8. Appendix

For the entirety of this section, assume $W = \mathbf{I}$ and $Z = \mathbf{I}_2$.

8.1. Derivation of gradient of the loss with respect to prototypes

In this subsection, we will derive the gradient of the expected loss function with respect to B under the general mixture of Gaussian model. That is, $\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\mu_+, \mathbf{\Sigma}_+) + \frac{1}{2}\mathcal{N}(\mu_-, \mathbf{\Sigma}_-)$.
The loss function with 2 prototypes and rbf-kernel ($\gamma^2 = \frac{1}{2}$) is the following:

$$\mathcal{R}_{emp} = \frac{1}{n}\sum_{i=1}^{n} \left[\|\mathbf{y}_i - \mathbf{e}_1 \exp\left\{-\tfrac{1}{2}\|\mathbf{b}_+ - \mathbf{x}_i\|^2\right\} - \mathbf{e}_2 \exp\left\{-\tfrac{1}{2}\|\mathbf{b}_- - \mathbf{x}_i\|^2\right\}\|^2\right]$$

In the infinite sample case ($n \to \infty$), with points being draw from mixture of Gaussian, we have

$$\begin{aligned}\mathcal{R} = \mathbb{E}[\mathcal{R}_{emp}] =& 0.5[\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mu_+,\mathbf{\Sigma}_+)} \left[(1 - \exp\left\{-\tfrac{1}{2}\|\mathbf{b}_+ - \mathbf{x}\|^2\right\})^2\right] \\ &+ 0.5\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mu_-,\mathbf{\Sigma}_-)} \left[(\exp\left\{-\tfrac{1}{2}\|\mathbf{b}_+ - \mathbf{x}\|^2\right\})^2\right]] \\ & 0.5[\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mu_-,\mathbf{\Sigma}_-)} \left[(1 - \exp\left\{-\tfrac{1}{2}\|\mathbf{b}_- - \mathbf{x}\|^2\right\})^2\right] \\ &+ 0.5\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mu_+,\mathbf{\Sigma}_+)} \left[(\exp\left\{-\tfrac{1}{2}\|\mathbf{b}_- - \mathbf{x}\|^2\right\})^2\right]]\end{aligned}$$

Notice that the loss function decomposes into independent terms with respect to $\mathbf{b}_+$ and $\mathbf{b}_-$. Given this observation, we constrain our focus on the analysis with respect to only $\mathbf{b}_+$. All theorems follow analogously for $\mathbf{b}_-$.

We introduce some notation that we will use in the rest of the section.

- $\mathbf{\Delta}_+ := (\mathbf{b}_+ - \mu_+)$, $\mathbf{\Delta}_- := (\mathbf{b}_+ - \mu_-)$, $\bar{\mu} := (\mu_+ - \mu_-)$.
- For p.s.d. matrix M, vector $\mathbf{v}$, $\|\mathbf{v}\|_M^2 = \mathbf{v}^T M\mathbf{v}$.
- $\mathbf{\Sigma}'_+ := (\mathbf{I} + \mathbf{\Sigma}_+)^{-1}, g'_+ := \exp\left\{-\frac{1}{2}\|\mathbf{\Delta}_+\|^2_{\mathbf{\Sigma}'_+}\right\}$.
- $\mathbf{\Sigma}''_+ := (\mathbf{I} + 2\mathbf{\Sigma}_+)^{-1}, g''_+ := \exp\left\{-\|\mathbf{\Delta}_+\|^2_{\mathbf{\Sigma}''_+}\right\}$.
- $\mathbf{\Sigma}'_- := (\mathbf{I} + 2\mathbf{\Sigma}_-)^{-1}, g'_- := \exp\left\{-\|\mathbf{\Delta}_-\|^2_{\mathbf{\Sigma}'_-}\right\}$.
- $\mathbf{\Sigma}''_- := (\mathbf{I} + 2\mathbf{\Sigma}_-)^{-1}, g''_- := \exp\left\{-\|\mathbf{\Delta}_-\|^2_{\mathbf{\Sigma}''_-}\right\}$.

Theorem 3 (Gradient of the loss)**.** *In the infinite sample case:*

$$\nabla_{\mathbf{b}_+}\mathcal{R} = \left(\mathbf{\Sigma}'_+|\mathbf{\Sigma}'_+|g'_+ - \mathbf{\Sigma}''_+|\mathbf{\Sigma}''_+|g''_+\right)\mathbf{\Delta}_+ - \left(\mathbf{\Sigma}''_-|\mathbf{\Sigma}''_-|g''_-\right)\mathbf{\Delta}_-.$$

Proof. The loss function decomposes as a sum over datapoints. Hence, using the fact that the expectation and gradient operators both distribute over sums, we can write down the gradient as,

$$\begin{aligned}&\nabla_{\mathbf{b}_+}\mathcal{R} = \mathbf{E}_{\mathbf{x}}\left[\frac{d\mathcal{R}(\mathbf{x})}{d\mathbf{b}_+}\right] = \\ &= \tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_+|}}\int (\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}\|(\mathbf{b}_+ - \mathbf{x})\|^2\right\}(1 - \exp\left\{-\tfrac{1}{2}\|(\mathbf{b}_+ - \mathbf{x})\|^2\right\})\exp\left\{-\tfrac{1}{2}(\mathbf{x} - \mu_+)^T(\mathbf{\Sigma}_+)^{-1}(\mathbf{x} - \mu_+)\right\}d\mathbf{x} \\ &- \tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_-|}}\int (\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}\|(\mathbf{b}_+ - \mathbf{x})\|^2\right\}(\exp\left\{-\tfrac{1}{2}\|(\mathbf{b}_+ - \mathbf{x})\|^2\right\})\exp\left\{-\tfrac{1}{2}(\mathbf{x} - \mu_-)^T(\mathbf{\Sigma}_-)^{-1}(\mathbf{x} - \mu_-)\right\}d\mathbf{x} \\ &= \tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_+|}}\int (\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}(\mathbf{x}^T(\mathbf{\Sigma}_+\mathbf{\Sigma}'_+)^{-1}\mathbf{x} - 2\mathbf{x}^T(\mathbf{b}_+ + \mathbf{\Sigma}_+^{-1}\mu_+) + \|\mathbf{b}_+\|^2 + {\mu_+}^T\mathbf{\Sigma}_+^{-1}\mu_+)\right\}d\mathbf{x} \\ &+ \tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_+|}}\int (\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}(\mathbf{x}^T(\mathbf{\Sigma}_+\mathbf{\Sigma}''_+)^{-1}\mathbf{x} - 2\mathbf{x}^T(2\mathbf{b}_+ + \mathbf{\Sigma}_+^{-1}\mu_+) + 2\|\mathbf{b}_+\|^2 + {\mu_+}^T\mathbf{\Sigma}_+^{-1}\mu_+)\right\}d\mathbf{x} \\ &- \tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_-|}}\int (\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}(\mathbf{x}^T(\mathbf{\Sigma}_-\mathbf{\Sigma}''_-)^{-1}\mathbf{x} - 2\mathbf{x}^T(2\mathbf{b}_+ + \mathbf{\Sigma}_-^{-1}\mu_-) + 2\|\mathbf{b}_+\|^2 + {\mu_-}^T\mathbf{\Sigma}_-^{-1}\mu_-)\right\}d\mathbf{x}.\end{aligned}$$

The last equality follows by completing the *square* and separating out constants with respect to $\mathbf{x}$. We thus have Gaussians with the following means,

$$\mu'_+ := \mathbf{\Sigma}_+\mathbf{\Sigma}'_+(\mathbf{b}_+ + \mathbf{\Sigma}_+^{-1}\mu_+),\ \mu''_+ := \mathbf{\Sigma}_+\mathbf{\Sigma}''_+(2\mathbf{b}_+ + \mathbf{\Sigma}_+^{-1}\mu_+),\ \mu''_- := \mathbf{\Sigma}_-\mathbf{\Sigma}''_-(2\mathbf{b}_+ + \mathbf{\Sigma}_-^{-1}\mu_-),$$

The following constants come out as factors,

$$g'_+ = \exp\left\{-\tfrac{1}{2}\|\mathbf{\Delta}_+\|^2_{\mathbf{\Sigma}'_+}\right\},\ g''_+ = \exp\left\{-\|\mathbf{\Delta}_+\|^2_{\mathbf{\Sigma}''_+}\right\},\ g''_- = \exp\left\{-\|\mathbf{\Delta}_-\|^2_{\mathbf{\Sigma}''_-}\right\}.$$

Then, the expression for the gradient can be re-written as:

$$\begin{aligned}\nabla_{\mathbf{b}_+}\mathcal{R} &= \left[\tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_+|}}\int(\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mu'_+)^T(\mathbf{\Sigma}_+\mathbf{\Sigma}'_+)^{-1}(\mathbf{x}-\mu'_+)\right\}d\mathbf{x}\right]g'_+ \\ &- \left[\tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_+|}}\int(\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mu''_+)^T(\mathbf{\Sigma}_+\mathbf{\Sigma}''_+)^{-1}(\mathbf{x}-\mu''_+)\right\}d\mathbf{x}\right]g''_+ \\ &- \left[\tfrac{1}{\sqrt{|2\pi\mathbf{\Sigma}_-|}}\int(\mathbf{b}_+ - \mathbf{x})\exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mu''_-)^T(\mathbf{\Sigma}_-\mathbf{\Sigma}''_-)^{-1}(\mathbf{x}-\mu''_-)\right\}d\mathbf{x}\right]g''_-.\end{aligned}$$

Using expectation of Gaussians, we get:

$$\begin{aligned}\nabla_{\mathbf{b}_+}\mathcal{R} &= \left[(\mathbf{b}_+ - \mu'_+)\left(|\mathbf{\Sigma}'_+|\right)g'_+ - (\mathbf{b}_+ - \mu''_+)\left(|\mathbf{\Sigma}''_+|\right)g''_+ - (\mathbf{b}_+ - \mu''_-)\left(|\mathbf{\Sigma}''_-|\right)g''_-\right] \\ &= \left(\mathbf{\Sigma}'_+|\mathbf{\Sigma}'_+|g'_+ - \mathbf{\Sigma}''_+|\mathbf{\Sigma}''_+|g''_+\right)\mathbf{\Delta}_+ - \left(\mathbf{\Sigma}''_-|\mathbf{\Sigma}''_-|g''_-\right)\mathbf{\Delta}_-.\end{aligned}$$

To establish the last equality, we claim that $(\mathbf{b}_+ - \mu'_+) = \mathbf{\Sigma}'_+\mathbf{\Delta}_+, (\mathbf{b}_+ - \mu''_+) = \mathbf{\Sigma}''_+\mathbf{\Delta}_+$, and $(\mathbf{b}_+ - \mu''_-) = \mathbf{\Sigma}''_-\mathbf{\Delta}_-$ as a consequence of their definitions. We show the first of these three equalities and the other two proofs are similar. Note that $\mathbf{\Sigma}_+$, $\mathbf{\Sigma}'_+$, $\mathbf{\Sigma}_+^{-1}$ and $(\mathbf{\Sigma}'_+)^{-1}$ commute, when $\mathbf{\Sigma}_+$ is non-singular. Therefore

$$\begin{aligned}(\mathbf{b}_+ - \mu'_+) &= \mathbf{\Sigma}'_+(\mathbf{\Sigma}'_+)^{-1}(\mathbf{b}_+ - \mathbf{\Sigma}_+\mathbf{\Sigma}'_+(\mathbf{b}_+ + \mathbf{\Sigma}_+^{-1}\mu_+)) \\ &= \mathbf{\Sigma}'_+((\mathbf{\Sigma}'_+)^{-1} - \mathbf{\Sigma}_+)\mathbf{b}_+ - \mathbf{\Sigma}'_+(\mathbf{\Sigma}'_+)^{-1}\mathbf{\Sigma}_+\mathbf{\Sigma}'_+\mathbf{\Sigma}_+^{-1}\mu_+ \\ &= \mathbf{\Sigma}'_+((\mathbf{I} + \mathbf{\Sigma}_+) - \mathbf{\Sigma}_+)\mathbf{b}_+ - \mathbf{\Sigma}'_+\mu_+ \\ &= \mathbf{\Sigma}'_+(\mathbf{b}_+ - \mu_+) \\ &= \mathbf{\Sigma}'_+\mathbf{\Delta}_+.\end{aligned}$$

□

Corollary 1. *Let $\mathbf{x}$ be sampled from a mixture of 2 spherical Gaussians, i.e., $\mathbf{\Sigma}_+ = \mathbf{\Sigma}_- = \mathbf{I}$ and $\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\mu_+, I) + \frac{1}{2}\mathcal{N}(\mu_-, I)$. Then, the gradient of the loss is given by,*

$$\nabla_{\mathbf{b}_+}\mathcal{R} = \left(\frac{1}{2^{d/2+1}}\right)g'_+ - \left(\frac{1}{3^{d/2+1}}\right)g''_+ - \left(\frac{1}{3^{d/2+1}}\right)g''_-$$

where,

$$g'_+ = \mathbf{\Delta}_+\exp\left\{-\frac{1}{4}\|\mathbf{\Delta}_+\|^2\right\}, g''_+ = \exp\left\{-\frac{1}{3}\|\mathbf{\Delta}_+\|^2\right\}, g''_- = \exp\left\{-\frac{1}{3}\|\mathbf{\Delta}_-\|^2\right\}.$$

For the rest of this section, unless otherwise stated, assume $\mathbf{\Sigma}_+ = \mathbf{\Sigma}_- = \mathbf{I}$.

8.2. Lemmas

Lemma 1. *Let assumptions of Theorem 1 hold. In particular, let ${\mathbf{\Delta}_+}^T\bar{\mu} \geq -\frac{(1-\delta)}{2}\|\bar{\mu}\|^2$, for some small constant $\delta > 0$, then:*

$$g''_- \leq g'_+\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\},$$

where $g''_- := \exp\left\{-\frac{\|\mathbf{\Delta}_-\|^2}{3}\right\}$, $g'_+ := \exp\left\{-\frac{\|\mathbf{\Delta}_+\|^2}{4}\right\}$.

Proof.

$$\begin{aligned} g''_- &= \exp\left\{-\frac{\|\boldsymbol{\Delta}_-\|^2}{3}\right\} \leq \exp\left\{-\frac{\|\boldsymbol{\Delta}_-\|^2}{4}\right\} = \exp\left\{-\frac{\|\boldsymbol{\Delta}_+ + \bar{\mu}\|^2}{4}\right\} = \exp\left\{-\frac{\|\boldsymbol{\Delta}_+\|^2 + \|\bar{\mu}\|^2 + 2{\boldsymbol{\Delta}_+}^T\bar{\mu}}{4}\right\} \\ &\overset{\zeta_1}{\geq} \exp\left\{-\frac{\|\boldsymbol{\Delta}_+\|^2 + \|\bar{\mu}\|^2 - (1-\delta)\|\bar{\mu}\|^2}{4}\right\} = \exp\left\{-\frac{\|\boldsymbol{\Delta}_+\|^2}{4}\right\}\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\} = g'_+\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}. \end{aligned}$$

Here, ζ_1 follows by replacing ${\boldsymbol{\Delta}_+}^T\bar{\mu}$ with its lower bound $-\frac{(1-\delta)}{2}\|\bar{\mu}\|^2$ as specified in the assumption. □

Lemma 2. *Let assumptions of Theorem 1 hold. In particular, if for some fixed* $\delta \geq 0$, ${\boldsymbol{\Delta}_+}^T\bar{\mu} \geq -\frac{(1-\delta)}{2}\|\bar{\mu}\|^2$,*and* $\|\boldsymbol{\Delta}_+\| \geq 8\|\bar{\mu}\|\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}$ *for some fixed* $\alpha > 0$. *Also, let* $d \geq 8(\alpha-\delta)\|\bar{\mu}\|^2$. *Then we have:*

$${\boldsymbol{\Delta}_+}^T\mathbf{E}\left[\frac{d\mathcal{R}}{d\mathbf{b}_+}\right] > 0.1\cdot(\tfrac{1}{2})^{\frac{d}{2}+1}g'_+\|\boldsymbol{\Delta}_+\|\|\boldsymbol{\Delta}_-\|\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\} > 0,$$

where where $g''_- := \exp\left\{-\frac{\|\boldsymbol{\Delta}_-\|^2}{3}\right\}$, $g'_+ := \exp\left\{-\frac{\|\boldsymbol{\Delta}_+\|^2}{4}\right\}$.

Proof. Using $\boldsymbol{\Delta}_- = \boldsymbol{\Delta}_+ + \bar{\mu}$ and triangle inequality, we have:

$$\frac{\|\boldsymbol{\Delta}_+\|}{\|\boldsymbol{\Delta}_-\|} = \frac{\|\boldsymbol{\Delta}_+\|}{\|\boldsymbol{\Delta}_+ + \bar{\mu}\|} \geq \frac{\|\boldsymbol{\Delta}_+\|}{\|\boldsymbol{\Delta}_+\| + \|\bar{\mu}\|}.$$

The above quantity is monotonically increasing in $\boldsymbol{\Delta}_+$. Hence, for all $\|\boldsymbol{\Delta}_+\| \geq 8\|\bar{\mu}\|\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}$, we have:

$$\frac{\|\boldsymbol{\Delta}_+\|}{\|\boldsymbol{\Delta}_-\|} \geq \frac{8\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}}{8\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}+1} \geq 4\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}. \tag{3}$$

Using Lemma 1 and the fact that $g'_+ \geq g''_+$, we have

$$\begin{aligned} {\boldsymbol{\Delta}_+}^T\mathbf{E}\left[\frac{d\mathcal{R}}{d\mathbf{b}_+}\right] &= {\boldsymbol{\Delta}_+}^T\left[\left((\tfrac{1}{2})^{\frac{d}{2}+1}g'_+ - (\tfrac{1}{3})^{\frac{d}{2}+1}g''_+\right)\boldsymbol{\Delta}_+ - \left((\tfrac{1}{3})^{\frac{d}{2}+1}g''_-\right)\boldsymbol{\Delta}_-\right] \\ &= \left[\|\boldsymbol{\Delta}_+\|^2\left((\tfrac{1}{2})^{\frac{d}{2}+1}g'_+ - (\tfrac{1}{3})^{\frac{d}{2}+1}g''_+\right)\right] - \left[{\boldsymbol{\Delta}_+}^T\boldsymbol{\Delta}_-(\tfrac{1}{3})^{\frac{d}{2}+1}g'_+\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right] \\ &\geq \left[\|\boldsymbol{\Delta}_+\|^2\left((\tfrac{1}{2})^{\frac{d}{2}+1} - (\tfrac{1}{3})^{\frac{d}{2}+1}\right)g'_+\right] - \left[{\boldsymbol{\Delta}_+}^T\boldsymbol{\Delta}_-(\tfrac{1}{3})^{\frac{d}{2}+1}g'_+\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right] \\ &\geq \left[\|\boldsymbol{\Delta}_+\|^2\left((\tfrac{1}{2})^{\frac{d}{2}+1} - (\tfrac{1}{3})^{\frac{d}{2}+1}\right)g'_+\right] - \left[\|\boldsymbol{\Delta}_+\|\|\boldsymbol{\Delta}_-\|(\tfrac{1}{3})^{\frac{d}{2}+1}g'_+\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right] \\ &= g'_+\|\boldsymbol{\Delta}_+\|\|\boldsymbol{\Delta}_-\|\left[\frac{\|\boldsymbol{\Delta}_+\|}{\|\boldsymbol{\Delta}_-\|}\left((\tfrac{1}{2})^{\frac{d}{2}+1} - (\tfrac{1}{3})^{\frac{d}{2}+1}\right) - (\tfrac{1}{3})^{\frac{d}{2}+1}\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right]. \\ &\overset{\zeta_1}{\geq} g'_+\|\boldsymbol{\Delta}_+\|\|\boldsymbol{\Delta}_-\|\left[4\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}\left((\tfrac{1}{2})^{\frac{d}{2}+1} - (\tfrac{1}{3})^{\frac{d}{2}+1}\right) - (\tfrac{1}{3})^{\frac{d}{2}+1}\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right], \\ &= g'_+\|\boldsymbol{\Delta}_+\|\|\boldsymbol{\Delta}_-\|\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}\left[4\left((\tfrac{1}{2})^{\frac{d}{2}+1} - (\tfrac{1}{3})^{\frac{d}{2}+1}\right) - (\tfrac{1}{3})^{\frac{d}{2}+1}\exp\left\{-\frac{(\delta-\alpha)\|\bar{\mu}\|^2}{4}\right\}\right], \\ &\geq g'_+\|\boldsymbol{\Delta}_+\|\|\boldsymbol{\Delta}_-\|\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}\left[(\tfrac{1}{2})^{\frac{d}{2}+1} - (\tfrac{1}{3})^{\frac{d}{2}+1}\exp\left\{-\frac{(\delta-\alpha)\|\bar{\mu}\|^2}{4}\right\}\right], \end{aligned}$$

where ζ_1 follows from (3) and the last inequality follows by simple calculations.

Lemma now follows by using $d \geq 8(\alpha-\delta)\|\bar{\mu}\|^2$ and $d \geq 1$.

□

8.3. Proof of Theorem 1

Proof. Note that $\nabla_{\mathbf{b}_+}\mathcal{R} = c_1\boldsymbol{\Delta}_+ - c_2\boldsymbol{\Delta}_-$, where $c_1 = \frac{1}{2^{d/2+1}}\exp(-\frac{1}{4}\|\boldsymbol{\Delta}_+\|^2) - \frac{1}{3^{d/2+1}}\exp(-\frac{1}{3}\|\boldsymbol{\Delta}_+\|^2)$, $c_2 = \frac{1}{3^{d/2+1}}\exp(-\frac{1}{3}\|\boldsymbol{\Delta}_-\|^2)$.

Let ${\mathbf{b}_+}' = \mathbf{b}_+ - \eta\nabla_{\mathbf{b}_+}\mathcal{R}$. Then, ${\mathbf{b}_+}' - \mu_+ = \boldsymbol{\Delta}_+ - \eta\nabla_{\mathbf{b}_+}\mathcal{R}$.

$$\|{\mathbf{b}_+}' - \mu_+\|^2 = \|\boldsymbol{\Delta}_+\|^2 + \eta^2\|\nabla_{\mathbf{b}_+}\mathcal{R}\|^2 - 2\eta{\boldsymbol{\Delta}_+}^T\nabla_{\mathbf{b}_+}\mathcal{R}. \tag{4}$$

Note that ${\boldsymbol{\Delta}_+}^T\nabla_{\mathbf{b}_+}\mathcal{R} > 0$. Hence, setting η appropriately, we get:

$$\|{\mathbf{b}_+}' - \mu_+\|^2 \leq \|\bar{\mu}\|^2\left(1 - \frac{({\boldsymbol{\Delta}_+}^T\nabla_{\mathbf{b}_+}\mathcal{R})^2}{\|\boldsymbol{\Delta}_+\|\|\nabla_{\mathbf{b}_+}\mathcal{R}\|^2}\right). \tag{5}$$

Using Lemma 2, we have:

$$\|{\mathbf{b}_+}' - \mu_+\|^2 \leq \|\boldsymbol{\Delta}_+\|^2\left(1 - \frac{0.01\cdot(\frac{1}{2})^{d+2}\cdot(g'_+)^2\|\boldsymbol{\Delta}_-\|^2\cdot\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{2}\right\}}{2c_1^2\|\boldsymbol{\Delta}_+\|^2 + 2c_2^2\|\boldsymbol{\Delta}_-\|^2}\right). \tag{6}$$

Using $\boldsymbol{\Delta}_- = \boldsymbol{\Delta}_+ + \bar{\mu}$ and $2{\boldsymbol{\Delta}_+}^T\bar{\mu} \geq -(1-\delta)\|\bar{\mu}\|^2$, we have: $\|\boldsymbol{\Delta}_-\|^2 = \|\boldsymbol{\Delta}_+ + \bar{\mu}\|^2 = \|\boldsymbol{\Delta}_+\|^2 + \|\bar{\mu}\|^2 + 2{\boldsymbol{\Delta}_+}^T\bar{\mu} \geq \|\boldsymbol{\Delta}_+\|^2 + \delta\|\bar{\mu}\|^2$. Using monotonicity of the above function wrt $\|\boldsymbol{\Delta}_-\|$, and using $\|\boldsymbol{\Delta}_-\| \geq \|distp\|$, we have:

$$\|{\mathbf{b}_+}' - \mu_+\|^2 \leq \|\boldsymbol{\Delta}_+\|^2\left(1 - \frac{0.01\cdot(g'_+)^2\cdot(\frac{1}{2})^{d+2}\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}}{2c_1^2 + 2c_2^2}\right). \tag{7}$$

Using Lemma 1, we have: $c_2^2 \leq c_1^2$. Moreover, using $(g'_+)^2(\frac{1}{2})^{d+2} \geq 4c_1^2$, we get:

$$\|{\mathbf{b}_+}' - \mu_+\|^2 \leq \|\boldsymbol{\Delta}_+\|^2\left(1 - 0.01\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}\right). \tag{8}$$

That is, $\|{\mathbf{b}_+}' - \mu_+\|^2$ decreases geometrically until $\|{\mathbf{b}_+}'\| \leq 8\exp\left\{-\frac{\alpha\|\bar{\mu}\|^2}{4}\right\}$. □

8.4. Proof of Theorem 2

Proof. We wish to analyze the Hessian $\nabla^2_{\mathbf{b}_+}\mathcal{R}$. The loss function decomposes as a sum over datapoints. Hence, using the fact that the expectation and Hessian operators both distribute over sums, we can write down the Hessian as,

$$\nabla^2_{\mathbf{b}_+}\mathcal{R} = \mathbf{E}_{\mathbf{x}}\left[\frac{d^2\mathcal{R}(\mathbf{x})}{d{\mathbf{b}_+}^2}\right] = c_1g'_+(\mathbf{I} - \tfrac{1}{2}\boldsymbol{\Delta}_+{\boldsymbol{\Delta}_+}^T) - c_2g''_+(\mathbf{I} - \tfrac{2}{3}\boldsymbol{\Delta}_+{\boldsymbol{\Delta}_+}^T) - c_2g''_-(\mathbf{I} - \tfrac{2}{3}\boldsymbol{\Delta}_-{\boldsymbol{\Delta}_-}^T).$$

Here, $c_1 = (\frac{1}{2})^{\frac{d}{2}+1}, c_2 = (\frac{1}{3})^{\frac{d}{2}+1}$. Now,

$$\nabla^2_{\mathbf{b}_+}\mathcal{R} \succcurlyeq c_1g'_+\mathbf{I} - \left[c_1g'_+\boldsymbol{\Delta}_+{\boldsymbol{\Delta}_+}^T + c_2g''_+\mathbf{I} + c_2g''_-\mathbf{I}\right] \succcurlyeq c_1g'_+\left(\left[1 - \frac{c_2}{c_1}\left(\frac{g''_+ + g''_-}{g'_+}\right)\right]\mathbf{I} - \boldsymbol{\Delta}_+{\boldsymbol{\Delta}_+}^T\right).$$

From lemma 1, $g''_- \leq g'_+\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}$. Also, let $c := \frac{c_2}{c_1} = \left(\frac{2}{3}\right)^{d/2+1}$. It can be seen that if $d \geq 1$, $c \leq 0.6$. Thus,

$$\begin{aligned}\nabla^2_{\mathbf{b}_+}\mathcal{R} &\succcurlyeq c_1g'_+\left(\left[1 - 0.6\frac{g''_+}{g'_+} - 0.6\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right]\mathbf{I} - \boldsymbol{\Delta}_+{\boldsymbol{\Delta}_+}^T\right)\\ &= c_1g'_+\left(\left[1 - 0.6\exp\left\{-\frac{\|\boldsymbol{\Delta}_+\|^2}{12}\right\} - 0.6\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\}\right]\mathbf{I} - \boldsymbol{\Delta}_+{\boldsymbol{\Delta}_+}^T\right)\end{aligned}$$

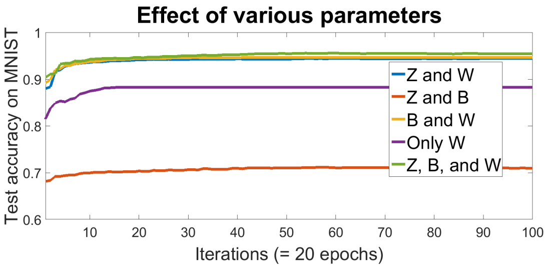

Figure 5. Iteration vs. Test Accuracy plot on mnist binary dataset. The legend shows the variables that are optimized (e.g., *Z and W* corresponds to case where we fix B and optimize over Z, W).

From assumption, $\exp\left\{-\frac{\delta\|\bar{\mu}\|^2}{4}\right\} \leq \exp\left\{-\frac{\delta\left(\frac{4}{(\ln 0.1)\delta}\right)}{4}\right\} = 0.1$. Thus,

$$\nabla^2_{\mathbf{b}_+}\mathcal{R} \succcurlyeq c_1 g'_+ \left(\left[0.9 - 0.6\exp\left\{-\frac{\|\mathbf{\Delta}_+\|^2}{12}\right\}\right]\mathbf{I} - \mathbf{\Delta}_+{\mathbf{\Delta}_+}^T\right)$$

For $\|\mathbf{\Delta}_+\|^2 \leq 0.5$, the following facts can be seen by simple one-dimensional arguments:

$$\left(0.9 - 0.6\exp\left\{-\frac{\|\mathbf{\Delta}_+\|^2}{12}\right\}\right) - \|\mathbf{\Delta}_+\|^2 \geq \frac{1}{20}.$$

$$\left(0.9 - 0.6\exp\left\{-\frac{\|\mathbf{\Delta}_+\|^2}{12}\right\}\right) \leq 1$$

$\|\mathbf{\Delta}_+\|^2$ is the only eigen value of $\mathbf{\Delta}_+{\mathbf{\Delta}_+}^T$, and all eigen values of the scaled identity matrix are $\left[0.9 - 0.6\exp\left\{-\frac{\|\mathbf{\Delta}_+\|^2}{12}\right\}\right]$. Thus the ratio of the largest eigen value to the smallest eigen value of $\nabla^2_{\mathbf{b}_+}\mathcal{R}$ is smaller than 20. Thus the condition number is bounded by 20, and the theorem follows. □

9. Experiments

9.1. Joint training of Z, B, and W

A major reason ProtoNN achieves state-of-the-art performance is because of the joint optimization problem over Z, B, W that ProtoNN solves. Instead of limiting ourselves to a projection matrix that's fixed beforehand on the basis of some unknown objective function (LMNN, SLEEC), we incorporate it into our objective and learn it along with the prototypes and the label vectors. To show that this joint optimization in fact helps improve the accuracy of ProtoNN, we conducted the following experiment, where we don't optimize one or more of Z, B, W in our algorithm and instead fix them to their initial values. We use the following hyper parameters for ProtoNN: $d = 10$, $s_W = 0.1$, $s_Z = s_B = 1.0$ and $m = 20$. We initialize ProtoNN using LMNN. If W is not begin trained, then we sparsify it immediately at the beginning of the experiment.

Figure 5 shows the results from this experiment on mnist binary dataset. The X-axis denotes iterations of alternating minimization. One iteration denotes 20 epochs each over each of the e parameters W, B, Z. From the plots, we can see that if W is fixed to its initial value, then the performance of ProtoNN drops significantly.

9.2. Datasets

Dataset	n	d	L	Links
cifar	50000	400	2	http://manikvarma.org/
character recognition	4397	400	2	https://www.kaggle.com/
eye	456	8192	2	https://rd.springer.com/chapter/10.1007/978-3-540-25976-3_23
mnist	60000	784	2	http://manikvarma.org/
usps	7291	256	2	http://manikvarma.org/
ward	4503	1000	2	https://www.kaggle.com/
letter-26	19500	16	26	https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
mnist-10	4397	784	10	https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
usps-10	7291	256	10	https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
curet-61	4209	610	61	http://www.manikvarma.org/
aloi	97200	128	1000	https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/multiclass.html
mediamill	30993	101	120	http://manikvarma.org/downloads/XC/XMLRepository.html
delicious	12920	500	983	http://manikvarma.org/downloads/XC/XMLRepository.html
eurlex	15539	5000	3993	http://manikvarma.org/downloads/XC/XMLRepository.html

Table 3. Dataset statistics and links