{2 \cdot 4d^2}{(d-3)} + d(d-1) \cdot \frac{(d-1)4d^2}{(d-3)d} \\ &= 4d^3\left(\frac{2}{(d-3)} + \frac{(d-1)^2}{(d-3)d}\right) \leq 16d^3,\end{aligned}$$

since we assume that $d \geq 5$. This inequality holds since the left-hand side is decreasing for $d \geq 5$: the derivative of $\frac{2}{(d-3)} + \frac{(d-1)^2}{(d-3)d}$ is $\frac{-3d^2-2d+3}{(d-3)^2d^2}$, and the numerator is negative for $d \geq 5$. ■

We are now in place to prove the main lower bound (Lemma 2.2).

Proof (of Lemma 2.2) Define $g(\Sigma) = \mathrm{E}_{|\Sigma}\left[\, \mathcal{M}(X) \,\right]$. By Theorem 2.1, together with the lower bound proven in Lemma 3.2, we have

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}\left[\, \mathcal{A}(x_i, \mathcal{M}(X)) \,\right] &= \sum_{i,j\in[d]\colon i\geq j} \mathrm{E}\left[\, \frac{\partial}{\partial \Sigma_{ij}} g(\Sigma)_{ij} \,\right] = \mathrm{E}\left[\, \langle \mathbb{D}_\Sigma, g(\Sigma) \rangle \,\right] \\ &\geq \frac{d(d+1)}{2} - 2d^{1.5}\sqrt{\mathrm{E}\left[\, \|\Sigma - g(\Sigma)\|_F^2 \,\right]}.\end{aligned}$$

Note that by the conditional Jensen's inequality and the definition of α_Σ^2 from (5) we have

$$\mathrm{E}[\,\|g(\Sigma)-\Sigma\|_F^2\,] \le \mathrm{E}[\,\|\mathcal{M}(X)-\Sigma\|_F^2\,] = \mathrm{E}\left[\,\alpha_\Sigma^2\,\right].$$

By assumption, $d \ge 20$ and $\alpha^2 \le d/(15)^2 \le d/200$. Thus, by Lemma B.1 we have

$$\mathrm{E}\left[\,\alpha_\Sigma^2\,\right] \le \alpha^2 + \frac{d}{200} \le \frac{d}{100}.$$

Combining the above facts yields

$$\frac{d(d+1)}{2} - 2d^{1.5}\sqrt{\mathrm{E}\left[\,\|\Sigma - g(\Sigma)\|_F^2\,\right]} \ge \frac{d(d+1)}{2} - \frac{2d^2}{10} \ge \frac{3d^2}{10} \ge \frac{d^2}{4}.$$

■

4. Upper Bound on the Correlation Satistics

This section proves Lemma 2.3, an upper-bound on the correlation statistics when $\mathcal{M}$ is differentially private. These bounds are parameterized the covariance matrix Σ and, thus, are random variables. We shall do so by using the upper bound of Theorem 2.1, which depends on the tails of the correlation statistics. We give tail bounds for this statistics in the following lemma.

Lemma 4.1 *Assume $\mathcal{M}(X) \preceq \beta_u I$ and set $\gamma \coloneqq 1 + \beta_u/\lambda_{\min}(\Sigma)$. Then, for any $i \in [n]$ and $T \ge 6\gamma d^{3/2}$,*

$$\int_T^\infty \Pr_{|\Sigma}(|\mathcal{A}(x_i, \mathcal{M}(X))| \ge t)\,\mathrm{d}t \;\le\; 9\gamma\sqrt{d}\exp\!\left(-\frac{T}{9\gamma\sqrt{d}}\right).$$

Proof To simplify notation, let $A_i = |\mathcal{A}(x_i, \mathcal{M}(X))|$ and $z \coloneqq \Sigma^{-1/2}x_i$. Note that z has the distribution $\mathcal{N}(0, I)$. Then,

$$\begin{aligned}
A_i &= |\langle \mathcal{M}(X) - \Sigma, \nabla_\Sigma\, p(x_i \,|\, \Sigma)\rangle| && \text{(by definition in (2))}\\
&= |\langle \mathcal{M}(X) - \Sigma, \Sigma^{-1}xx^\mathsf{T}\Sigma^{-1} - \Sigma^{-1}\rangle| && \text{(by Proposition C.2)}\\
&= |\langle \Sigma^{-1/2}\mathcal{M}(X)\Sigma^{-1/2} - I, zz^\mathsf{T} - I\rangle| && \text{(cyclic property of trace)}\\
&\le \left\|\Sigma^{-1/2}\mathcal{M}(X)\Sigma^{-1/2} - I\right\|_F \cdot \left\|zz^\mathsf{T} - I\right\|_F\\
&\le \sqrt{d}\underbrace{\left(\frac{\beta_u}{\lambda_{\min}(\Sigma)} + 1\right)}_{=\gamma} \cdot \left\|zz^\mathsf{T} - I\right\|_F, && (7)
\end{aligned}$$

using the triangle inequality and $\Sigma^{-1/2}\mathcal{M}(X)\Sigma^{-1/2} \preceq \beta_u\Sigma^{-1} \preceq \frac{\beta_u}{\lambda_{\min}(\Sigma)}I$.

To complete the proof, it suffices to prove a tail bound on (7). To begin, observe that $zz^\mathsf{T} - I$ has eigenvalues $\|z\|_2^2 - 1$ with multiplicity 1, and -1 with multiplicity $d-1$, so

$$\left\|zz^\mathsf{T} - I\right\|_F \;=\; \sqrt{(\|z\|_2^2 - 1)^2 + d - 1} \;\le\; \sqrt{\|z\|_2^4 + d}.$$

Therefore,

$$\Pr_{|\Sigma}[\,A_i \ge t\,] \;\le\; \Pr_{|\Sigma}\left[\gamma\sqrt{d}\cdot\sqrt{\|z\|_2^4 + d} \ge t\right] \;=\; \Pr_{|\Sigma}\left[\|z\|_2^2 \ge \sqrt{\frac{t^2}{\gamma^2 d} - d}\right].$$

For any $t \geq 6\gamma d^{3/2}$, we may write $t \geq t/2 + 6\gamma d^{3/2}/2$. Squaring then dividing by $\gamma^2 d$, we get

$$\frac{t^2}{\gamma^2 d} \;\geq\; \frac{t^2}{4\gamma^2 d} + \frac{36d^2}{4} \;\geq\; \frac{2}{9} \cdot \frac{t^2}{\gamma^2 d} + 8d^2 + d \;=\; 18x^2 + 8d^2 + d,$$

where we have defined $x = t/9\gamma\sqrt{d}$. Thus, we have

$$\Pr_{|\Sigma}[\, A_i \geq t \,] \;\leq\; \Pr_{|\Sigma}\left[\|z\|_2^2 \geq \sqrt{\frac{t^2}{\gamma^2 d} - d} \right] \;\leq\; \Pr_{|\Sigma}\left[\|z\|_2^2 \geq \sqrt{8d^2 + 18x^2} \right] \;\leq\; e^{-x},$$

by Corollary D.2. We assume $T \geq 6\gamma d^{3/2}$, so

$$\int_T^\infty \Pr_{|\Sigma}[\, A_i \geq t \,]\, \mathrm{d}t \leq \int_T^\infty \exp\Big(- \frac{t}{9\gamma\sqrt{d}} \Big)\, \mathrm{d}t = 9\gamma\sqrt{d} \exp\Big(- \frac{T}{9\gamma\sqrt{d}} \Big).$$

■

We are now in position to prove Lemma 2.3.

Proof (of Lemma 2.3) Let $z \sim \mathcal{N}(0, \Sigma)$ be independent of X and X_i' be identical to the matrix X except with its i-th column replace by z. By Theorem 2.1 we have, for any $T > 0$,

$$\sum_{i=1}^n \mathrm{E}_{|\Sigma}\left[\mathcal{A}(x_i, \mathcal{M}(X)) \right] \leq 2\varepsilon n \alpha_\Sigma \sqrt{\lambda_{\max}(\mathcal{I}(\Sigma))} + 2n\delta T + \sum_{i=1}^n \int_T^\infty \Pr_{|\Sigma}[A_i \geq t]\, \mathrm{d}t$$

where, as before, we let $A_i = |\mathcal{A}(x_i, \mathcal{M}(X))|$. Let us first bound the latter 2 terms in the right-hand side. Set $T := 9\gamma d^{3/2} \ln(1/\delta)$ where $\gamma := \beta_u/\lambda_{\min}(\Sigma) + 1$. Since $\delta \leq 1/e$, we have $T \geq 6\gamma d^{3/2}$. Thus, by Lemma 4.1 we have

$$\begin{aligned} 2n\delta T + \sum_{i=1}^n \int_T^\infty \Pr_{|\Sigma}[A_i \geq t]\, \mathrm{d}t \;&\leq\; 9\gamma n \Big(2d^{3/2}\delta \ln(1/\delta) + \sqrt{d} \exp\big(-d \ln(1/\delta) \big) \Big) \\ &\leq\; 18\gamma n d^{3/2}(\delta \ln(1/\delta) + \delta) \;\leq\; 36\gamma d^{3/2} \end{aligned}$$

since $\delta \ln(1/\delta) \leq 1/n$ for all $n \geq 1$, by our choice of δ in (1).

To complete the proof of the desired inequality, note that Lemma C.5 yields

$$2\varepsilon n \alpha_\Sigma \sqrt{\lambda_{\max}(\mathcal{I}(\Sigma))} \leq 2\varepsilon n \frac{\alpha_\Sigma}{\lambda_{\min}(\Sigma)}.$$

■

Acknowledgments

We thank Matthew Scott, Sharvaj Kubal, Yaniv Plan, and Naomi Graham for several helpful technical discussions. We thank Abner Turkieltaub, Betty Shea, Curtis Fox, Chris Liaw and Frederik Kunstner for their feedback on early versions of the manuscript. We also thank the anonymous reviewers for their feedback on this work. This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2024/09381-1.

References

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private Assouad, Fano, and Le Cam. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *ALT 2021*, volume 132 of *Proceedings of Machine Learning Research*, pages 48–78. PMLR, 2021.

Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *ALT 2021*, volume 132 of *Proceedings of Machine Learning Research*, pages 185–216. PMLR, 2021.

Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 1018–1040. PMLR, 2023.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In *COLT 2022*, volume 178 of *Proceedings of Machine Learning Research*, pages 1075–1076. PMLR, 2022.

Timothy D. Barfoot. Multivariate gaussian variational inference by natural gradient descent, 2020. arXiv:2001.10025v2 [stat.ML].

Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Trans. Inform. Theory*, 44(5):1897–1905, 1998.

Mark Bun, Thomas Steinke, and Jonathan R. Ullman. Make up your mind: The price of online queries in differential privacy. In *SODA 2017*, pages 1306–1325. SIAM, 2017.

Mark Bun, Jonathan R. Ullman, and Salil P. Vadhan. Fingerprinting Codes and the Price of Approximate Differential Privacy. *SIAM J. Comput.*, 47(5):1888–1938, 2018.

Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *IEEE Trans. Inf. Theory*, 67(3):1981–2000, 2021.

T. Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning, 2023. arXiv:2303.07152v1 [math.ST].

Zizhong Chen and Jack J. Dongarra. Condition numbers of Gaussian random matrices. *SIAM J. Matrix Anal. Appl.*, 27(3):603–620, 2005.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In Frank Neven, Catriel Beeri, and Tova Milo, editors, *PODS 2003*, pages 202–210. ACM, 2003.

Wei Dong, Yuting Liang, and Ke Yi. Differentially private covariance revisited. In *NeurIPS 2022*, 2022.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC 2006*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *STOC 2014*. ACM, May 2014.

Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *FOCS 2015*, pages 650–669, 2015.

Gerald B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-31716-0. Modern techniques and their applications, A Wiley-Interscience Publication.

L. R. Haff. An identity for the Wishart distribution with applications. *J. Multivariate Anal.*, 9(4):531–544, 1979.

Moritz Hardt and Jonathan R. Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS 2014*, pages 454–463. IEEE Computer Society, 2014.

Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *STOC 2023*, pages 497–506. ACM, 2023.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan R. Ullman. Privately learning high-dimensional distributions. In *COLT 2019*, volume 99 of *Proceedings of Machine Learning Research*, pages 1853–1902. PMLR, 2019.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *NeurIPS 2022*, 2022a.

Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan R. Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *COLT 2022*, volume 178 of *Proceedings of Machine Learning Research*, pages 544–572. PMLR, 2022b.

Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. In *ITCS 2018*, volume 94 of *LIPIcs*, pages 44:1–44:9. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *COLT 2022*, volume 178 of *Proceedings of Machine Learning Research*, pages 1167–1246. PMLR, 2022.

Xin Lyu and Kunal Talwar. Fingerprinting codes meet geometry: Improved lower bounds for private query release and adaptive data analysis, 2024. arXiv:2412.14396v1 [cs.DS].

Jan R. Magnus and H. Neudecker. The elimination matrix: some lemmas and applications. *SIAM J. Algebraic Discrete Methods*, 1(4):422–449, 1980. ISSN 0196-5212.

S. Muthukrishnan and Aleksandar Nikolov. Optimal private halfspace counting via discrepancy. In *STOC 2012*, pages 1285–1292. ACM, 2012.

Shyam Narayanan. Better and simpler lower bounds for differentially private statistical estimation, 2023. arXiv:2310.06289v2 [math.ST].

Shyam Narayanan. Better and simpler lower bounds for differentially private statistical estimation. *IEEE Transactions on Information Theory*, 71(2):1376–1388, 2025. doi: 10.1109/TIT.2024.3511624. Journal version of Narayanan (2023).

Naty Peter, Eliad Tsfadia, and Jonathan R. Ullman. Smooth lower bounds for differentially private algorithms via padding-and-permuting fingerprinting codes. In *COLT 2024*, volume 247 of *Proceedings of Machine Learning Research*, pages 4207–4239. PMLR, 2024.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Shriram Srinivasan and Nishant Panda. What is the gradient of a scalar function of a symmetric matrix? *Indian J. Pure Appl. Math.*, 54(3):907–919, 2023. ISSN 0019-5588,0975-7465.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume II, pages 583–602, 1971.

Charles Stein, Persi Diaconis, Susan Holmes, and Gesine Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 1–26. Inst. Math. Statist., Beachwood, OH, 2004.

Thomas Steinke and Jonathan R. Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *COLT 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1588–1628. JMLR.org, 2015.

Thomas Steinke and Jonathan R. Ullman. Between pure and approximate differential privacy. *J. Priv. Confidentiality*, 7(2), 2016.

Gábor Tardos. Optimal probabilistic fingerprint codes. *Journal of the ACM (JACM)*, 55(2):1–24, 2008.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 21828–21863. PMLR, 2022.

Hisayuki Tsukuma and Tatsuya Kubokawa. *Shrinkage estimation for mean and covariance matrices*. SpringerBriefs in Statistics. Springer, Singapore, 2020. JSS Research Series in Statistics.

Salil P. Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, 2017.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Appendix A. Conditions for the Stein-Haff Identity

Let us now describe the conditions a function $g\colon \mathbb{S}^d_+ \to \mathbb{S}^d_+$ ought to satisfy for the Stein-Haff identity (Theorem 3.1) to hold, where $\mathbb{S}^d_{++}$ is the set of positive definite matrices. These are fairly mild yet technical conditions, thus one may skip this section is such details are not of interest. Ultimately, we will see we only need to assume that the mechanism $\mathcal{M}$ is measurable in order to use the Stein-Haff identity.

In this section we will not require the function g to be defined for singular matrices, as they will not arise in our application of the identity. For the remainder of this section let $p_{\mathcal{W}}$ be the density of the Wishart distribution $\mathcal{W}_d(D; V)$ for some non-singular $V \in \mathbb{S}^d_{++}$. That is, for all $\Sigma \in \mathbb{S}^d_+$ define

$$p_{\mathcal{W}}(\Sigma) \coloneqq \frac{1}{2^{Dd/2}\det(V)^{D/2}\Gamma_d(D/2)} \cdot \det(\Sigma)^{(D-d-1)/2} \exp\Big(-\frac{1}{2}\operatorname{Tr}(V^{-1}\Sigma)\Big).$$

The conditions in Haff (1979, Theorem 2.1) state that

(i) For any *strictly positive* numbers ρ_1, ρ_2, the function $g \cdot p_{\mathcal{W}}$ should be continuously differentiable (or at least Lipchitz continuous) over the set

$$B(\rho_1, \rho_2) \coloneqq \Big\{\ A \in \mathbb{S}^d_+ \ :\ \rho_1 < \|A\|_F < \rho_2\ \Big\}.$$

Note that the matrices in $B(\rho_1, \rho_2)$ can be singular.

(ii) Define $B(\rho) \coloneqq \{\ A \in \mathbb{S}^d_+ \ :\ \|A\|_F = \rho\ \}$. We need g to not grow too fast at the boundaries of $\mathbb{S}^d_+$, in the sense that

$$\lim_{\rho \to 0} \frac{\sup_{A \in B(\rho)} \|g(A)\|_F}{\rho^{1-dD/2}} = 0 \qquad \text{and} \qquad \lim_{\rho \to +\infty} \frac{\sup_{A \in B(\rho)} \|g(A)\|_F}{\rho^{1-dD/2}\exp(\gamma\rho)} = 0 \text{ for every } \gamma > 0.$$

Let us now verify that the function $g\colon \mathbb{S}^d_{++} \to \mathbb{S}^d_+$ given[7] by

$$g(\Sigma) \coloneqq \mathrm{E}\,[\,\mathcal{M}(X)\ \mid\ \Sigma\,] = \int_{\mathbb{R}^{d\times n}} \mathcal{M}(X) \cdot \Big(\prod_{i=1}^n p(x_i \,|\, \Sigma)\Big)\,\mathrm{d}X, \qquad \forall \Sigma \in \mathbb{S}^d_{++} \tag{8}$$

satisfy the above conditions, where $p(\cdot \,|\, \Sigma)$ is the density of a normal distribution with mean 0 and covariance matrix Σ, and $\mathcal{M}$ is a mechanism that is measurable and such that $\mathcal{M}(X) \in \mathcal{W}$ (we can assume this last condition holds since, if it does not, we may project onto $\mathcal{W}$ as argued in the proof of Theorem 1.1). Recall that $\mathcal{W} = \{\ A \in \mathbb{S}^d_+ \ :\ 0.09I \preceq A \preceq 10I\ \}$. In this case, $g(A) \in \mathcal{W}$ for any $A \in \mathbb{S}^d_+$. Therefore, condition (ii) is easily satisfied since $\sup_{A \in \mathbb{S}^d_+} \|g(A)\|_F$ is bounded by $10\sqrt{d}$.

Condition (i) is used to guarantee that we can apply Stokes' theorem (or a special case of it, the Gauss divergence theorem) to the function $g \cdot p_{\mathcal{W}}$ over $\mathbb{S}^d_+$ (or, actually, on $B(\rho_1, \rho_2)$ and then take the limits with ρ_1 vanishing and ρ_2 tending to infinity) and requires care to verify. Note that $B(\rho_1, \rho_2)$ is not an open set (it is an open ball intersected with $\mathbb{S}^d_+$, and there may be singular matrices in this ball). Thus, we need the existence of a continuously differentiable extension of $g \cdot p_{\mathcal{W}}$ over an open set $D \supseteq B(\rho_1, \rho_2)$ with $D \subseteq \mathbb{S}^d$. This is often not immediate since g could be defined only over the set of positive definite matrices (e.g., $g(\Sigma) = \Sigma^{-1}$) and its value (or even only the value of its derivatives) would tend to infinity for any sequence approaching the boundary of $\mathbb{S}^d_+$. Luckily, $p_{\mathcal{W}}$ decreases fast enough at the boundary of $\mathbb{S}^d_+$ that we can easily extend $g \cdot p_{\mathcal{W}}$ over *the entire set of symmetric matrices* by setting its value to zero.

[7] One could define the function g over $\mathbb{S}^d_+$, but for Σ singular the distribution $\mathcal{N}(0, \Sigma)$ does not have a density over $\mathbb{R}^d$. In this case, reasoning about differentiability of g with respect to Σ would be more challenging.

Intuitively, as $\lambda_{\min}(\Sigma_k)$ goes to 0, we have that $p_{\mathcal{W}}(\Sigma_k)$ goes to 0 as fast as $\lambda_{\min}(\Sigma_k)^{D-d-1}$, while the derivative of $g(\Sigma)$ will tend to infinity at a speed similar to $\lambda_{\min}(\Sigma_k)^{-1}$. We formally show the details of this continuously differentiable extension in the next theorem. If the reader is convinced by the intuitive argument just given, then the proof of the next theorem may be skipped.

Theorem A.1 *Let g be defined as in (8) and define $F\colon \mathbb{S}^d \to \mathbb{S}^d_+$ by*

$$F(\Sigma) := \begin{cases} g(\Sigma)\cdot p_{\mathcal{W}}(\Sigma), & \text{if } \Sigma \succ 0,\\ 0 & \text{otherwise}. \end{cases}$$

Then F is continuously differentiable.

Proof Let $i, j \in [d]$. Since $\mathcal{M}$ is bounded (that is, always in $\mathcal{W}$) and $p_{\mathcal{W}}(\Sigma) = 0$ for $\Sigma \in \mathbb{S}^d_+ \setminus \mathbb{S}^d_{++}$, one can easily verify that F is continuous. Let us show that $\partial_{\Sigma_{ij}} F(\Sigma)$ exists and is continuous over $\mathbb{S}^d$. Over $\mathbb{S}^d \setminus \mathbb{S}^d_+$ we clearly have $\partial_{\Sigma_{ij}} F(\Sigma) = 0$. Let us first derive the derivatives at $\Sigma \in \mathbb{S}^d_{++}$, and then show that the limit at the boundary is 0.

First note that

$$\partial_{\Sigma_{ij}} F(\Sigma) = g(\Sigma)\cdot \partial_{\Sigma_{ij}}(p_{\mathcal{W}}(\Sigma)) + \partial_{\Sigma_{ij}}(g(\Sigma))p_{\mathcal{W}}(\Sigma). \tag{9}$$

To compute $\partial_{\Sigma_{ij}}(g(\Sigma)) = \partial_{\Sigma_{ij}} \mathrm{E}_{|\Sigma}[\,\mathcal{M}(X)\,]$, we will exchange the order of the differential and integration (the expectation). To do so, we may the Leibniz integral rule (see, e.g., Folland (1999, Theorem 2.27)), which requires us to check that the partial derivative of the integrand in the right-hand side (8) is bounded in absolute value by an integrable function for any Σ on an open ball contained in $\mathbb{S}^d_{++}$. That is, we will show

> given $\overline{\Sigma} \in \mathbb{S}^d_{++}$ and $\varepsilon > 0$ small enough, there is a function $H(X)$ and a constant $C > 0$ (that may depend on $\overline{\Sigma}$ and d) such that, for any $\Sigma \in \mathbb{S}^d_{++}$ with $\|\overline{\Sigma} - \Sigma\|_F < \varepsilon$, we have (10)
>
> $$\partial_{\Sigma_{ij}}\Big(\mathcal{M}(X)\cdot\Big(\prod_{k=1}^n p(x_k \mid \Sigma)\Big)\Big) \leq CH(X) \qquad \text{and} \qquad \int H(X)\,\mathrm{d}X < \infty.$$

Fix $\overline{\Sigma} \in \mathbb{S}^d_{++}$ and let $\varepsilon > 0$ be small enough such that $\mathbb{B}_\varepsilon := \{\Sigma \in \mathbb{S}^d \colon \|\Sigma - \overline{\Sigma}\|_F^2 < \varepsilon\} \subseteq \mathbb{S}^d_{++}$. Let $\Sigma \in \mathbb{B}_\varepsilon$. For any $x \in \mathbb{R}^d$ define $s(x)_{ij} := \partial_{\Sigma_{ij}}(\ln p(x \mid \Sigma)) = \partial_{\Sigma_{ij}}(p(x \mid \Sigma))/p(x \mid \Sigma_{ij})$. Then,

$$\begin{aligned}\partial_{\Sigma_{ij}}\Big(\mathcal{M}(X)\cdot\Big(\prod_{k=1}^n p(x_k \mid \Sigma)\Big)\Big) &= \mathcal{M}(X)\cdot \partial_{\Sigma_{ij}}\Big(\prod_{k=1}^n p(x_k \mid \Sigma)\Big)\\ &= \mathcal{M}(X)\cdot \sum_{k=1}^n \partial_{\Sigma_{ij}} p(x_i \mid \Sigma)\Big(\prod_{r\neq i} p(x_r \mid \Sigma)\Big)\\ &= \mathcal{M}(X)\cdot \Big(\sum_{k=1}^n s(x_k)_{ij}\Big)\Big(\prod_{k=1}^n p(x_k \mid \Sigma)\Big).\end{aligned}$$

To conclude the proof of (10), it suffices to show that (each entry of) the above expression is upper-bounded by an integrable function of $x_1, \ldots, x_n$ with respect to the Lebesgue measure. For that, it suffices to show that $\mathrm{E}_{|\Sigma}[\,\mathcal{M}(X)s(x_k)_{ij}\,]$ is finite with $x_1, \ldots, x_k \sim \mathcal{N}(0, \Sigma)$. Notice that any entry of $\mathcal{M}(X)s(x_k)_{ij}$ is upper bounded by its maximum eigenvalue. Using the formula for $s(x)$ (see Proposition C.2 and Lemma C.1) and the fact that $\mathcal{M}(X) \preceq 10I$ we have

$$\begin{aligned}\lambda_{\max}(\mathcal{M}(X)\cdot s(x_k)_{ij}) \lesssim |s(x_k)_{ij}| \lesssim \lambda_{\max}\Big(\Sigma^{-1}x_k x_k^{\mathsf{T}}\Sigma^{-1} - \Sigma^{-1}\Big)&\\ = \lambda_{\max}(\Sigma^{-1})(\lambda_{\max}((\Sigma^{-1/2}x_k)(\Sigma^{-1/2}x_k)^{\mathsf{T}}) - 1)&\\ = \lambda_{\max}(\Sigma^{-1})(\|\Sigma^{-1/2}x_k\|_2^2 - 1),&\end{aligned}$$

where, as usual, the $\lesssim$ omits global constants. Since $x_k \sim \mathcal{N}(0, \Sigma)$, we have $\Sigma^{-1/2} x_k \sim \mathcal{N}(0, I)$ and, thus, $\mathrm{E}[\,\|\Sigma^{-1/2} x_k\|_2^2\,] = d$. Moreover, since $\Sigma \in \mathbb{B}_\varepsilon$, we have the (loose) bound $\lambda_{\max}(\Sigma^{-1}) \leq d\varepsilon\lambda_{\max}(\overline{\Sigma}^{-1})$. This finishes the proof of (10).

Thus, applying the Leibniz integral rule, we can exchange differentiation and integration, obtaining

$$\partial_{\Sigma_{ij}} g(\Sigma) = \mathrm{E}_{|\Sigma}\left[\mathcal{M}(X) \cdot \sum_{k=1}^{n} s(x_k)_{ij}\right] \tag{11}$$

Additionally, the dominated convergence theorem (which is applicable due to (10)) also implies that the above function is continuous on $\mathbb{S}^d_{++}$.

Next, let us derive an expression for $\partial_{\Sigma_{ij}} p_{\mathcal{W}}(\Sigma)$. Recalling the formula for the gradient[8] of $p_{\mathcal{W}}$ in (6) is given by

$$\nabla p_{\mathcal{W}}(\Sigma) = \frac{1}{2}(V^{-1} - (D - d - 1)\Sigma^{-1}) p_{\mathcal{W}}(\Sigma). \tag{12}$$

Therefore, $\partial_{\Sigma_{ij}} p_{\mathcal{W}}$ is continuous on $\mathbb{S}^d_{++}$. This, together with the continuity of $\partial_{\Sigma_{ij}} g$ and (9), implies that $\partial_{\Sigma_{ij}} F$ is continuous on $\mathbb{S}^d_{++}$.

It only remains to show continuity of $\partial_{\Sigma_{ij}} F(\Sigma)$ at the boundary $\mathrm{bd}(\mathbb{S}^d_+) \coloneqq \mathbb{S}^d_+ \setminus \mathbb{S}^d_{++}$. Let $(\Sigma_k)_{k=1}^{+\infty}$ be a convergent sequence in $\mathbb{S}^d_{++}$ such that $\lim_{k\to\infty} \Sigma_k = \overline{\Sigma} \in \mathrm{bd}(\mathbb{S}^d_+)$. We shall show that

$$\lim_{k\to\infty} \partial_{\Sigma_{ij}}(g(\Sigma_k)) p_{\mathcal{W}}(\Sigma_k) = \lim_{k\to\infty} g(\Sigma_k) \partial_{\Sigma_{ij}}(p_{\mathcal{W}}(\Sigma_k)) = 0. \tag{13}$$

This, together with the expression in (9) implies that $\partial_{\Sigma_{ij}} F(\overline{\Sigma}) = 0$, as desired. For the remainder of the proof, we shall write $a \lesssim b$ if $a \leq C \cdot b$ where C is a constant that is independent of the sequence Σ_k (the constant C *may depend on parameters such as* d and n). Then, for any index k,

$$\begin{aligned}
&\left\|\partial_{\Sigma_{ij}}(g(\Sigma_k)) p_{\mathcal{W}}(\Sigma_k)\right\|_F \\
&\leq \sum_{r=1}^{n} \mathrm{E}_{|\Sigma}\left[\,\|\mathcal{M}(X) s(x_r)_{ij}\|_F^2\,\right]^{1/2} p_{\mathcal{W}}(\Sigma_k) && \text{(By (11) and Jensen's ineq.)} \\
&\lesssim \sum_{r=1}^{n} \mathrm{E}_{|\Sigma}\left[\,s(x_r)_{ij}^2\,\right]^{1/2} p_{\mathcal{W}}(\Sigma_k) && \text{($\mathcal{M}(X)$ is bounded)} \\
&\lesssim \mathrm{E}_{|\Sigma}\left[\,s(x_1)_{ij}^2\,\right]^{1/2} p_{\mathcal{W}}(\Sigma_k) && \text{($x_1, \dots, x_n$ are i.i.d.)} \\
&= \mathcal{I}(\Sigma_k)_{ij,ij}^{1/2} \cdot p_{\mathcal{W}}(\Sigma_k) && \text{(By the def. of Fisher info. from (18))} \\
&\lesssim \lambda_{\max}(\mathcal{I}(\Sigma_k))^{1/2} \cdot p_{\mathcal{W}}(\Sigma_k) \\
&\leq \lambda_{\min}(\Sigma_k)^{-1} \cdot p_{\mathcal{W}}(\Sigma_k) && \text{(Lemma C.5).}
\end{aligned}$$

For any matrix $A \in \mathbb{S}^d$, let $\lambda_1(A), \cdots, \lambda_d(A)$ be the eigenvalues of A in non-increasing order. Since $\Sigma_k \to \overline{\Sigma}$ as $k \to \infty$, we have $\lambda_r(\Sigma_k) \to \lambda_r(\overline{\Sigma})$ for every $r \in [d]$. In particular, we have the limit $\lim_{k\to\infty} \prod_{r<n} \lambda_r(\Sigma_k) = \prod_{r<n} \lambda_r(\overline{\Sigma})$. Therefore, if $D > d - 2$

$$\begin{aligned}
\lambda_{\min}(\Sigma_k)^{-1} \cdot p_{\mathcal{W}}(\Sigma_k) \lesssim \lambda_{\min}(\Sigma_k)^{-1} \cdot \det(\Sigma_k)^{D-d-1} &= \det(\Sigma_k)^{D-d-2} \prod_{r<n} \lambda_r(\Sigma_k) \\
&\xrightarrow{k\to\infty} \underbrace{\det(\overline{\Sigma})}_{=0}{}^{D-d-2} \prod_{r<n} \lambda_r(\overline{\Sigma}) = 0.
\end{aligned}$$

[8] See Section C.1 for a discussion on gradients for functions of symmetric matrices and Lemma C.1 for a result showing that the partial derivatives are scaled entries of the gradient.

So far we have analyzed one of the limits in (13). The other limit can be analyzed by similar arguments, also under the assumption that $D > d - 2$. Using the formula for the gradient in (12), but omitting details for the sake of conciseness, we obtain

$$\left\|g(\Sigma)\cdot\partial_{\Sigma_{ij}}(p_{\mathcal{W}}(\Sigma))\right\|_F \lesssim \left(\|V\|_F + \left\|\Sigma_k^{-1}\right\|_F\right)\cdot p_{\mathcal{W}}(\Sigma_k) \lesssim \frac{p_{\mathcal{W}}(\Sigma_k)}{\lambda_{\min}(\Sigma)} \overset{k\to\infty}{\longrightarrow} 0.$$

This concludes the proof of (13). ■

Appendix B. Omitted proofs

B.1. Omitted material from Section 2

Proof (of Lemma 2.4) Taking the expectation with respect to Σ on the inequality of Lemma 2.3, then using the tower property of conditional expectation, yields

$$\begin{aligned}\mathrm{E}\left[\sum_{i=1}^{n}\mathrm{E}_{|\Sigma}\left[\mathcal{A}(x_i,\mathcal{M}(X))\right]\right] &= \sum_{i=1}^{n}\mathrm{E}\left[\mathcal{A}(x_i,\mathcal{M}(X))\right]\\ &\leq 2n\varepsilon\,\mathrm{E}\left[\frac{\alpha_\Sigma}{\lambda_{\min}(\Sigma)}\right] + 36\left(10\,\mathrm{E}\left[\frac{1}{\lambda_{\min}(\Sigma)}\right]+1\right)d^{3/2}.\end{aligned}$$

A standard bounds for Wishart matrices (see Lemma D.8) is that $\mathrm{E}\left[\lambda_{\min}^{-1}(\Sigma)\right] \leq 6.5$, so the second term on the right-hand side is at most $36\cdot 66\cdot d^{3/2}$. Thus, it remains to show that there is a constant $C > 0$ such that

$$\mathrm{E}\left[\alpha_\Sigma/\lambda_{\min}(\Sigma)\right] \leq 10\alpha + C2^{-d}. \tag{14}$$

We will do so by separately bounding $\mathrm{E}\left[(\alpha_\Sigma/\lambda_{\min}(\Sigma))\cdot\mathbb{1}_{\mathcal{E}}\right]$ and $\mathrm{E}\left[(\alpha_\Sigma/\lambda_{\min}(\Sigma))\cdot\mathbb{1}_{\bar{\mathcal{E}}}\right]$, where $\mathcal{E}$ is the event $\{\Sigma\in\mathcal{W}\}$. Under $\mathcal{E}$ we have $\lambda_{\min}(\Sigma)\geq 0.09 \geq 1/12$ and, by the definition of α (see (5)), we have $\mathrm{E}\left[\alpha_\Sigma\cdot\mathbb{1}_{\mathcal{E}}\right] \leq \mathrm{E}\left[\alpha_\Sigma \,|\, \mathcal{E}\right] = \alpha$. Therefore, $\mathrm{E}\left[(\alpha_\Sigma/\lambda_{\min}(\Sigma))\cdot\mathbb{1}_{\mathcal{E}}\right] \leq 12\alpha$. For the other term, first we can use the fact that $\mathcal{M}(X) \preceq 10I$ to get

$$\begin{aligned}\alpha_\Sigma &= \sqrt{\mathrm{E}_{|\Sigma}\left[\|\mathcal{M}(X)-\Sigma\|_F^2\right]} \leq \sqrt{\mathrm{E}_{|\Sigma}\left[\|\mathcal{M}(X)-\Sigma\|_F^2\right]} \leq \sqrt{\mathrm{E}_{|\Sigma}\left[2\,\|\mathcal{M}(X)\|_F^2 + 2\,\|\Sigma\|_F^2\right]}\\ &\leq \sqrt{2d(10^2+\lambda_{\max}(\Sigma)^2)} \leq \sqrt{2d}(10+\lambda_{\max}(\Sigma)).\end{aligned}$$

Define $\kappa(\Sigma) \coloneqq \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. Then,

$$\mathrm{E}\left[\frac{\alpha_\Sigma}{\lambda_{\min}(\Sigma)}\cdot\mathbb{1}_{\bar{\mathcal{E}}}\right] \leq 10\sqrt{2d}\,\mathrm{E}\left[\lambda_{\min}(\Sigma)^{-1}\cdot\mathbb{1}_{\bar{\mathcal{E}}}\right] + \sqrt{2d}\,\mathrm{E}\left[\kappa(\Sigma)\cdot\mathbb{1}_{\bar{\mathcal{E}}}\right].$$

Let us start by noticing that Lemmas D.6 and D.8 together yield

$$\Pr\left[\bar{\mathcal{E}}\right] \leq \Pr\left[\lambda_{\min}(\Sigma) < 0.09\right] + \Pr\left[\lambda_{\max}(\Sigma) > 10\right] \leq \frac{2}{\sqrt{d}}2^{-d}. \tag{15}$$

Using the tail bound (20) we have

$$\begin{aligned}\mathrm{E}\left[\frac{1}{\lambda_{\min}(\Sigma)}\cdot\mathbb{1}_{\bar{\mathcal{E}}}\right] &= \int_0^\infty \Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t \,\wedge\, \bar{\mathcal{E}}\right] \leq 4e\cdot\Pr\left[\bar{\mathcal{E}}\right] + \int_{4e}^\infty \Pr\left[\lambda_{\min}(\Sigma)^{-1} > t\right]\mathrm{d}t\\ &\leq 4e\cdot\Pr\left[\bar{\mathcal{E}}\right] + \int_{4e}^\infty \frac{1}{\sqrt{2\pi d}}(2e)^{d+1}\frac{1}{t^{d+1}}\,\mathrm{d}t = 4e\cdot\Pr\left[\bar{\mathcal{E}}\right] + \frac{1}{\sqrt{2\pi d}}\frac{1}{d}\frac{(2e)^{d+1}}{(4e)^d}\\ &\lesssim \frac{1}{\sqrt{d}}2^{-d} + \frac{1}{\sqrt{d}}2^{-d} \leq \frac{2}{\sqrt{d}}2^{-d},\end{aligned}$$

where in the last step we used $d \geq 2e$. Similarly, using the tail bound in Lemma D.10,

$$
\begin{aligned}
\mathrm{E}\left[\,\kappa(\Sigma)\cdot \mathbb{1}_{\bar{\mathcal{E}}}\,\right] &\leq 10^4 \Pr\left[\,\bar{\mathcal{E}}\,\right] + \int_{10^4}^{\infty} \frac{(13)^{d+1}}{\sqrt{2\pi}} \cdot \frac{1}{t^{(d+1)/2}}\,\mathrm{d}t \\
&\lesssim \frac{1}{\sqrt{d}} 2^{-d} + \int_{10^4}^{\infty} (13)^{d+1} \cdot \frac{1}{t^{(d+1)/2}}\,\mathrm{d}t \\
&= \frac{1}{\sqrt{d}} 2^{-d} + 13\frac{2}{d+1}\frac{13^d}{(10^4)^{(d-1)/2}} \\
&\lesssim \frac{1}{\sqrt{d}} 2^{-d} + \frac{1}{d}\left(\frac{13}{10^2}\right)^d \\
&\leq \frac{1}{\sqrt{d}} 2^{-d} + \frac{1}{\sqrt{d}} 2^{-d}.
\end{aligned}
$$

Putting everything together yields

$$
\mathrm{E}\left[\frac{\alpha_\Sigma}{\lambda_{\min}(\Sigma)} \cdot \mathbb{1}_{\bar{\mathcal{E}}}\right] \lesssim \sqrt{d} \cdot \frac{1}{\sqrt{d}} 2^{-d} \leq 2^{-d},
$$

which finishes the proof of (14). ■

B.2. Omitted material from Section 3

Lemma B.1 *For α and α_Σ defined as in (5) we have*

$$
\mathrm{E}\left[\,\alpha_\Sigma^2\,\right] = \mathrm{E}\left[\,\|\mathcal{M}(X) - \Sigma\|_F^2\,\right] \leq \alpha^2 + 600d \cdot \frac{2^{-d}}{\sqrt{d}}.
$$

In particular, if $d \geq 19$, then $\mathrm{E}\left[\,\alpha_\Sigma^2\,\right] \leq \alpha^2 + d/200.$

Proof Define the event $\mathcal{E} := \{\Sigma \in \mathcal{W}\}$. Then, by the definition of α,

$$
\mathrm{E}\left[\,\alpha_\Sigma^2\,\right] = \mathrm{E}\left[\,\alpha_\Sigma^2 \mathbb{1}_{\mathcal{E}}\,\right] + \mathrm{E}\left[\,\alpha_\Sigma^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right] \leq \alpha^2 + \mathrm{E}\left[\,\alpha_\Sigma^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right].
$$

Moreover, similarly to what we did in the proof of Lemma 2.4, we have

$$
\begin{aligned}
\mathrm{E}\left[\,\alpha_\Sigma^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right] \leq \mathrm{E}\left[\,2d(10^2 + \lambda_{\max}(\Sigma)^2)\mathbb{1}_{\bar{\mathcal{E}}}\,\right] &= 200d \Pr\left[\,\bar{\mathcal{E}}\,\right] + d\,\mathrm{E}\left[\,\lambda_{\max}(\Sigma)^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right] \\
&\overset{(15)}{\leq} 200d\frac{2}{\sqrt{d}}2^{-d} + d\,\mathrm{E}\left[\,\lambda_{\max}(\Sigma)^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right]
\end{aligned}
$$

Let us now give an $O(2^{-d}/\sqrt{d})$ upper bound the latter term. We have,

$$
\begin{aligned}
\mathrm{E}\left[\,\lambda_{\max}(\Sigma)^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right] &\leq 80 \cdot \Pr\left[\,\bar{\mathcal{E}}\,\right] + \int_{80}^{\infty} \Pr\left[\,\lambda_{\max}(\Sigma)^2 \geq t\,\right] \mathrm{d}t \\
&= 80 \cdot \Pr\left[\,\bar{\mathcal{E}}\,\right] + \int_{8}^{\infty} \Pr\left[\,\lambda_{\max}(\Sigma) \geq \sqrt{72+s}\,\right] \mathrm{d}s \\
&\leq 80 \cdot \Pr\left[\,\bar{\mathcal{E}}\,\right] + \int_{8}^{\infty} \Pr\left[\lambda_{\max}(\Sigma) \geq \underbrace{\sqrt{72/2}}_{=6} + \sqrt{s/2}\right] \mathrm{d}s
\end{aligned}
$$

$$
\begin{aligned}
&\leq 80 \cdot \Pr\left[\,\bar{\mathcal{E}}\,\right] + \int_8^\infty \exp\left(\frac{-d\sqrt{s}}{2\sqrt{2}}\right) \mathrm{d}s && \text{(By Lemma D.6)}\\
&\leq 80 \cdot \Pr\left[\,\bar{\mathcal{E}}\,\right] + \frac{4\sqrt{2}(d\sqrt{8} + 2\sqrt{2})}{d^2} \exp\left(\frac{-d\sqrt{8}}{2\sqrt{2}}\right)\\
&\leq 80 \cdot \Pr\left[\,\bar{\mathcal{E}}\,\right] + \frac{16(d+1)}{d^2} e^{-d}\\
&\leq 160 \cdot \frac{1}{\sqrt{d}} 2^{-d} + \frac{160}{9d} e^{-d} && (d \geq 9)\\
&\leq 160 \cdot \frac{10}{9\sqrt{d}} 2^{-d}.
\end{aligned}
$$

Thus,

$$
\mathrm{E}\left[\,\alpha_\Sigma^2 \mathbb{1}_{\bar{\mathcal{E}}}\,\right] \leq d \cdot \frac{2^{-d}}{\sqrt{d}}\left(400 + \frac{1600}{9}\right) \leq 600d \cdot \frac{2^{-d}}{\sqrt{d}}
$$

In particular, one may verify that for $d \geq 19$ we have $600\sqrt{d}2^{-d} \leq 1/200$ ■

Appendix C. Score and Fisher Information of Gaussian with Unknown Covariance

In this section we shall rigorously derive the formulas for the score function and the Fisher information matrix, with respect to the parameter $\Sigma \in \mathbb{S}^d_+$, for the Gaussian density

$$
p(x \,|\, \Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\Big(-\frac{1}{2} x^\mathsf{T} \Sigma^{-1} x\Big), \qquad \forall x \in \mathbb{R}^d.
$$

Since the function is meant to be evaluated only for symmetric, positive definite matrices Σ, one should take into account the symmetry of Σ when manipulating the derivatives and gradients of $p(\cdot \,|\, \Sigma)$ with respect to Σ (see, e.g., the discussions in Srinivasan and Panda, 2023).

If fact, these results are already known (see Magnus and Neudecker, 1980, §5 or Barfoot, 2020), but they are not always derived with the adequate amount of rigor. Furthermore, one may find a variety of different versions of the Fisher information, both depending on the parameterization and how one takes symmetry in account during differentiation. It shall be important for us to rigorously derive the score and Fisher information to then connect these results to the Stein-Haff identity in Section 3. Moreover, we shall derive a formula for the Fisher information directly from its definition as the covariance matrix of the score, without making use of its connection to the second derivative of the score (which needs care when taking symmetry into account).

We shall first discuss how to properly take the symmetry in differentiation. Then we shall derive the formulas of the score function (including differentiating between "matrix score" and the classical score function) and the Fisher information matrix of the Gaussian density.

C.1. Gradients of Functions of Symmetric Matrices

Let $F \colon \mathbb{S}^d \to \mathbb{R}$ be a real valued function over symmetric matrices. (Later we will focus on the function $\Sigma \in \mathbb{S}^d_+ \mapsto \ln p(x \,|\, \Sigma)$.) Due to symmetry, we may actually consider F to be a function of the lower triangular portion of the matrix, that is, restricted to $\mathbb{S}^d$, the function F is a real function over a space isomorphic to $\mathbb{R}^{d(d+1)/2}$. Formalizing this turns out to be crucial to correctly define and compute the score function and Fisher information matrices of a Gaussian with unknown covariance.

More formally, define $\binom{[d]}{k} := \{\, S \subseteq [d] : |S| = k \,\}$. For any matrix $A \in \mathbb{R}^{d \times d}$ (not necessarily symmetric), we shall denote the vector containing the entries in the lower triangular portion of A by $\mathrm{vech}(\Sigma)$

which we parameterize it in the natural way by $\binom{[d]}{2} \cup \binom{[d]}{1}$. Formally, for any matrix $A \in \mathbb{R}^{d \times d}$ we define the vector $\operatorname{vech}(A) \colon \binom{[d]}{2} \cup \binom{[d]}{1} \to \mathbb{R}$ by

$$\operatorname{vech}(A)_{\{i,j\}} \coloneqq A_{ij} \qquad \forall i, j \in [d] \text{ with } i \geq j.$$

Analogously, we define the d^2 dimensional vector[9] $\operatorname{vec}(A) \in \mathbb{R}^{[d] \times [d]}$ by

$$\operatorname{vec}(A)_{(i,j)} \coloneqq A_{ij}, \qquad \forall i, j \in [d].$$

The idea of the duplication matrix (Magnus and Neudecker, 1980) will be useful in navigating these different spaces. The *Duplication matrix* is the matrix $D \colon ([d] \times [d]) \times (\binom{[d]}{2} \cup \binom{[d]}{1}) \to \{0, 1\}$ given by

$$D((r,s), \{i,j\}) = \mathbf{1}[\{r,s\} = \{i,j\}], \qquad \forall i, j, r, s \in [d].$$

In words, for any matrix $A \in \mathbb{R}^{d \times d}$, the vector $D \operatorname{vech}(A)$ is the d^2 dimensional vector comprised of the lower triangle entries of A, with the off-diagonal entries duplicated. If A is symmetric then $D \operatorname{vech}(A) = \operatorname{vec}(A)$. Following Srinivasan and Panda (2023), define the *symmetric gradient of* F by[10]

$$\nabla_{\text{sym}} F(A) = \frac{1}{2}(\nabla F(A) + \nabla F(A)^{\mathsf{T}}), \tag{16}$$

where $\nabla F(A) \in \mathbb{R}^{d \times d}$ is the traditional gradient of F, that is, $(\nabla F(A))_{ij}$ is the derivative of F evaluated at A *not taking into account symmetry*. This definition is far from arbitrary: Srinivasan and Panda (2023) show that $\nabla_{\text{sym}} F(\Sigma)$ is the gradient that satisfies the definition of *Frechét derivative* over $\mathbb{S}^d$, the space of symmetric $d \times d$ matrices. This identity shall be useful later to compute the score function of a Gaussian with respect to the covariance matrix using traditional matrix calculus.

Finally, the following lemma connects the symmetric gradient of F and the gradient of the function f given by $f(\operatorname{vech}(A)) \coloneqq F(A)$ for any $A \in \mathbb{S}^d$. Namely, it shows that $\nabla_{\text{sym}} F$ is defined in such a way such that $\langle \nabla_{\text{sym}} F(A), B \rangle$ agrees with $\nabla f(\operatorname{vech}(A))^{\mathsf{T}} \operatorname{vech}(B)$.

Lemma C.1 *Let $F \colon \mathbb{S}^d \to \mathbb{R}$ and $f \colon \mathbb{R}^{\binom{[d]}{2} \cup \binom{[d]}{1}} \to \mathbb{R}$ be such that $F(A) = f(\operatorname{vech}(A))$ for all $A \in \mathbb{S}^d$. Then, for all $i, j \in [d]$ and any $A \in \mathbb{S}^d$, we have*

$$\nabla_{\text{sym}} F(A)_{ij} = \frac{(1 + \mathbf{1}[i \neq j])}{2} \nabla f(\operatorname{vech}(A))_{\{i,j\}}.$$

In particular, for any symmetric matrix B we have $\langle B, \nabla_{\text{sym}} F(A) \rangle = \operatorname{vech}(B)^{\mathsf{T}} \nabla f(\operatorname{vech}(A))$.

Proof This is a combination of Srinivasan and Panda (2023, Theorem 3.8) and the formula for the pseudo-inverse of D given by Magnus and Neudecker (1980, Lemma 3.6.iv). ■

[9] For the sake of preciseness, we shall for this section differentiate the space $\mathbb{R}^{d \times d}$ of $d \times d$ matrices and the space $\mathbb{R}^{[d] \times [d]}$ of vectors indexed by ordered pairs of elements in $[d]$ so that we can use vector operations in elements of $\mathbb{R}^{[d] \times [d]}$ without requiring us to overload notation in the space of matrices.

[10] This definition requires F to be well-defined and differentiable over non-symmetric matrices. Srinivasan and Panda (2023) also show how to write this gradient when we have a function F that is *only* defined over symmetric matrices and cannot be extended to $\mathbb{R}^{d \times d}$.

C.2. Score Function Derivation

Let us start by discussing the definition and derivation of the score function of the Gaussian distribution $\mathcal{N}(0, \Sigma)$ with respect to the covariance matrix $\Sigma \in \mathbb{S}^d_+$. As discussed in the previous section, to properly take into account the symmetry of the covariance matrix Σ, we should look at the function $f_x \colon \mathbb{R}^{\binom{[d]}{2} \cup \binom{[d]}{1}} \to \mathbb{R}$ defined by

$$f_x(\mathrm{vech}(\Sigma)) \coloneqq \ln p(x \,|\, \Sigma), \qquad \forall \Sigma \in \mathbb{R}^{d \times d}, \qquad \forall x \in \mathbb{R}^d, \tag{17}$$

where we let $f_x(\mathrm{vech}(\Sigma))$ evaluate to $+\infty$ whenever $p(x \,|\, \Sigma)$ is not well-defined. Then, the *score function* of $p(x \,|\, \Sigma)$ with respect to Σ is $s(x) \coloneqq (\nabla f_x)(\mathrm{vech}(\Sigma))$. Here we use extra parentheses to make it clear that the latter expression if the gradient of f_x evaluated at $\mathrm{vech}(\Sigma)$, not the gradient of $f \circ \mathrm{vech}$ evaluated at Σ. One issue that could slightly complicate the computation of the score is that we know the formulas for $p(x \,|\, \Sigma)$ in matrix notation, and differentiating with respect to each entry individually could be cumbersome. Moreover, as demonstrated by Srinivasan and Panda (2023), for any $i, j \in [d]$, we probably have $s(x)_{\{i,j\}} = (\nabla f_x)(\mathrm{vech}(\Sigma))_{\{i,j\}} \neq \nabla F_x(\Sigma)_{ij}$ where

$$F_x(\Sigma) \coloneqq \ln p(x \,|\, \Sigma) \qquad \forall \Sigma \in \mathbb{R}^{d \times d}, \qquad \forall x \in \mathbb{R}^d$$

and the gradient $\nabla F_x(\Sigma)$ does not take into account symmetry (that is, it is the gradient of F_x as a function from $\mathbb{R}^{d \times d}$ to $\mathbb{R}$). In the literature symmetry has been taken into account in different ways that may disagree with each other, and we refer the interested reader to the discussion in Srinivasan and Panda (2023).

Nonetheless, the identity in (16) allows us to compute the symmetric gradient using matrix calculus rules, and Lemma C.1 allows us to compute the actual score function from the symmetric gradient. In the next proposition, we show that the symmetric gradient and the classical gradient (that does not take into account symmetry) luckily agree in your case and have a simple formula.

Proposition C.2 *Let $\Sigma \in \mathbb{S}^d_+$ be positive definite and $x \in \mathbb{R}^d$. Then,*

$$\nabla_{\mathrm{sym}} F_x(\Sigma) = \nabla F_x(\Sigma) = \frac{1}{2}(\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1}).$$

Proof First, note that

$$\ln p(x \,|\, \Sigma) = -\frac{1}{2} x^{\mathsf{T}} \Sigma^{-1} x - \frac{1}{2} \ln \det \Sigma - \frac{d}{2} \ln 2\pi.$$

Therefore, to compute the score function $\nabla_\Sigma \ln p(x \,|\, \Sigma)$ it suffices to compute the gradient of each of the terms above individually. The last term is constant with respect to Σ, so its derivative is zero. Moreover, can compute the gradients (with respect to Σ) of each term above. Namely,by equations 57 and 61 of Petersen and Pedersen (2008) we have

$$\nabla\big(x^{\mathsf{T}} \Sigma^{-1} x\big) = -\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} \qquad \text{and} \qquad \nabla(\ln \det \Sigma) = \Sigma^{-1}.$$

Putting everything together yields the formula we desired. Moreover, since the gradient is already symmetric, the identity (16) yields the first equation in the claim. ■

C.3. Fisher Information Derivation

In the previous section we have obtained a formula for the "matrix score function" which, thanks to Lemma C.1, allows us to obtain a formula for the actual score function. In this section we shall derive a formula for the Fisher information matrix. As mentioned at the beginning of this section, although the formula for the Fisher information is known in the literature, we did not find a direct proof from the definition of

Fisher information. Moreover, since the use of derivatives and gradients of symmetric gradient has not always been rigorous in previous work, we include a rigorous derivation of the formula.

The *Fisher information matrix* of the Gaussian distribution $\mathcal{N}(0, \Sigma)$ with respect to $\Sigma \in \mathbb{S}_+^d$ is the $(d(d+1)/2) \times (d(d+1)/2)$ dimensional matrix given by

$$\mathcal{I} \equiv \mathcal{I}(\Sigma) \coloneqq \mathrm{E}\left[(\nabla f_x)(\mathrm{vech}(\Sigma))(\nabla f_x)(\mathrm{vech}(\Sigma))^\mathsf{T} \right] \quad \text{where } x \sim \mathcal{N}(0, \Sigma), \tag{18}$$

where f_x is the log-density defined in (17). To compute a formula for $\mathcal{I}$, we will need the following corollary of Isserli's Formula (Isserlis, 1918).

Lemma C.3 *For any $B \in \mathbb{R}^{d\times d}$ and $x \sim \mathcal{N}(0, \Sigma)$ we have*

$$\mathrm{E}\left[xx^\mathsf{T} B xx^\mathsf{T} \right] = \Sigma(B + B^\mathsf{T})\Sigma + \Sigma\,\mathrm{Tr}(B\Sigma).$$

Proof By Isserli's Theorem (Isserlis, 1918), for any $i, j, r, s \in [d]$ we have

$$\mathrm{E}\left[x_i x_r x_s x_j \right] = \Sigma_{ir}\Sigma_{sj} + \Sigma_{is}\Sigma_{rj} + \Sigma_{ij}\Sigma_{rs}.$$

Therefore, for any $i, j \in [d]$, we have

$$\begin{aligned}\mathrm{E}\left[xx^\mathsf{T} B xx^T \right]_{ij} &= \sum_{r,s\in[d]} \mathrm{E}\left[x_i x_r B_{r,s} x_s x_j \right] = \sum_{r,s\in[d]} B_{rs}(\Sigma_{ir}\Sigma_{sj} + \Sigma_{is}\Sigma_{rj} + \Sigma_{ij}\Sigma_{rs}) \\ &= \sum_{r,s\in[d]} B_{rs}(\Sigma_{ir}\Sigma_{sj} + \Sigma_{is}\Sigma_{rj}) + \Sigma_{ij} \underbrace{\sum_{r,s\in[d]} B_{rs}\Sigma_{rs}}_{=\mathrm{Tr}(B\Sigma^\mathsf{T})=\mathrm{Tr}(B\Sigma)} .\end{aligned}$$

Finally, if $e_i \in \{0,1\}^d$ denotes the indicator vector given by $e_i(j) \coloneqq \mathbf{1}[i = j]$ for any $j \in [d]$, then

$$\sum_{r,s\in[d]} B_{rs}(\Sigma_{ir}\Sigma_{sj} + \Sigma_{is}\Sigma_{rj}) = \sum_{r,s\in[d]} \Sigma_{is}(B_{sr} + B_{rs})\Sigma_{rj} = e_i^\mathsf{T}\Sigma(B + B^\mathsf{T})\Sigma e_j^\mathsf{T},$$

which concludes the proof of the desired identity. ∎

Finally, let us derive the formula for the Fisher information matrix $\mathcal{I}$. We note that the next theorem can be found in the literature (e.g., Magnus and Neudecker 1980, Lemma 5.2). Yet, most of the derivations rely on looking at the second derivate of the log-density, which requires care to properly take symmetry into account. Ours is a direct derivation from the definition of Fisher information. In the next proposition, $A \otimes B$ denotes the *Kronecker product* of the matrices A and B with appropriate dimensions.

Proposition C.4 *Let $\Sigma \in \mathbb{S}_+^d$ be non-singular. Then,*

$$\mathcal{I}(\Sigma) = \frac{1}{2} D^\mathsf{T}(\Sigma^{-1} \otimes \Sigma^{-1})D$$

Proof Let $x \sim \mathcal{N}(0, \Sigma)$. Define the (matrix) score function $S(x) \coloneqq \nabla_\Sigma(\ln p(x \mid \Sigma))$, by Proposition C.2, we have $S(x) = \frac{1}{2}(\Sigma^{-1}xx^\mathsf{T}\Sigma^{-1} - \Sigma^{-1})$. Note, however, that the definition of Fisher information depends on the vector score $s(x) \coloneqq \nabla f_x(\Sigma)$. By Lemma C.1, we have (abusing notation when indexing $s(x)$)

$$s(x)_{ij} = (1 + \mathbf{1}[i \neq j])S(x)_{ij}, \qquad \forall i, j \in [d]$$

Therefore, for any $i, j, r, s \in [d]$ we have

$$\begin{aligned}\mathcal{I}_{ij,rs} &= \mathrm{E}\left[\, s(x)_{ij} s(x)_{rs} \,\right] = (1 + \mathbf{1}[i \neq j])(1 + \mathbf{1}[r \neq s])\, \mathrm{E}\left[\, S(x)_{ij} S(x)_{rs} \,\right] \\ &= \frac{(1 + \mathbf{1}[i \neq j])(1 + \mathbf{1}[r \neq s])}{4} \cdot \mathrm{E}\left[\, (\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1})_{ij} (\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1})_{rs} \,\right].\end{aligned}$$

Let $e_i \in \{0,1\}^m$ (for dimension $m > 0$ clear from context) denote the indicator vector given by $e_i(j) :- \mathbf{1}[i = j]$ for any $j \in [m]$. For ease of notation, define $\Psi := \Sigma^{-1}$ and $\psi_i := \Sigma^{-1} e_i$ for each $i \in [d]$. Then we have

$$\begin{aligned}&\mathrm{E}\left[\, (\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1})_{ij} (\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1})_{rs} \,\right] \\ =&\, \mathrm{E}\left[\, (\psi_i^{\mathsf{T}} x x^{\mathsf{T}} \psi_j - \Psi_{ij})(\psi_r^{\mathsf{T}} x x^{\mathsf{T}} \psi_s - \Psi_{rs}) \,\right] \\ =&\, \mathrm{E}\left[\, \psi_i^{\mathsf{T}} x x^{\mathsf{T}} \psi_j \psi_r^{\mathsf{T}} x x^{\mathsf{T}} \psi_s \,\right] - \Psi_{ij} \cdot \mathrm{E}\left[\, \psi_r^{\mathsf{T}} x x^{\mathsf{T}} \psi_s \,\right] - \Psi_{rs} \cdot \mathrm{E}\left[\, \psi_i^{\mathsf{T}} x x^{\mathsf{T}} \psi_j \,\right] + \Psi_{ij} \Psi_{rs} \\ =&\, \psi_i^{\mathsf{T}}\, \mathrm{E}\left[\, x x^{\mathsf{T}} \psi_j \psi_r^{\mathsf{T}} x x^{\mathsf{T}} \,\right] \psi_s - \Psi_{ij} \Psi_{rs},\end{aligned}$$

where in the last equation we used that $\mathrm{E}[\, x x^{\mathsf{T}} \,] = \Sigma$ and the fact that $\psi_p^{\mathsf{T}} \Sigma \psi_q = \Psi_{pq}$ for any $p, q \in [d]$. By Lemma C.3 with $B = \psi_j \psi_r^{\mathsf{T}}$ yields

$$\begin{aligned}\psi_i^{\mathsf{T}}\, \mathrm{E}\left[\, x x^{\mathsf{T}} \psi_j \psi_r^{\mathsf{T}} x x^{\mathsf{T}} \,\right] \psi_s &= \psi_i^{\mathsf{T}} (\Sigma(\psi_j \psi_r^{\mathsf{T}} + \psi_r \psi_j^{\mathsf{T}})\Sigma + \Sigma \operatorname{Tr}(\psi_j \psi_r^{\mathsf{T}} \Sigma)) \psi_s \\ &= \psi_i^{\mathsf{T}} (\Sigma(\psi_j \psi_r^{\mathsf{T}} + \psi_r \psi_j^{\mathsf{T}})\Sigma + \Sigma \Psi_{rj}) \psi_s \\ &= \Psi_{ij} \Psi_{rs} + \Psi_{ir} \Psi_{js} + \Psi_{is} \Psi_{rj}.\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}&\mathrm{E}\left[\, (\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1})_{ij} (\Sigma^{-1} x x^{\mathsf{T}} \Sigma^{-1} - \Sigma^{-1})_{rs} \,\right] \\ &= \Psi_{ij} \Psi_{rs} + \Psi_{ir} \Psi_{js} + \Psi_{is} \Psi_{rj} - \Psi_{ij} \Psi_{rs} \\ &= \Psi_{ir} \Psi_{js} + \Psi_{is} \Psi_{rj} \\ &= (\Psi \otimes \Psi)_{ij,rs} + (\Psi \otimes \Psi)_{ij,sr} \\ &= \frac{1}{2} \mathrm{vec}\Big(e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}}\Big)^{\mathsf{T}} (\Psi \otimes \Psi) \mathrm{vec}\Big(e_r e_s^{\mathsf{T}} + e_s e_r^{\mathsf{T}}\Big)\end{aligned}$$

where in the last equation we used that Ψ is symmetric and, thus, $\Psi_{ir} \Psi_{js} + \Psi_{is} \Psi_{rj} = \Psi_{jr} \Psi_{is} + \Psi_{js} \Psi_{ri}$. Finally, assuming without loss of generality that $i \leq j$, the claim follows since

$$\frac{(1 + \mathbf{1}[i \neq j])}{2} \mathrm{vec}(e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}}) = \mathrm{vec}(e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}} - e_i e_j \odot I) = D\, \mathrm{vech}(e_i e_j^{\mathsf{T}}),$$

where $\odot$ the last equation follows from the definition of the duplication matrix and since $i \leq j$ (see Magnus and Neudecker, 1980, Def. 3.2a) ■

Lemma C.5 *Let $\Sigma \succ 0$. Then $\mathcal{I}(\Sigma) = \frac{1}{2} D^{\mathsf{T}} \Sigma^{-1} \otimes \Sigma^{-1} D$ where $\otimes$ denotes the Kronecker product between matrices. In particular, we have $\lambda_{\max}(\mathcal{I}(\Sigma)) \leq \lambda_{\min}(\Sigma)^{-2}$.*

Proof From Proposition C.4 we have

$$\mathcal{I}(\Sigma) = \frac{1}{2} D^{\mathsf{T}} \Sigma^{-1} \otimes \Sigma^{-1} D.$$

Moreover, one may note that $\|Dx\|_2^2 \leq 2\,\|x\|_2^2$ for any $x \in \mathbb{R}^{d(d+1)/2}$. Therefore,

$$\begin{aligned} \lambda_{\max}(\mathcal{I}(\Sigma)) &= \max\left\{\ \tfrac{1}{2} x^T D^{\mathsf{T}} \Sigma^{-1} \otimes \Sigma^{-1} D x \ :\ x \in \mathbb{R}^{d(d+1)/2}, \|x\|_2 \leq 1\ \right\} \\ &\leq \max\left\{\ z^T \Sigma^{-1} \otimes \Sigma^{-1} z \ :\ x \in \mathbb{R}^{d(d+1)/2}, \|z\|_2 \leq 1\ \right\} = \lambda_{\max}(\Sigma^{-1} \otimes \Sigma^{-1}) \end{aligned}$$

Finally, the result then follows by Horn and Johnson (1991, Thm. 4.2.12) which shows that the set of all eigenvalues of $\Sigma^{-1} \otimes \Sigma^{-1}$ is $\{\ \lambda_i(\Sigma^{-1}) \cdot \lambda_j(\Sigma^{-1}) \ :\ i, j \in [d]\ \}$. ■

Appendix D. Mathematical Background

D.1. Results from Probability and Statistics

We will use the following tail bound for the χ^2 distribution.

Lemma D.1 (Laurent and Massart, 2000, Lemma 1) *Let $Z \sim \chi^2(d)$ and $x > 0$. Then*

$$\Pr\left[\ Z - d \geq 2\sqrt{dx} + 2x\ \right] \leq e^{-x}.$$

We now derive a corollary of this bound that will be more convenient for our purposes.

Corollary D.2 *Let $Z \sim \chi^2(d)$ and $x > 0$. Then*

$$\Pr\left[\ Z \geq \sqrt{8d^2 + 18x^2}\ \right] \leq e^{-x}.$$

Proof We have

$$\begin{aligned} 0 \ \leq\ (2d - 3x)^2 \ &=\ 4d^2 - 12dx + 9x^2 \\ &=\ (8d^2 + 18x^2) - (4d^2 + 12dx + 9x^2) \\ &=\ (8d^2 + 18x^2) - (2d + 3x)^2. \end{aligned}$$

Rearranging and using the AM-GM inequality, we have

$$(8d^2 + 18x^2)^{1/2} \ \geq\ 2d + 3x \ =\ d + 2x + (d + x) \ \geq\ d + 2x + 2\sqrt{dx}.$$

Thus, $\Pr\left[\ Z \geq (8d^2 + 18x^2)^{1/2}\ \right] \leq \Pr\left[\ Z \geq d + 2\sqrt{dx} + 2x\ \right] \leq e^{-x}$ by Lemma D.1. ■

D.2. Properties of Wishart matrices

In this section we collect known results for Wishart matrices and derive corollaries for the normalized Wishart distribution as the one of Σ in (4). First, for the lower bound in Section 3 we shall need a few properties of the inverse Wishart distribution, which are collected in the following lemma.

Lemma D.3 (Haff, 1979, Theorem 3.2) *Let $\Sigma \sim \mathcal{W}_d(D; V)$ for some non-singular $V \in \mathbb{S}_+^d$ and with $D > d + 3$. Then $\mathrm{E}\left[\ \Sigma^{-1}\ \right] = \frac{1}{D-d-1} V^{-1}$. Moreover, for every distinct $i, j \in [d]$ we have*

$$\begin{aligned} \mathrm{Var}\left[\ \Sigma_{ii}^{-1}\ \right] &= \frac{2(V_{ii}^{-1})^2}{(D-d-1)^2(D-d-3)} \quad \textit{and} \\ \mathrm{Var}\left[\ \Sigma_{ij}^{-1}\ \right] &= \frac{(D-d+1)(V_{ij}^{-1})^2 + (D-d-1)V_{ii}^{-1}V_{jj}^{-1}}{(D-d-1)^2(D-d-3)(D-d)}. \end{aligned}$$

Let us know collect a few facts about the normalized Wishart distribution, mainly regarding concentrations of its eigenvalues.

Proposition D.4 *Let* $\Sigma \coloneqq \frac{1}{D} G G^{\mathsf{T}}$ *with the entries of* $G \in \mathbb{R}^{d \times D}$ *being i.i.d. standard Gaussians. Then* $\mathrm{E}\left[\, \Sigma \,\right] = I$ *and* $\mathrm{E}\left[\, \|\Sigma - I\|_F^2 \,\right] = d^2/D$*. In particular, if* $D = 2d$ *then* $\mathrm{E}\left[\, \|\Sigma - I\|_F^2 \,\right] = d/2$.

Proof Let $g_i \sim \mathcal{N}(0, I)$ be the i-th row of G. Then $\Sigma_{ij} = \frac{1}{D}\langle g_i, g_j \rangle$. Since g_i is independent of g_j for $i \neq j$, one can see that $\mathrm{E}\left[\, \Sigma \,\right]_{ij} = I$. Moreover,

$$\mathrm{E}\left[\, \|\Sigma - I\|_F^2 \,\right] = \sum_{i,j \in [d]} \mathrm{E}\left[\, \frac{1}{D}(\langle g_i, g_j \rangle - \mathbb{1}\mathbf{1}[i = j])^2 \,\right] = \frac{1}{D^2} \sum_{i,j \in [d]} \mathrm{E}\left[\, (\langle g_i, g_j \rangle - D\mathbb{1}\mathbf{1}[i = j])^2 \,\right]$$

If $i = j$, then $\mathrm{E}\left[\, (\langle g_i, g_j \rangle - D\mathbb{1}\mathbf{1}[i = j])^2 \,\right] = D$ since it is exactly the variance of a $\chi^2(D)$ distribution. For $i \neq j$, we have

$$\mathrm{E}\left[\, \langle g_i, g_j \rangle^2 \,\right] = \sum_{k=1}^{D} \underbrace{\mathrm{E}\left[\, g_i(k)^2 g_j(k)^2 \,\right]}_{=1 \cdot 1 = 1} + \sum_{r,s \in [D], r \neq s}^{d} \underbrace{\mathrm{E}\left[\, g_i(r) g_j(r) g_i(s) g_j(s) \,\right]}_{=0} = D.$$

Therefore, $\mathrm{E}\left[\, \|\Sigma - I\|_F^2 \,\right] = d^2 D / D^2 = d^2/D$ as desired. ■

The following theorem gives us tail bounds on the singular values of a random Gaussian matrix, which will yield sub-exponential tails for $\lambda_{\max}(\Sigma)$.

Theorem D.5 (Wainwright 2019, Thm. 6.1) *Let* $\Sigma \in \mathbb{R}^{d \times d}$ *be positive definite and let* G *be a* $d \times D$ *random matrix with i.i.d. columns each with distribution* $\mathcal{N}(0, \Sigma)$*. Then, for all* $\delta > 0$,

$$\Pr\left[\, \frac{\sigma_{\max}(G^{\mathsf{T}})}{\sqrt{D}} \geq \lambda_{\max}(\Sigma^{1/2})(1 + \delta) + \sqrt{\frac{\mathrm{Tr}(\Sigma^{1/2})}{D}} \,\right] \leq \exp\Big(-\frac{D}{2} \cdot \delta^2\Big),$$

where $\sigma_{\max}(G^{\mathsf{T}})$ *is the maximum singular value of* G^{T}*. Moreover, if* $D > d$*, then*

$$\Pr\left[\, \frac{\sigma_{\min}(G^{\mathsf{T}})}{\sqrt{D}} \leq \lambda_{\min}(\Sigma^{1/2})(1 - \delta) - \sqrt{\frac{\mathrm{Tr}(\Sigma^{1/2})}{D}} \,\right] \leq \exp\Big(-\frac{D}{2} \cdot \delta^2\Big),$$

where $\sigma_{\min}(G^{\mathsf{T}})$ *is the minimum singular value of* G^{T}.

Lemma D.6 *Let* G *be a* $d \times D$ *random matrix with* $D = 2d$ *and i.i.d. entries each with distribution* $\mathcal{N}(0, 1)$*. Define the matrix* $W \coloneqq \frac{1}{D} G G^{\mathsf{T}}$*. Then, for any* $\delta > 0$ *we have*

$$\Pr\left[\, \lambda_{\max}(\Sigma) \geq 6 + 2\delta^2 \,\right] \leq \Pr\left[\, \lambda_{\max}(\Sigma) \geq \left(1 + \frac{1}{\sqrt{2}} + \delta\right)^2 \,\right] \leq \exp(-d\delta^2). \tag{19}$$

In particular, we have

$$\Pr\left[\, \lambda_{\max}(\Sigma) \geq 10 \,\right] \leq e^{-2d} \leq \frac{1}{\sqrt{d}} e^{-d}.$$

Proof The first inequality in (19) holds since $(1+1/\sqrt{2}+\delta^2) \leq 2(1+1/\sqrt{2})^2+2\delta^2 \leq 6+2\delta^2$. The second inequality in (19) follows directly from Theorem D.5 by noticing that each column of G has distribution $\mathcal{N}(0, I)$. Thus, since $I^{1/2} = I$ and $\operatorname{Tr}(I) = d = D/2$, we have

$$\Pr\left[\lambda_{\max}(\Sigma) \geq \left(1+\frac{1}{\sqrt{2}}+\delta\right)^2\right] = \Pr\left[\left(\frac{1}{\sqrt{D}} \cdot \sigma_{\max}(G^{\mathsf{T}})\right)^2 \geq \left(\left(1+\sqrt{\frac{\operatorname{Tr}(I^{1/2})}{D}}\right)+\delta\right)^2\right].$$

In particular, define $\delta \coloneqq \sqrt{10}-1-1/\sqrt{2}$. Then, using the last inequality and the fact that $\delta^2 \leq 2$,

$$\Pr\left[\,\lambda_{\max}(\Sigma)\,\right] = \Pr\left[\lambda_{\max}(\Sigma) \geq \left(1+\frac{1}{\sqrt{2}}+\delta\right)^2\right] \leq \exp(-d\delta^2) \leq e^{-2d},$$

as desired. ■

For our purposes, the tails bounds from Theorem D.5 are too weak to usefully bound $\lambda_{\min}(\Sigma)$. The reason for that is that the tail bound on $\Pr\left[\,\lambda_{\min}(\Sigma) < t\,\right]$ does not vanish as t goes to 0, although we know $\lambda_{\min}(\Sigma) > 0$ almost surely. In other words, we want to provide tail bounds on $1/\lambda_{\min}(\Sigma)$. Thus, to better control $\lambda_{\min}(\Sigma)$ and the condition number $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ we will use results by Chen and Dongarra (2005).

Lemma D.7 (Chen and Dongarra 2005, Lemma 4.1) *Let W be a $d \times d$ Wishart matrix with D degrees of freedom. Then, for any $t > 0$*

$$\Pr\left[\lambda_{\min}(W) \leq \frac{D}{t^2}\right] < \frac{1}{\Gamma(D-d+2)}\left(\frac{D}{t}\right)^{D-d+1} = \frac{1}{(D-d+1)!}\left(\frac{D}{t}\right)^{D-d+1}.$$

Lemma D.8 *Define $\Sigma \coloneqq \frac{1}{D}GG^{\mathsf{T}}$ where G is a $d \times D$ random matrix with i.i.d. standard Gaussian entries with $D = 2d$ and $d \geq 10$. Then*

$$\Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t\right] \leq \frac{(2e)^{d+1}}{\sqrt{2\pi d}}\frac{1}{t^{d+1}}. \tag{20}$$

In particular, $\Pr\left[\,\lambda_{\min}(\Sigma) < 0.09\,\right] \leq d^{-1/2}2^{-d}$. *Moreover, we have*

$$\mathrm{E}\left[\frac{1}{\lambda_{\min}(\Sigma)}\right] \leq 2e+1 \leq 6.5.$$

Proof Since $D \cdot \Sigma = GG^{\mathsf{T}}$ follows a Wishart distribution with D degrees of freedom, Lemma D.7 yields for any $t > 0$ the bound

$$\Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t\right] \leq \frac{1}{(D-d+1)!}\left(\frac{D}{t}\right)^{D-d+1} = \frac{1}{(d+1)!}\left(\frac{2d}{t}\right)^{d+1}.$$

Using a non-asymptotic estimate of Stirling's approximation, we have

$$(d+1)! \geq \sqrt{2\pi(d+1)}\left(\frac{d+1}{e}\right)^{d+1} \geq \sqrt{2\pi d}\left(\frac{d}{e}\right)^{d+1}.$$

Therefore,

$$\Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t\right] \leq \frac{1}{\sqrt{2\pi d}}\left(\frac{2ed}{d+1}\right)^{d+1}\frac{1}{t^{d+1}} \leq \frac{1}{\sqrt{2\pi d}}(2e)^{d+1}\frac{1}{t^{d+1}},$$

which proves (20). In particular, we have

$$\Pr\left[\,\lambda_{\min}(\Sigma) \leq 0.09\,\right] \leq \Pr\left[\lambda_{\min}(\Sigma) \leq \frac{1}{4e}\right] \leq \frac{1}{\sqrt{2\pi d}}\left(\frac{2e}{4e}\right)^{d+1} = \frac{1}{\sqrt{2\pi d}} - 2^{-(d+1)} \leq \frac{1}{\sqrt{d}}2^{-d}.$$

Finally, we can use (20) to upper bound $\mathrm{E}\left[\,1/\lambda_{\min}(\Sigma)\,\right]$ by

$$\mathrm{E}\left[\frac{1}{\lambda_{\min}(\Sigma)}\right] \leq \int_0^\infty \Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t\right] \mathrm{d}t \leq 2e + \int_{2e}^\infty \Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t\right] \mathrm{d}t.$$

For the last integral in the right-hand side, we have

$$\int_{2e}^\infty \Pr\left[\frac{1}{\lambda_{\min}(\Sigma)} \geq t\right] \mathrm{d}t \leq \int_{2e}^\infty \frac{1}{\sqrt{2\pi d}}(2e)^{d+1}\frac{1}{t^{d+1}}\,\mathrm{d}t = \int_1^\infty \frac{1}{\sqrt{2\pi d}}\frac{1}{y^{d+1}} \cdot 2e\,\mathrm{d}y = \frac{2e}{\sqrt{\pi d}}\frac{1}{d} \leq 1,$$

where in the last inequality we used $d \geq 10 \geq 2e$. ■

Lemma D.9 (Chen and Dongarra 2005, Theorem 4.5) *Let W be a $d \times d$ Wishart matrix with D degrees of freedom. Then, there is a constant $C \leq 6.414$ independent of d and D such that, for any $t > 0$,*

$$\Pr\left[\sqrt{\frac{\lambda_{\max}(W)}{\lambda_{\min}(W)}} > \frac{D}{D-d+1}\cdot t\right] < \frac{1}{\sqrt{2\pi}}\left(\frac{C}{t}\right)^{D-d+1}.$$

Lemma D.10 *Let Σ follow a normalized Wishart distribution as in (4). Then*

$$\Pr\left[\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} > t\right] < \frac{(13)^{d+1}}{\sqrt{2\pi}} \cdot \frac{1}{t^{(d+1)/2}}$$

Proof Define $\kappa(\Sigma) \coloneqq \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. Since $D = 2d$, we have $D/(D-d+1) = 2d/(d+1) \geq 2$. Thus, for all $t > 0$ it follows from Lemma D.9 that

$$\begin{aligned}\Pr\left[\,\kappa(\Sigma) > t\,\right] &= \Pr\left[\sqrt{\kappa(\Sigma)} > \frac{2d}{d+1}\cdot\frac{d+1}{2d}\cdot\sqrt{t}\right] \\ &\leq \Pr\left[\sqrt{\kappa(\Sigma)} > \frac{2d}{d+1}\cdot\frac{1}{2}\cdot\sqrt{t}\right] < \frac{1}{\sqrt{2\pi}}\left(\frac{C}{\sqrt{t/2}}\right)^{d+1},\end{aligned}$$

where $C \leq 6.414$ is as in Lemma D.9. The final bound follows by noting that $2 \cdot C \leq 13$. ■