Dimension-free Analysis of Compressive Quadratic Classifiers: Supplementary Material

Efstratios Palias and Ata Kabán

1 Proofs

In this section we outline the proofs of our results in the main paper. Please refer to the paper for the notation and the statements of the theorems.

Proof of Theorem 1. We first write the singular value decomposition of $\mathbf{B}$ as

$$\mathbf{B} = \mathbf{W}\mathbf{D}\mathbf{V}^\top \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{d\times d}$ and $\mathbf{W} \in \mathbb{R}^{k\times k}$ are orthogonal, and $\mathbf{D} \in \mathbb{R}^{k\times d}$ is pseudodiagonal, containing the singular values of $\mathbf{B}$ in descending order. We can then write

$$\mathbf{B}^+\mathbf{B} = \mathbf{V}\mathbf{D}^+\mathbf{W}^\top\mathbf{W}\mathbf{D}\mathbf{V}^\top = \mathbf{V}\mathbf{D}^+\mathbf{D}\mathbf{V}^\top = \mathbf{V}\begin{bmatrix}\mathbf{I}_k & \\ & \mathbf{0}\end{bmatrix}\mathbf{V}^\top. \tag{2}$$

Since $\mathbf{A}$ is assumed low-rank, we can write its orthogonal eigendecomposition as $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{d\times s}$ is semi-orthogonal and $\mathbf{S} \in \mathbb{S}^s$ is diagonal, containing the non-zero eigenvalues of $\mathbf{A}$. We can then write

$$\mathbf{B}^+\mathbf{B}\mathbf{A}\mathbf{B}^+\mathbf{B} = \mathbf{V}\begin{bmatrix}\mathbf{I}_k & \\ & \mathbf{0}\end{bmatrix}\mathbf{V}^\top\mathbf{U}\mathbf{S}\mathbf{U}^\top\mathbf{V}\begin{bmatrix}\mathbf{I}_k & \\ & \mathbf{0}\end{bmatrix}\mathbf{V}^\top \tag{3}$$

By our assumption, the first k columns of $\mathbf{V}$ span the eigenvectors of $\mathbf{A}$ corresponding to non-zero eigenvalues. This means that the bottom $(d-k)$ rows of $\mathbf{V}^\top\mathbf{U}$ are zero. If we set its last $(d-k)$ rows to zero, nothing changes, and thus

$$\mathbf{V}\begin{bmatrix}\mathbf{I}_k & \\ & \mathbf{0}\end{bmatrix}\mathbf{V}^\top\mathbf{U} = \mathbf{V}\mathbf{V}^\top\mathbf{U} = \mathbf{U}. \tag{4}$$

Transposing both sides of (4) also implies that

$$\mathbf{U}^\top\mathbf{V}\begin{bmatrix}\mathbf{I}_k & \\ & \mathbf{0}\end{bmatrix}\mathbf{V}^\top = \mathbf{U}^\top. \tag{5}$$

Plugging (4) and (5) into (3) yields

$$\mathbf{B}^+\mathbf{B}\mathbf{A}\mathbf{B}^+\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{U}^\top = \mathbf{A}. \tag{6}$$

This completes the proof. □

Proof of Lemma 2. Let $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{Q}_{\mathbf{B}}$, and $t \in (0,1)$. Since $\mathbf{A}_1 \in \mathcal{Q}_{\mathbf{B}}$, there exists $\mathbf{A}'_1 \in \mathcal{Q}$, such that

$$(\mathbf{B}^\top)^+ \mathbf{A}'_1 \mathbf{B}^+ = \mathbf{A}_1, \tag{7}$$

and since $\mathbf{A}_2 \in \mathcal{Q}_{\mathbf{B}}$, there exists $\mathbf{A}'_2 \in \mathcal{Q}$, such that

$$(\mathbf{B}^\top)^+ \mathbf{A}'_2 \mathbf{B}^+ = \mathbf{A}_2. \tag{8}$$

We can then write

$$\begin{align} t\mathbf{A}_1 + (1-t)\mathbf{A}_2 &= t(\mathbf{B}^\top)^+ \mathbf{A}'_1 \mathbf{B}^+ (1-t)(\mathbf{B}^\top)^+ \mathbf{A}'_2 \mathbf{B}^+ \tag{9}\\ &= (\mathbf{B}^\top)^+ (t\mathbf{A}'_1 + (1-t)\mathbf{A}'_2)\mathbf{B}^+. \tag{10} \end{align}$$

But $t\mathbf{A}'_1 + (1-t)\mathbf{A}'_2 \in \mathcal{Q}$, since

$$\begin{align} \|t\mathbf{A}'_1 + (1-t)\mathbf{A}'_2\|_* &\le \|t\mathbf{A}'_1\|_* + \|(1-t)\mathbf{A}'_2\|_* \tag{11}\\ &= t\|\mathbf{A}'_1\|_* + (1-t)\|\mathbf{A}'_2\|_* \tag{12}\\ &\le ta + (1-t)a \tag{13}\\ &= a. \tag{14} \end{align}$$

Therefore, $(\mathbf{B}^\top)^+ (t\mathbf{A}'_1 + (1-t)\mathbf{A}'_2)\mathbf{B}^+ \in \mathcal{Q}_{\mathbf{B}}$, and thus, $\mathcal{Q}_{\mathbf{B}}$ is convex. □

Proof of Lemma 3. Let $\widetilde{\mathbf{I}}_k \in \mathbb{R}^{k \times d}$ be the matrix with the top-most k rows of $\mathbf{I}_d$. We first note that

$$\mathbf{\Sigma}^+ = \widetilde{\mathbf{I}}_k^\top \mathbf{\Sigma}_k^+, \tag{15}$$

and also that

$$\widetilde{\mathbf{I}}_k \mathcal{Q} \widetilde{\mathbf{I}}_k^\top = \widetilde{\mathcal{Q}}_{\mathbf{B}}, \tag{16}$$

and finally that $\widetilde{\mathcal{Q}}_{\mathbf{B}}$ is invariant to rotations, as they leave all singular values the same. We then have

$$\begin{align} \mathcal{Q}_{\mathbf{B}} &= (\mathbf{B}^\top)^+ \mathcal{Q} \mathbf{B}^+ \tag{17}\\ &= \mathbf{U}(\mathbf{\Sigma}^\top)^+ \mathbf{V}^\top \mathcal{Q} \mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^\top \tag{18}\\ &= \mathbf{U}(\mathbf{\Sigma}^\top)^+ \mathcal{Q} \mathbf{\Sigma}^+ \mathbf{U}^\top \tag{19}\\ &= \mathbf{U}\mathbf{\Sigma}_k^+ \widetilde{\mathbf{I}}_k \mathcal{Q} \widetilde{\mathbf{I}}_k^\top \mathbf{\Sigma}_k^+ \mathbf{U}^\top \tag{20}\\ &= \mathbf{U}\mathbf{\Sigma}_k^+ \widetilde{\mathcal{Q}}_{\mathbf{B}} \mathbf{\Sigma}_k^+ \mathbf{U}^\top \tag{21}\\ &= \mathbf{U}\mathbf{\Sigma}_k^+ \mathbf{U}^\top \widetilde{\mathcal{Q}}_{\mathbf{B}} \mathbf{U} \mathbf{\Sigma}_k^+ \mathbf{U}^\top \tag{22}\\ &= \mathbf{C} \widetilde{\mathcal{Q}}_{\mathbf{B}} \mathbf{C}. \tag{23} \end{align}$$

□

Proof of Lemma 4. It is trivial to show that $\mathcal{Q}_{\mathbf{B}} \subseteq \widetilde{\mathcal{Q}}_{\mathbf{B}}$, using Theorem 10. To show that $\mathcal{Q}_{\mathbf{B}} \supseteq \widetilde{\mathcal{Q}}_{\mathbf{B}}$, we let $\mathbf{A} \in \widetilde{\mathcal{Q}}_{\mathbf{B}}$ be arbitrary, and will show that $\mathbf{A} \in \mathcal{Q}_{\mathbf{B}}$. Define $\mathbf{A}' := \mathbf{B}^\top \mathbf{A} \mathbf{B}$. Since $\mathbf{B}$ is semi-orthogonal, $(d-k)$ singular values of $\mathbf{A}'$ are zero, and the others are the same as the singular values of $\mathbf{A}$. Since $\mathbf{A} \in \widetilde{\mathcal{Q}}_{\mathbf{B}}$, $\|\mathbf{A}\|_* \le a$, and thus $\|\mathbf{A}'\|_* \le a$. Therefore, $\mathbf{A}' \in \mathcal{Q}$, and thus $\mathbf{B}\mathbf{A}'\mathbf{B}^\top \in \mathcal{Q}_{\mathbf{B}}$. But $\mathbf{B}\mathbf{A}'\mathbf{B}^\top = \mathbf{B}\mathbf{B}^\top \mathbf{A} \mathbf{B}\mathbf{B}^\top = \mathbf{A}$, so $\mathbf{A} \in \mathcal{Q}_{\mathbf{B}}$. Therefore, $\mathbf{Q}_{\mathbf{B}} \supseteq \widetilde{\mathbf{Q}}_{\mathbf{B}}$. Combining the two results, we complete the proof. □

Proof of Lemma 5. Let $\mathbf{A}' \in \mathcal{Q}$, such that $\mathbf{A} = (\mathbf{B}^\top)^+\mathbf{A}'\mathbf{B}^+$. Also let $\mathbf{B}_o \in \mathbb{R}^{k\times d}$ be a semi-orthogonal matrix, such that

$$\mathbf{B}^+\mathbf{B} = \mathbf{B}_o^\top \mathbf{B}_o, \tag{24}$$

and define $\mathbf{A}_o := \mathbf{B}_o\mathbf{A}'\mathbf{B}_o^\top$, so that $\mathbf{A}_o \in \mathcal{Q}_{\mathbf{B}_o}$. For the true error, we have that

$$\begin{aligned}
L_{\mathcal{D}}^{\mathbf{B}}(\mathbf{A}) &= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\ell(X^\top\mathbf{B}^\top\mathbf{A}\mathbf{B}X, Y)] && (25)\\
&= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\ell(X^\top\mathbf{B}^\top(\mathbf{B}^\top)^+\mathbf{A}'\mathbf{B}^+\mathbf{B}X, Y)] && (26)\\
&= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\ell(X^\top(\mathbf{B}^+\mathbf{B})^\top\mathbf{A}'(\mathbf{B}^+\mathbf{B})X, Y)] && (27)\\
&= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\ell(X^\top\mathbf{B}_o^\top\mathbf{B}_o\mathbf{A}'\mathbf{B}_o^\top\mathbf{B}_oX, Y)] && (28)\\
&= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\ell(X^\top\mathbf{B}_o^\top\mathbf{A}_o\mathbf{B}_oX, Y)] && (29)\\
&= L_{\mathcal{D}}^{\mathbf{B}_o}(\mathbf{A}_o). && (30)
\end{aligned}$$

The same derivation can also be followed for the empirical error, by replacing the random variable (X, Y) with the elements of $\mathcal{T}$, and averaging the sum. □

Proof of Theorem 6. As shown in Lemma 4, we can consider the case where $\mathbf{B}$ is semi-orthogonal, without loss of generality. In this case, it follows from Lemma 5 that $\mathcal{Q}_{\mathbf{B}} = \widetilde{\mathcal{Q}}_{\mathbf{B}}$, which is simply $\mathbb{S}^k$ with a nuclear-norm constraint. We then define the following function class.

$$\mathcal{F}_{\mathbf{B}} := \{f_{\mathbf{A}} : \mathbf{x} \to (\mathbf{B}\mathbf{x})^\top\mathbf{A}(\mathbf{B}\mathbf{x}) : \mathbf{A} \in \mathcal{Q}_{\mathbf{B}} \text{ and } \mathbf{x} \in \mathcal{X}\}. \tag{31}$$

We would like to upper bound

$$\sup_{f_{\mathbf{A}}\in\mathcal{F}_{\mathbf{B}}} \left(\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\ell(f_{\mathbf{A}}(X), Y)] - \frac{1}{n}\sum_{i=1}^{n}\ell(f_{\mathbf{A}}(\mathbf{x}_i), y_i) \right), \tag{32}$$

with high-probability, with respect to the random draws of $\mathcal{T} \sim \mathcal{D}^n$. To this end, we upper bound the Rademacher complexity (Definition 8) of $\mathcal{F}_{\mathbf{B}}$. Since $\mathcal{F}_{\mathbf{B}}$ is operating in the k-dimensional space, we note that the compressed covariance, under the mapping $\mathbf{x} \mapsto \mathbf{B}\mathbf{x}$, becomes $\mathbf{B}\mathbf{\Sigma}\mathbf{B}^\top$. We have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_{\mathbf{B}}) &\le \beta a \left(\sqrt{\frac{r(\mathbf{B}\mathbf{\Sigma}\mathbf{B}^\top)\ln k}{n}} + \frac{r(\mathbf{B}\mathbf{\Sigma}\mathbf{B}^\top)\ln k}{n} \right) \sigma_{\max}(\mathbf{B}\mathbf{\Sigma}\mathbf{B}^\top) && (33)\\
&\le \beta a \left(\sqrt{\frac{k\ln k}{n}} + \frac{k\ln k}{n} \right) \sigma_{\max}(\mathbf{\Sigma}), && (34)
\end{aligned}$$

where $\beta \in \mathbb{R}$ is an absolute constant. We used the bound of [1, eq. (11)] to obtain (33), and the trivial fact that $r(\mathbf{B}\mathbf{\Sigma}\mathbf{B}^\top) \le k$, along with Theorem 10, to obtain (34). To complete the proof, we then invoke Theorem 9. □

Proof of Theorem 7. As in the proof of Theorem 6, we can consider the case where $\mathbf{B}$ is semi-orthogonal, without loss of generality. For ease of notation, we define the matrix $\hat{\mathbf{A}}' := \mathbf{B}^\top\mathbf{B}\hat{\mathbf{A}}\mathbf{B}^\top\mathbf{B}$.

Note that $\hat{\mathbf{A}}' \in \mathcal{Q}$, due to Theorem 10. We then have

$$
\begin{aligned}
\hat{L}^{\mathbf{B}}_{\mathcal{T}}(\hat{\mathbf{A}}_0) - \hat{L}_{\mathcal{T}}(\hat{\mathbf{A}}) &\le \hat{L}^{\mathbf{B}}_{\mathcal{T}}(\mathbf{B}\hat{\mathbf{A}}\mathbf{B}^\top) - \hat{L}_{\mathcal{T}}(\hat{\mathbf{A}}) && (35)\\
&= \hat{L}_{\mathcal{T}}(\hat{\mathbf{A}}') - \hat{L}_{\mathcal{T}}(\hat{\mathbf{A}}) && (36)\\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(\mathbf{x}_i^\top \hat{\mathbf{A}}'\mathbf{x}_i, y_i) - \ell(\mathbf{x}_i^\top \hat{\mathbf{A}}\mathbf{x}_i, y_i)\Big) && (37)\\
&\le \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i^\top(\hat{\mathbf{A}}' - \hat{\mathbf{A}})\mathbf{x}_i| && (38)\\
&\le \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2 \max\{-\lambda_{\min}(\hat{\mathbf{A}}' - \hat{\mathbf{A}}), \lambda_{\max}(\hat{\mathbf{A}}' - \hat{\mathbf{A}})\} && (39)\\
&\le b^2 \max\{-\lambda_{\min}(\hat{\mathbf{A}}' - \hat{\mathbf{A}}), \lambda_{\max}(\hat{\mathbf{A}}' - \hat{\mathbf{A}})\} && (40)\\
&\le b^2 \max\{-\lambda_{\min}(\hat{\mathbf{A}}') - \lambda_{\min}(-\hat{\mathbf{A}}), \lambda_{\max}(\hat{\mathbf{A}}') + \lambda_{\max}(-\hat{\mathbf{A}})\} && (41)\\
&= b^2 \max\{-\lambda_{\min}(\hat{\mathbf{A}}') + \lambda_{\max}(\hat{\mathbf{A}}), \lambda_{\max}(\hat{\mathbf{A}}') - \lambda_{\min}(\hat{\mathbf{A}})\} && (42)\\
&\le b^2 \max\{(-\lambda_{\min}(\hat{\mathbf{A}}))_+ + \lambda_{\max}(\hat{\mathbf{A}}), (\lambda_{\max}(\hat{\mathbf{A}}))_+ - \lambda_{\min}(\hat{\mathbf{A}})\} && (43)\\
&\le b^2((\lambda_{\max}(\hat{\mathbf{A}}))_+ + (-\lambda_{\min}(\hat{\mathbf{A}}))_+). && (44)
\end{aligned}
$$

To obtain (35), we used the fact that $\mathbf{B}\hat{\mathbf{A}}\mathbf{B}^\top \in \mathcal{Q}_{\mathbf{B}}$, and $\hat{\mathbf{A}}_0$ is the ERM in $\mathcal{Q}_{\mathbf{B}}$. We used the 1-Lipschitz property of ℓ, to obtain (38), Theorem 12 (combined with the fact that if $x_1 < x < x_2$, then $|x| \le \max\{-x_1, x_2\}$), to obtain (39), Theorem 11 to obtain (41), and Theorem 10 (also accounting for the fact that $d-k$ eigenvalues of $\mathbf{A}$ are zero) to obtain (43). □

2 Supplementary Results

In this section we include some supplementary results that we used in our proofs.

Definition 8 (Rademacher complexity [2, Definitions 3.1 and 3.2])**.** *Let $Z_1, \dots, Z_n$ be i.i.d. Bernoulli random variables, that is*

$$\Pr\{Z_i = 1\} = \Pr\{Z_i = -1\} = 1/2, \text{ for all } i = 1, \dots, n. \tag{45}$$

Let $\mathcal{D}$ be a distribution over $\mathcal{X}$ and $\mathcal{H}$ be a class of hypotheses $h : \mathcal{X} \to \mathbb{R}$. Given a sample $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^n$ drawn i.i.d. from $\mathcal{D}$, the empirical Rademacher complexity *of $\mathcal{H}$ given $\mathcal{T}$ is defined as*

$$\hat{\mathcal{R}}_{\mathcal{T}}(\mathcal{H}) := \frac{1}{n}\mathop{\mathbb{E}}_{Z_1,\dots,Z_n}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n} Z_i h(\mathbf{x}_i)\right], \tag{46}$$

and the Rademacher complexity *of $\mathcal{H}$ with respect to $\mathcal{D}$ is defined by taking the expectation of the above quantity, with respect to the sample $\mathcal{T}$, as*

$$\mathcal{R}_n(\mathcal{F}) := \mathop{\mathbb{E}}_{\mathcal{T}\sim\mathcal{D}^n}[\hat{\mathcal{R}}_{\mathcal{T}}(\mathcal{H})]. \tag{47}$$

Theorem 9 (Rademacher bound [3])**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$ and let $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a sample of size n drawn i.i.d. from $\mathcal{D}$. Given a hypothesis class $\mathcal{H}$, a loss function $\ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_+$ such that $|\ell(y', y)| \le c$, for all $y', y \in \mathbb{R}$ and ℓ is ρ-Lipschitz in its* *first argument, then, for any $0 < \delta < 1$, with probability at least $1-\delta$ (with respect to the draw of $\mathcal{T}$), for all $h \in \mathcal{H}$, we have*

$$\underset{(X,Y)\sim\mathcal{D}}{\mathbb{E}}[\ell(h(X),Y)] - \frac{1}{|\mathcal{T}|}\sum_{(\mathbf{x},y)\in\mathcal{T}} \ell(h(\mathbf{x}),y) \leq 2\rho\mathcal{R}_n(\mathcal{H}) + c\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}. \tag{48}$$

Theorem 10 (Poincaré separation theorem [4, Corollary 4.3.37])**.** *Let $\mathbf{A} \in \mathbb{S}^d$ and $\mathbf{B} \in \mathbb{R}^{k\times d}$, where $k \leq d$, such that $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_k$. Then, for all $i \in [k]$, we have*

$$\lambda_{d-k+i}(\mathbf{A}) \leq \lambda_i(\mathbf{B}\mathbf{A}\mathbf{B}^\top) \leq \lambda_i(\mathbf{A}). \tag{49}$$

Theorem 11 (Weyl's inequality [4, Theorem 4.3.1])**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{S}^d$. Then, for all $i \in [d]$, we have*

$$\lambda_i(\mathbf{A}) + \lambda_{\min}(\mathbf{B}) \leq \lambda_i(\mathbf{A}+\mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_{\max}(\mathbf{B}). \tag{50}$$

Theorem 12 (Rayleigh quotient [4, Theorem 4.2.2])**.** *Let $\mathbf{A} \in \mathbb{S}^d$ and $\mathbf{x} \in \mathbb{R}^d$. We then have*

$$\|\mathbf{x}\|^2\lambda_{\min}(\mathbf{A}) \leq \mathbf{x}^\top\mathbf{A}\mathbf{x} \leq \|\mathbf{x}\|^2\lambda_{\max}(\mathbf{A}). \tag{51}$$

References

[1] Fabian Latorre et al. "The Effect of the Intrinsic Dimension on the Generalization of Quadratic Classifiers". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[2] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. 2nd ed. Cambridge university press, 2014.

[3] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[4] Roger A Horn and Charles R Johnson. *Matrix analysis*. 2nd ed. Cambridge university press, 2012.