Supplementary Material for "Sharp Statistical Guarantees for Adversarially Robust Gaussian Classification"

Abstract

This document provides the complete proofs and additional details for the main results stated in the ICML publication "Sharp Statistical Guarantees for Adversarially Robust Gaussian Classification".

Notation For positive semi-definite matrix A, we use $\|x\|_A := \sqrt{x^T A x}$. Let $\Phi(\cdot)$ the CDF of standard Gaussian distribution $\mathcal{N}(0,1)$ and $\bar{\Phi}(x) := 1 - \Phi(x)$. The notation $f(n,d) = O\big(g(n,d)\big)$ means that there exist a universal constant $c > 0$ that does not depend on the problem parameters such as n, d etc, such that $|f(n,d)| \leq c|g(n,d)|$. Similarly, we define $f(n,d) = \Omega\left(g(n,d)\right)$ when there exist constants $c_1, c_2 > 0$ such that $c_1|g(n,d)| \leq |f(n,d)| \leq c_2|g(n,d)|$. Notation O_P, Ω_P are used if the corresponding relations happen with probability converges to 1 as $n \to \infty$. We define the ℓ_p norm $\|x\|_p = (\sum_{i=1}^d x_i^p)^{1/p}$ and the corresponding ℓ_p-ball as $\{x \in^d |\|x\|_p \leq 1\}$.

A. Proof of Theorem 2.1

For completeness, in this section, we present the proof of Theorem 2.1. This result follows from combining Theorem 1, Theorem 2 and Lemma 1 in (Bhagoji et al., 2019). The proof is mainly a simplified presentation of their proofs (e.g. without using the language of optimal transport) which make some of their results explicit to interpret for our case (e.g. they did not provide the expression for optimal linear classifier, which is useful to our algorithmic results).

To start with, let us define $w_1 := \frac{w_0}{\|w_0\|_\Sigma} = \frac{\Sigma^{-1}(\mu - z_\Sigma(\mu))}{\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}}$ be the normalized version of w_0 so that $\|w_1\|_\Sigma = 1$. The following lemma is implicit in (Bhagoji et al., 2019):

Lemma A.1. *Suppose we define*

$$G(z, w) = w^T(\mu - z),$$

then $(z_\Sigma(\mu), w_1)$ *is solution of the following minimax optimization problem:*

$$\min_{\|z\|_B \leq \varepsilon} \max_{\|w\|_\Sigma \leq 1} G(z, w). \tag{14}$$

Proof. We first show that the optimal value of the inner maximization problem can be written as:

$$\max_{\|w\|_\Sigma \leq 1} w^T(\mu - z) = \|\mu - z\|_{\Sigma^{-1}}, \tag{15}$$

and the maximum is achieved when

$$w = \frac{\Sigma^{-1}(\mu - z)}{\|\mu - z\|_{\Sigma^{-1}}}. \tag{16}$$

In fact, for any w such that $\|w\|_\Sigma \leq 1$, Cauchy-Schwarz inequality gives

$$\begin{aligned} w^T(\mu - z) = (\Sigma^{1/2}w)^T \Sigma^{-1/2}(\mu - z) &\leq \|\Sigma^{1/2}w\|_2 \|\Sigma^{-1/2}(\mu - z)\|_2 \\ &= \|w\|_\Sigma \|\mu - z\|_{\Sigma^{-1}} \\ &\leq \|\mu - z\|_{\Sigma^{-1}}. \end{aligned}$$

Furthermore, it is easy to check that the choice $w = \frac{\Sigma^{-1}(\mu - z)}{\|\mu - z\|_{\Sigma^{-1}}}$ directly yields $w^T(\mu - z) = \|\mu - z\|_{\Sigma^{-1}}$ achieving the equality. Therefore we have proved (15) and (16).

Using (15), the minimax problem (14) therefore simplifies to:

$$\min_{\|z\|_B \le \varepsilon} \|\mu - z\|_{\Sigma^{-1}}.$$

Recall that we define $z_\Sigma(\mu)$ (cf. (4)) as

$$z_\Sigma(\mu) = \operatorname*{argmin}_{\|z\|_B \le \varepsilon} \|\mu - z\|^2_{\Sigma^{-1}},$$

which is the optimal solution to this outer minimization problem. Combining with the optimality condition for the inner maximization (16), we conclude that $(z_\Sigma(\mu), w_1)$ is solution of the minimax problem (14) and complete the proof. □

Corollary A.1. *The following relation is satisfied for quantities w_1 and $z_\Sigma(\mu)$:*

$$w_1^T\mu - \varepsilon\|w_1\|_{B*} = \|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}.$$

Proof. Since $G(z, w)$ is linear in both z and w and both constraint sets $\{\|z\|_B \le \varepsilon\}$ and $\{\|w\|_\Sigma \le 1\}$ are convex, the minimax problem (14) satisfies strong duality by Von Neumann's Minimax Theorem. In other words, we can switch the order of the min and max, namely,

$$\min_{\|z\|_B \le \varepsilon} \max_{\|w\|_\Sigma \le 1} G(z, w) = \max_{\|w\|_\Sigma \le 1} \min_{\|z\|_B \le \varepsilon} G(z, w),$$

and $(z_\Sigma(\mu), w_1)$ is the solution to both sides. By the stationary condition of the minimax problem,

$$z_\Sigma(\mu) = \operatorname*{argmin}_{\|z\|_B \le \varepsilon} G(z, w_1).$$

By the definition of dual norm, we also have

$$\min_{\|z\|_B \le \varepsilon} G(z, w_1) = \min_{\|z\|_B \le \varepsilon} w_1^T(\mu - z) = w_1^T\mu - \varepsilon\|w_1\|_{B*}.$$

Hence,

$$\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}} = G(z_\Sigma(\mu), w_1) = \min_{\|z\|_B \le \varepsilon} G(z, w_1) = w_1^T\mu - \varepsilon\|w_1\|_{B*}.$$

Thus we completed the proof. □

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. The proof can be divided into two parts:

1. Show that f_{w_0} has robust risk $R^{B,\varepsilon}_{\mu,\Sigma}(f_{w_0}) = \bar{\Phi}(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}})$.
2. Show that no classifier can achieve robust risk smaller than $\bar{\Phi}(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}})$.

The first part is a consequence of Corollary A.1. In order to see this, we first note that since w_1 is a rescaling of w_0, the induced linear classifiers are the same, hence,

$$R^{B,\varepsilon}_{\mu,\Sigma}(f_{w_0}) = R^{B,\varepsilon}_{\mu,\Sigma}(f_{w_1}).$$

By Lemma 6.2, the robust risk of f_{w_1} is

$$R^{B,\varepsilon}_{\mu,\Sigma}(f_{w_1}) = \bar{\Phi}(\frac{w_1^T\mu - \varepsilon\|w_1\|_{B*}}{\|w_1\|_\Sigma}) = \bar{\Phi}(w_1^T\mu - \varepsilon\|w_1\|_{B*}).$$

By Corollary A.1,

$$\bar{\Phi}(w_1^T\mu - \varepsilon\|w_1\|_{B*}) = \bar{\Phi}(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}).$$

Therefore, we have proved the first part.

For the second part, we invoke Lemma 6.4. By setting $\mu' = \mu - z_\Sigma(\mu)$ in Lemma 6.4, we have that for any classifier f,

$$R^{B,\varepsilon}_{\mu,\Sigma}(f) \geq R^{\text{std}}_{\mu - z_\Sigma(\mu),\Sigma}(f).$$

We also know that no classifier can achieve standard risk smaller than the Bayes Risk in $P_{\mu - z_\Sigma(\mu),\Sigma}$. Recall that for a conditional Gaussian kmodel $P_{\mu',\Sigma}$, the standard Bayes Risk is $\bar{\Phi}(\|\mu'\|_{\Sigma^{-1}})$. In other words, for any classifier f, we have

$$R^{\text{std}}_{\mu - z_\Sigma(\mu),\Sigma}(f) \geq \bar{\Phi}(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}).$$

Combining the two inequalities, we conclude that

$$R^{B,\varepsilon}_{\mu,\Sigma}(f) \geq \bar{\Phi}(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}) \tag{17}$$

holds for all classifiers f. Therefore, we prove the second part and thus complete the proof. □

B. Proof of Proposition 5.1

Proof of Proposition 5.1. Recall that the setting of interest here is $\Sigma = I$ and $\|\cdot\|_B$ corresponds to the ℓ_2 norm. In this setting, we show that $z_\Sigma(\mu)$ has a simplified form. In fact, directly invoking

$$z_\Sigma(\mu) = \operatorname*{argmin}_{\|z\|_B \leq \varepsilon} \|\mu - z\|^2_{\Sigma^{-1}} = \operatorname*{argmin}_{\|z\|_2 \leq \varepsilon} \|\mu - z\|^2_2,$$

gives $z_\Sigma(\mu) = \min(\varepsilon, \|\mu\|_2)\frac{\mu}{\|\mu\|_2}$, and

$$\mu - z_\Sigma(\mu) = \max(0, \frac{\|\mu\|_2 - \varepsilon}{\|\mu\|_2})\mu.$$

From this expression, we can see that when $\varepsilon > \|\mu\|_2$, the Adversarial Signal-to-Noise Ratio of $P_{\mu,\Sigma}$ is $2\|\mu - z_\Sigma(\mu)\|_2 = 0$. Hence, no classifier can achieve accuracy better than $\frac{1}{2}$. Below we only consider the case when $\varepsilon < \|\mu\|_2$.

Recall that we want to compare the minimax rate in adversarial and standard setting. As we showed earlier, the minimax rates are $O(\exp(-\frac{1}{2}\|\mu - z_\Sigma(\mu)\|_2^2)\frac{d}{n})$ and $O(\exp(-\frac{1}{2}\|\mu\|_2^2)\frac{d}{n})$ respectively. The ratio between the two quantities equals to:

$$\frac{\exp(-\frac{1}{2}\|\mu - z_\Sigma(\mu)\|_2^2)\frac{d}{n}}{\exp(-\frac{1}{2}\|\mu - z_\Sigma(\mu)\|_2^2)\frac{d}{n}} = \exp(\frac{1}{2}((\|\mu\|_2 - \varepsilon)^2 - \|\mu\|_2^2)) = \exp(\varepsilon\|\mu\|_2 - \frac{1}{2}\varepsilon^2). \tag{18}$$

Since $0 \leq \varepsilon < \|\mu\|_2$, we have

$$\varepsilon\|\mu\|_2 - \frac{1}{2}\varepsilon^2 = \varepsilon(\|\mu\|_2 - \frac{1}{2}\varepsilon) \in \left[\frac{1}{2}\varepsilon\|\mu\|_2, \varepsilon\|\mu\|_2\right].$$

Equipped with the above relation, we are in the position of establishing Proposition 5.1.

- When $\varepsilon \leq O(\frac{1}{\|\mu\|_2})$, one has

  $$\varepsilon\|\mu\|_2 - \frac{1}{2}\varepsilon^2 \leq \varepsilon\|\mu\|_2 \leq O(1),$$

  thereby, the adversarial rate is at most $\exp(O(1)) = O(1)$ times slower than the standard rate.

- When $\|\mu\|_2 \geq \Omega(\log d)$ and $\varepsilon \geq \Omega(\frac{\log d}{\|\mu\|_2})$, we conclude

  $$\varepsilon\|\mu\|_2 - \frac{1}{2}\varepsilon^2 \geq \frac{1}{2}\varepsilon\|\mu\|_2 \geq \Omega(\log d),$$

  the adversarial rate can be slower than the standard rate by an $\Omega(\exp(\log d)) = \Omega(poly(d))$ factor.

- When $\|\mu\|_2 \geq \Omega(\sqrt{d})$ and $\varepsilon \geq \Omega(\frac{d}{\|\mu\|_2})$, it is guaranteed that

  $$\varepsilon\|\mu\|_2 - \frac{1}{2}\varepsilon^2 \geq \frac{1}{2}\varepsilon\|\mu\|_2 \geq \Omega(d),$$

  therefore, the adversarial rate can be slower than the standard rate by an $\Omega(\exp(d))$ factor.

□

C. Improved analysis when Σ is known

Meticulous readers may find a tiny gap between our bounds: the upper bound in Theorem 3.1 is $O_P\left(e^{-\frac{1}{8}r^2} \cdot r \cdot \frac{d}{n}\right)$, while the lower bound above gives $\Omega_P\left(e^{-\frac{1}{8}r^2} \cdot \frac{1}{r} \cdot \frac{d}{n}\right)$. Since the dominant factor is $e^{-\frac{1}{8}r^2}$ and $r = \Omega(1)$, this difference is only in a lower order term. This gap is due to the fact that (Li et al., 2017) assumed the covariance matrix Σ is known to the learner. In this section, we will prove that under the same assumption, there is a modified version of Algorithm 1 that achieves the truly optimal rate which matches the lower bound even with lower order term in r.

The only modification we made in Algorithm 1 is to replace the sample covariance matrix by the true covariance Σ. The modified algorithm is presented below in Algorithm 2.

Algorithm 2 An improved estimator for w_0 when Σ is known

```
Input: Data pairs $\{(x_i, y_i)\}_{i=1}^n$.
Output: $\widehat{w}$.
Step 1: Define $\widehat{\mu}$ and $\widehat{\Sigma}$ as
    $$\widehat{\mu} := \frac{1}{n}\sum_{i=1}^n y_i x_i, \qquad \widehat{\Sigma} := \Sigma.$$
Step 2: Solve for $\widehat{z}$ in the following
    $$\widehat{z} := z_{\widehat{\Sigma}}(\widehat{\mu}) = \operatorname*{argmin}_{\|z\|_B \le \varepsilon} \|\widehat{\mu} - z\|_{\widehat{\Sigma}^{-1}}^2.$$
Step 3: Define $\widehat{w} := \widehat{\Sigma}^{-1}(\widehat{\mu} - \widehat{z})$.
```

Theorem C.1. *For the* $(\|\cdot\|_B, \varepsilon)$ *adversary, suppose the adversarial signal-to-noise ratio* $\mathrm{AdvSNR}_{B,\varepsilon}(\mu, \Sigma) = r$*, then the excess risk of* $f_{\widehat{w}}$ *defined in Algorithm 2 is upper bounded by*

$$R^{B,\varepsilon}_{\mu,\Sigma}(f_{\widehat{w}}) - R^{B,\varepsilon}_{\mu,\Sigma}* \le O_P\left(e^{-\frac{1}{8}r^2} \cdot \frac{1}{r} \cdot \frac{d}{n}\right).$$

This improved rate can be proved by some simple modification to the proof of Theorem 3.1.

Proof. We demonstrate that in this setting, there is a stronger upper bound $\delta_n = O_P\left(\frac{1}{r} \cdot \frac{d}{n}\right)$ and the rest of proof follows the same as that of Theorem 3.1. To this end, let us recall that by Lemma 6.3 and one has the decomposition,

$$\|\widehat{w}\|_\Sigma \delta_n = \underbrace{-\frac{1}{2}\left(\|w_0\|_\Sigma - \|\widehat{w}\|_\Sigma\right)^2}_{T_1} \underbrace{+ w_0^T(\widehat{z} - z_\Sigma(\mu))}_{T_2} \underbrace{- \frac{1}{2}\|\widehat{z} - z_\Sigma(\mu)\|_{\Sigma^{-1}}^2}_{T_3} \underbrace{+ \frac{1}{2}\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|_{\Sigma^{-1}}^2}_{T_4}.$$

Similar to the proof of Theorem 3.1, we shall establish that

$$T_1 \le 0,\ T_2 \le 0,\ T_3 \le 0,\ T_4 \le O_P\left(\frac{d}{n}\right).$$

Note that the only difference here is that we can now give a tighter upper bound for T_4: $O_P\left(\frac{d}{n}\right)$ instead of $O_P\left(r^2 \frac{d}{n}\right)$.

Since $\Sigma = \hat{\Sigma}$, by Lemma 6.1, we have

$$T_4 = \frac{1}{2}\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|_{\Sigma^{-1}}^2 = \frac{1}{2}\|(\widehat{\mu} - \mu)\|_{\Sigma^{-1}}^2 = O_P\left(\frac{d}{n}\right). \tag{19}$$

Hence, we have proved that $T_4 = O_P\left(\frac{d}{n}\right)$, and

$$\delta_n = O_P\left(\frac{1}{r} \cdot \frac{d}{n}\right).$$

Therefore we have completed the proof. □

D. Proofs of auxiliary lemmas

D.1. Proof of Lemma 6.3

Proof of Lemma 6.3. Recall that our goal is to establish

$$\begin{aligned}\|\widehat{w}\|_{\Sigma}\delta_n &= \|\widehat{w}\|_{\Sigma}\|w_0\|_{\Sigma} - \left(\widehat{w}^T\mu - \varepsilon\|\widehat{w}\|_{B*}\right)\\ &= \underbrace{-\frac{1}{2}\left(\|w_0\|_{\Sigma} - \|\widehat{w}\|_{\Sigma}\right)^2}_{T_1}\underbrace{+w_0^T(\widehat{z} - z_{\Sigma}(\mu))}_{T_2}\underbrace{-\frac{1}{2}\|\widehat{z} - z_{\Sigma}(\mu)\|^2_{\Sigma^{-1}}}_{T_3}\underbrace{+\frac{1}{2}\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|^2_{\Sigma^{-1}}}_{T_4}.\end{aligned} \tag{20}$$

Since $\widehat{w} = \widehat{\Sigma}^{-1}(\widehat{\mu} - z_{\widehat{\Sigma}}(\widehat{\mu}))$, by Theorem 2.1, $f_{\widehat{w}}$ is the optimal robust classifier for $P_{\widehat{\mu},\widehat{\Sigma}}$, therefore, one can observe

$$\frac{\widehat{w}^T\widehat{\mu} - \varepsilon\|\widehat{w}\|_{B*}}{\|\widehat{w}\|_{\widehat{\Sigma}}} = \|\widehat{w}\|_{\widehat{\Sigma}}.$$

Hence, direct calculations yield

$$\begin{aligned}\|\widehat{w}\|_{\Sigma}\delta_n &= \|w_0\|_{\Sigma}\|\widehat{w}\|_{\Sigma} - \|\widehat{w}\|^2_{\widehat{\Sigma}} - \widehat{w}^T(\mu - \widehat{\mu})\\ &= \|w_0\|_{\Sigma}\|\widehat{w}\|_{\Sigma} - (\widehat{\mu} - \widehat{z})^T\widehat{\Sigma}^{-1}(\widehat{\mu} - \widehat{z}) + (\widehat{\mu} - \widehat{z})^T\widehat{\Sigma}^{-1}(\widehat{\mu} - \mu)\\ &= \|w_0\|_{\Sigma}\|\widehat{w}\|_{\Sigma} + \widehat{w}^T(\widehat{z} - \mu).\end{aligned}$$

Now by use of the relation $\mu = \Sigma w_0 + z_{\Sigma}(\mu)$, we can further obtain

$$\begin{aligned}\|\widehat{w}\|_{\Sigma}\delta_n &= \|w_0\|_{\Sigma}\|\widehat{w}\|_{\Sigma} + \widehat{w}^T(\widehat{z} - \Sigma w_0 - z_{\Sigma}(\mu))\\ &= \|w_0\|_{\Sigma}\|\widehat{w}\|_{\Sigma} - \widehat{w}^T\Sigma w_0 + \widehat{w}^T(\widehat{z} - z_{\Sigma}(\mu))\\ &= -\frac{1}{2}\left(\|w_0\|_{\Sigma} - \|\widehat{w}\|_{\Sigma}\right)^2 + \frac{1}{2}\|w_0\|^2_{\Sigma} + \frac{1}{2}\|\widehat{w}\|^2_{\Sigma} - \widehat{w}^T\Sigma w_0 + \widehat{w}^T(\widehat{z} - z_{\Sigma}(\mu))\\ &= T_1 + \frac{1}{2}(\widehat{w} - w_0)^T\Sigma(\widehat{w} - w_0) + w_0^T(\widehat{z} - z_{\Sigma}(\mu)) + (\widehat{w} - w_0)^T(\widehat{z} - z_{\Sigma}(\mu))\\ &= T_1 + \frac{1}{2}(\widehat{w} - w_0)^T\Sigma(\widehat{w} - w_0) + T_2 + (\widehat{w} - w_0)^T(\widehat{z} - z_{\Sigma}(\mu)),\end{aligned}$$

where the last equality invokes the definitions in expression (20). To finish the proof, we make the observation about $\Sigma(\widehat{w} - w_0)$ in the following

$$\begin{aligned}\Sigma(\widehat{w} - w_0) &= (\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\Sigma}\widehat{w} - \Sigma w_0)\\ &= \underbrace{(\Sigma - \widehat{\Sigma})\widehat{w}}_{U_1} + \underbrace{(\widehat{\mu} - \mu)}_{U_2} - \underbrace{(\widehat{z} - z_{\Sigma}(\mu))}_{U_3} := U_1 + U_2 - U_3.\end{aligned}$$

Therefore, putting everything together and rearranging terms, it is guaranteed that

$$\begin{aligned}\|\widehat{w}\|_{\Sigma}\delta_n &= T_1 + T_2 + \frac{1}{2}(\widehat{w} - w_0)^T\Sigma(\widehat{w} - w_0) + (\widehat{w} - w_0)^T(\widehat{z} - z_{\Sigma}(\mu))\\ &= T_1 + T_2 + \frac{1}{2}(\Sigma(\widehat{w} - w_0))^T\Sigma^{-1}(\Sigma(\widehat{w} - w_0)) + (\Sigma(\widehat{w} - w_0))^T\Sigma^{-1}(\widehat{z} - z_{\Sigma}(\mu))\\ &= T_1 + T_2 + \frac{1}{2}(U_1 + U_2 - U_3)^T\Sigma^{-1}(U_1 + U_2 - U_3) + (U_1 + U_2 - U_3)\Sigma^{-1}U_3\\ &= T_1 + T_2 + \frac{1}{2}(U_1 + U_2 - U_3)^T\Sigma^{-1}(U_1 + U_2 + U_3)\\ &= T_1 + T_2 - \frac{1}{2}U_3^T\Sigma^{-1}U_3 + \frac{1}{2}(U_1 + U_2)^T\Sigma^{-1}(U_1 + U_2)\\ &= T_1 + T_2 + T_3 + T_4.\end{aligned}$$

Thus we have finished the proof. □

References

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7496–7508. 2019.

Li, T., Yi, X., Carmanis, C., and Ravikumar, P. Minimax gaussian classification & clustering. In *Artificial Intelligence and Statistics*, pp. 1–9, 2017.