# Appendix A

# Proofs omitted from the main text

## A.1 Proof of McDiarmid's theorem

**Theorem A.1 (McDiarmid [94])** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $A$, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i \in A} |f(x_1,\ldots,x_n) - f(x_1,\ldots,\hat{x}_i,x_{i+1},\ldots,x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

*Then for all $\epsilon > 0$,*

$$P\left\{f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) \geq \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

*Proof* Let $V_i = V_i(X_1,\ldots,X_i) = \mathbb{E}[f|X_1,\ldots,X_i] - \mathbb{E}[f|X_1,\ldots,X_{i-1}]$, where we have denoted $f(X_1,\ldots,X_n)$ with a simple $f$. Hence

$$f - \mathbb{E}[f] = \sum_{i=1}^n V_i.$$

We will denote the probability distribution of $X_i$ by $P_i$, while with $P$ we denote as above the overall distribution. So, for any $s > 0$, we have

$$\begin{aligned}
P\{f - \mathbb{E}[f] \geq \varepsilon\} &= P\left\{\sum_{i=1}^n V_i \geq \varepsilon\right\} = P\left\{\exp\left(s\sum_{i=1}^n V_i\right)\exp(-s\varepsilon) \geq 1\right\} \\
&\leq \mathbb{E}\left[\exp\left(s\sum_{i=1}^n V_i\right)\right]\exp(-s\varepsilon) \\
&= \exp(-s\varepsilon)\,\mathbb{E}\left[\prod_{i=1}^n \exp(sV_i)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left(-s\varepsilon\right) \mathbb{E}_{X_1 \ldots X_{n-1}} \mathbb{E}_{X_n}\left[\prod_{i=1}^{n} \exp\left(sV_i\right) | X_1, \ldots, X_{n-1}\right] \\
&= \exp\left(-s\varepsilon\right) \mathbb{E}_{X_1 \ldots X_{n-1}}\left(\left[\prod_{i=1}^{n-1} \exp\left(sV_i\right)\right]\right. \qquad \text{(A.1)} \\
&\qquad \left. \mathbb{E}_{X_n}\left[\exp\left(sV_n\right) | X_1, \ldots, X_{n-1}\right]\right),
\end{aligned}
$$

where we have used the independence of the $V_i$ from $X_n$, for $i = 1, \ldots, n-1$ and the fact that the expectation of a product of independent variables equals the product of their expectations. The random variables $V_i$ satisfy

$$
\begin{aligned}
\mathbb{E}[V_i|X_1, \ldots, X_{i-1}] &= \mathbb{E}[\mathbb{E}[f|X_1, \ldots, X_i]|X_1, \ldots, X_{i-1}] - \mathbb{E}[f|X_1, \ldots, X_{i-1}] \\
&= \mathbb{E}[f|X_1, \ldots, X_{i-1}] - \mathbb{E}[f|X_1, \ldots, X_{i-1}] = 0.
\end{aligned}
$$

while their range can be bounded by

$$
L_i = \inf_a V_i(X_1, \ldots, X_{i-1}, a) \leq V_i\left(X_1, \ldots, X_i\right) \leq \sup_a V_i(X_1, \ldots, X_{i-1}, a) = U_i.
$$

If $a_l$ and $a_u$ are the values at which the inf and sup are attained, we have

$$
\begin{aligned}
&|U_i - L_i| \\
&\quad = |\mathbb{E}[f|X_1, \ldots, X_{i-1}, X_i = a_u] - \mathbb{E}[f|X_1, \ldots, X_{i-1}, X_i = a_l]| \\
&\quad = \left|\int_{A^{n-i}} f\left(X_1, \ldots, X_{i-1}, a_u, x_{i+1}, \ldots x_n\right) dP_{i+1}(x_{i+1}) \ldots dP_n(x_n)\right. \\
&\qquad \left. - \int_{A^{n-i}} f\left(X_1, \ldots, X_{i-1}, a_l, x_{i+1}, \ldots, x_n\right) dP_{i+1}(x_{i+1}) \ldots dP_n(x_n)\right| \\
&\quad \leq \int_{A^{n-i}} dP_{i+1}(x_{i+1}) \ldots dP_n(x_n) \left|f\left(X_1, \ldots, X_{i-1}, a_u, x_{i+1}, \ldots, x_n\right)\right. \\
&\qquad \left. - f\left(X_1, \ldots, X_{i-1}, a_l, x_{i+1}, \ldots, x_n\right)\right| \\
&\quad \leq |c_i|.
\end{aligned}
$$

Letting $Z(X_i) = V_i(X_1, \ldots, X_{i-1}, X_i)$ be the random variable depending only on $X_i$ for given fixed values of $X_1, \ldots, X_{i-1}$, note that

$$
\exp\left(sZ\right) \leq \frac{Z - L_i}{U_i - L_i} \exp(sU_i) + \frac{U_i - Z}{U_i - L_i} \exp(sL_i),
$$

by the convexity of the exponential function. Using the fact that

$$
\mathbb{E}[Z] = \mathbb{E}[V_i|X_1, \ldots, X_{i-1}] = 0,
$$

it follows that

$$
\mathbb{E}\left[\exp\left(sV_i\right) | X_1, \ldots, X_{i-1}\right] = \mathbb{E}[\exp\left(sZ\right)]
$$

$$\leq \frac{-L_i}{U_i - L_i}\exp(sU_i) + \frac{U_i}{U_i - L_i}\exp(sL_i)$$
$$= \exp(\psi(s)),$$

where $\psi(s) = \ln(\frac{-L_i}{U_i - L_i}\exp(sU_i) + \frac{U_i}{U_i - L_i}\exp(sL_i))$. It is not hard to check that $\psi(0) = \psi'(0) = 0$, while $\psi''(s) \leq 0.25(U_i - L_i)^2 \leq 0.25c_i^2$ for $s \geq 0$. Hence, taking three terms of the Taylor series with remainder, we have that

$$\mathbb{E}\left[\exp(sV_i)\,|X_1, \dots, X_{i-1}\right] \leq \exp\left(\frac{s^2 c_i^2}{8}\right).$$

Plugging this into inequality (A.1) for $i = n$ gives

$$P\left\{f - \mathbb{E}[f] \geq \varepsilon\right\} \leq= \exp(-s\varepsilon)\exp\left(\frac{s^2 c_n^2}{8}\right)\mathbb{E}_{X_1 \dots X_{n-1}}\left[\prod_{i=1}^{n-1}\exp(sV_i)\right].$$

By iterating the same argument for $n-1, n-2, \dots, 1$, we can show that

$$\begin{aligned}
& P\left\{f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n) \geq \epsilon\right\} \\
\leq\; & \exp(-s\varepsilon)\prod_{i=1}^{n}\exp\left(\frac{s^2 c_i^2}{8}\right) \\
=\; & \exp\left(-s\varepsilon + \frac{s^2}{8}\sum_{i=1}^{n} c_i^2\right) \\
=\; & \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right),
\end{aligned}$$

where we have chosen $s = 4\epsilon\left(\sum_{i=1}^{n} c_i^2\right)^{-1}$ to minimise the expression. □

## A.2 Stability of principal components analysis

In this appendix we prove the following theorem from Chapter 6.

**Theorem A.2 (Theorem 6.14)** *If we perform PCA in the feature space defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ then with probability greater than $1 - \delta$, for any $1 \leq k \leq \ell$, if we project new data onto the space $U_k$, the expected squared residual is bounded by*

$$\begin{aligned}
\mathbb{E}\left[\left\|P_{U_k}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\right\|^2\right] \;\leq\; & \min_{1 \leq t \leq k}\left[\frac{1}{\ell}\lambda^{>t}(S) + \frac{8}{\ell}\sqrt{(t+1)\sum_{i=1}^{\ell}\kappa(\mathbf{x}_i, \mathbf{x}_i)^2}\right] \\
& + 3R^2\sqrt{\frac{\ln(2\ell/\delta)}{2\ell}},
\end{aligned}$$

*where the support of the distribution is in a ball of radius $R$ in the feature space.*

The observation that makes the analysis possible is contained in the following theorem.

**Theorem A.3** *The projection norm $\|P_{U_k}(\phi(\mathbf{x}))\|^2$ is a linear function $\hat{f}$ in a feature space $\hat{F}$ for which the kernel function is given by*

$$\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{x}, \mathbf{z})^2.$$

*Furthermore the 2-norm of the function $\hat{f}$ is $\sqrt{k}$.*

*Proof* Let $\mathbf{X}' = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'$ be the singular value decomposition of the matrix $\mathbf{X}'$ whose rows are the images of the training examples in the feature space. The projection norm is then given by

$$\hat{f}(\mathbf{x}) = \|P_{U_k}(\boldsymbol{\phi}(\mathbf{x}))\|^2 = \boldsymbol{\phi}(\mathbf{x})'\mathbf{U}_k\mathbf{U}_k'\boldsymbol{\phi}(\mathbf{x}),$$

where $\mathbf{U}_k$ is the matrix containing the first $k$ columns of $\mathbf{U}$. Hence we can write

$$\|P_{U_k}(\boldsymbol{\phi}(\mathbf{x}))\|^2 = \sum_{i,j=1}^{N} \alpha_{ij}\boldsymbol{\phi}(\mathbf{x})_i\boldsymbol{\phi}(\mathbf{x})_j = \sum_{i,j=1}^{N} \alpha_{ij}\hat{\boldsymbol{\phi}}(\mathbf{x})_{ij},$$

where $\hat{\boldsymbol{\phi}}$ is the mapping into the feature space $\hat{F}$ composed of all pairs of $F$ features and $\alpha_{ij} = (\mathbf{U}_k\mathbf{U}_k')_{ij}$. The standard polynomial construction gives the corresponding kernel $\hat{\kappa}$ as

$$\begin{aligned}
\hat{\kappa}(\mathbf{x}, \mathbf{z}) &= \kappa(\mathbf{x}, \mathbf{z})^2 = \left(\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x})_i\boldsymbol{\phi}(\mathbf{z})_i\right)^2 \\
&= \sum_{i,j=1}^{N} \boldsymbol{\phi}(\mathbf{x})_i\boldsymbol{\phi}(\mathbf{z})_i\boldsymbol{\phi}(\mathbf{x})_j\boldsymbol{\phi}(\mathbf{z})_j = \sum_{i,j=1}^{N} (\boldsymbol{\phi}(\mathbf{x})_i\boldsymbol{\phi}(\mathbf{x})_j)(\boldsymbol{\phi}(\mathbf{z})_i\boldsymbol{\phi}(\mathbf{z})_j) \\
&= \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle .
\end{aligned}$$

It remains to show that the norm of the linear function is $\sqrt{k}$. The norm satisfies (note that $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{u}_i$, $i = 1, \ldots, N$, the orthonormal columns of $\mathbf{U}$)

$$\|\hat{f}\|^2 = \sum_{i,j=1}^{N} \alpha_{ij}^2 = \|\mathbf{U}_k\mathbf{U}_k'\|_F^2 = \left\langle \sum_{i=1}^{k} \mathbf{u}_i\mathbf{u}_i', \sum_{j=1}^{k} \mathbf{u}_j\mathbf{u}_j' \right\rangle_F = \sum_{i,j=1}^{k} (\mathbf{u}_i'\mathbf{u}_j)^2 = k,$$

as required. □

Since the norm of the residual can be viewed as a linear function we can now apply the methods developed in Chapter 4.

**Theorem A.4** *If we perform PCA on a training set $S$ of size $\ell$ in the feature space defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ and project new data onto the space $U_k$ spanned by the first $k$ eigenvectors, with probability greater than $1 - \delta$ over the generation of the sample $S$ the expected squared residual is bounded by*

$$\mathbb{E}\left[\|P_{U_k}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\|^2\right] \leq \frac{1}{\ell}\lambda^{>k}(S) + \frac{8}{\ell}\sqrt{(k+1)\sum_{i=1}^{\ell}\kappa(\mathbf{x}_i, \mathbf{x}_i)^2} + 3R^2\sqrt{\frac{\ln(2/\delta)}{2\ell}},$$

*where*

$$R^2 = \max_{\mathbf{x} \in \mathrm{supp}(\mathcal{D})} \kappa(\mathbf{x}, \mathbf{x}).$$

*Proof* Prompted by Theorem A.3 we consider the linear function class

$$\hat{\mathcal{F}}_{\sqrt{k}} = \left\{\mathbf{x} \to \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle : \|\mathbf{w}\| \leq \sqrt{k}\right\}$$

with respect to the kernel

$$\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{x}, \mathbf{z})^2 = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z})\rangle,$$

with corresponding feature mapping $\hat{\boldsymbol{\phi}}$. However, we further augment the corresponding primal weight vectors with one further dimension while augmenting the corresponding feature vectors with a feature

$$\|\boldsymbol{\phi}(\mathbf{x}))\|^2 = \kappa(\mathbf{x}, \mathbf{x}) = \sqrt{\hat{\kappa}(\mathbf{x}, \mathbf{x})} = \|\hat{\boldsymbol{\phi}}(\mathbf{x}))\|$$

that is the norm squared in the original feature space. We now apply Theorem 4.9 to the loss class

$$\begin{aligned}\hat{F}_{\mathcal{L}} &= \left\{f_{\mathcal{L}} : (\hat{\boldsymbol{\phi}}(\mathbf{x}), \|\hat{\boldsymbol{\phi}}(\mathbf{x}))\|) \mapsto \mathcal{A}(\|\hat{\boldsymbol{\phi}}(\mathbf{x}))\| - f(\hat{\boldsymbol{\phi}}(\mathbf{x}))) \mid f \in \hat{\mathcal{F}}_{\sqrt{k}}\right\} \qquad (A.2)\\ &\subseteq \mathcal{A} \circ \hat{\mathcal{F}}'_{\sqrt{k+1}},\end{aligned}$$

where $\hat{\mathcal{F}}'_{\sqrt{k+1}}$ is the class of linear functions with norm bounded by $\sqrt{k+1}$ in the feature space defined by the kernel

$$\hat{\kappa}'(\mathbf{x}, \mathbf{z}) = \hat{\kappa}(\mathbf{x}, \mathbf{z}) + \kappa(\mathbf{x}, \mathbf{x})\kappa(\mathbf{z}, \mathbf{z}) = \kappa(\mathbf{x}, \mathbf{z})^2 + \kappa(\mathbf{x}, \mathbf{x})\kappa(\mathbf{z}, \mathbf{z})$$

and $\mathcal{A}$ is the function

$$\mathcal{A}(x) = \begin{cases} 0 & \text{if } x \leq 0; \\ x/R^2 & \text{if } 0 \leq x \leq R^2; \\ 1 & \text{otherwise.} \end{cases}$$

The theorem is applied to the pattern function $\mathcal{A} \circ \hat{f}_{\mathcal{L}}$ where $\hat{f}$ is the projection function of Theorem A.3 and $\hat{f}_{\mathcal{L}}$ is defined in (A.2). We conclude that with probability $1 - \delta$

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{A} \circ \hat{f}_{\mathcal{L}}(\mathbf{x})\right] \leq \hat{\mathbb{E}}\left[\mathcal{A} \circ \hat{f}_{\mathcal{L}}(\mathbf{x})\right] + \hat{R}_{\ell}(\mathcal{A} \circ \hat{\mathcal{F}}'_{\sqrt{k+1}}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}. \qquad (A.3)$$

First note that the left-hand side of the inequality is equal to

$$\frac{1}{R^2}\mathbb{E}\left[\|P_{U_k}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\|^2\right],$$

since $\mathcal{A}$ acts as the identity in the range achieved by the function $\hat{f}_{\mathcal{L}}$. Hence, to obtain the result it remains to evaluate the first two expressions on the right-hand side of equation (A.3). Again observing that $\mathcal{A}$ acts as the identity in the range achieved, the first is a scaling of the squared residual of the training set when projecting into the space $U_k$, that is

$$\frac{1}{R^2}\hat{\mathbb{E}}\left[\|P_{U_k}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\|^2\right] = \frac{1}{\ell R^2}\sum_{i=k+1}^{\ell}\lambda_i = \frac{1}{\ell R^2}\lambda^{>k}(S).$$

The second expression is $\hat{R}_{\ell}(\mathcal{A} \circ \hat{\mathcal{F}}'_{\sqrt{k+1}})$. Here we apply Theorem 4.12 and Theorem 4.15 part 4 to obtain

$$\hat{R}_{\ell}(\mathcal{A} \circ \hat{\mathcal{F}}'_{\sqrt{k+1}}) \leq \frac{4\sqrt{k+1}}{\ell R^2}\sqrt{\operatorname{tr}\left(\hat{\mathbf{K}}'\right)} = \frac{4}{R^2}\sqrt{\frac{k+1}{\ell}}\sqrt{\frac{4}{\ell}\sum_{i=1}^{\ell}\kappa(\mathbf{x}_i, \mathbf{x}_i)^2}.$$

Assembling all the components and multiplying by $R^2$ gives the result. □

We now apply the bound $\ell$ times to obtain a proof of Theorem 6.14.

*Proof* [Proof of Theorem 6.14] We apply Theorem A.4 for $k = 1, \ldots, \ell$, in each case replacing $\delta$ by $\delta/\ell$. This ensures that with probability $1 - \delta$ the assertion holds for all $\ell$ applications. The result follows from the observation that for $k \geq t$

$$\mathbb{E}\left[\|P_{U_k}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\|^2\right] \leq \mathbb{E}\left[\|P_{U_t}^{\perp}(\boldsymbol{\phi}(\mathbf{x}))\|^2\right].$$

□

## A.3 Proofs of diffusion kernels

**Proposition A.5** *Provided $\mu < \|\mathbf{K}\|^{-1} = \|\mathbf{G}\|^{-1}$, the kernel $\hat{\mathbf{K}}$ that solves the recurrences (10.2) is $\mathbf{K}$ times the von Neumann kernel over the base* *kernel* $\mathbf{K}$*, while the matrix* $\hat{\mathbf{G}}$ *satisfies*

$$\hat{\mathbf{G}} = \mathbf{G}(\mathbf{I} - \mu\mathbf{G})^{-1}.$$

*Proof* First observe that

$$\begin{aligned}
\mathbf{K}(\mathbf{I} - \mu\mathbf{K})^{-1} &= \mathbf{K}(\mathbf{I} - \mu\mathbf{K})^{-1} - \frac{1}{\mu}(\mathbf{I} - \mu\mathbf{K})^{-1} + \frac{1}{\mu}(\mathbf{I} - \mu\mathbf{K})^{-1} \\
&= -\frac{1}{\mu}(\mathbf{I} - \mu\mathbf{K})(\mathbf{I} - \mu\mathbf{K})^{-1} + \frac{1}{\mu}(\mathbf{I} - \mu\mathbf{K})^{-1} \\
&= \frac{1}{\mu}(\mathbf{I} - \mu\mathbf{K})^{-1} - \frac{1}{\mu}\mathbf{I}.
\end{aligned}$$

Now if we substitute the second recurrence into the first we obtain

$$\begin{aligned}
\hat{\mathbf{K}} &= \mu^2\mathbf{D}\mathbf{D}'\hat{\mathbf{K}}\mathbf{D}\mathbf{D}' + \mu\mathbf{D}\mathbf{D}'\mathbf{D}\mathbf{D}' + \mathbf{K} \\
&= \mu^2\mathbf{K}(\mathbf{K}(\mathbf{I} - \mu\mathbf{K})^{-1})\mathbf{K} + \mu\mathbf{K}^2 + \mathbf{K} \\
&= \mu^2\mathbf{K}(\frac{1}{\mu}(\mathbf{I} - \mu\mathbf{K})^{-1} - \frac{1}{\mu}\mathbf{I})\mathbf{K} + \mu\mathbf{K}^2 + \mathbf{K} \\
&= \mu\mathbf{K}(\mathbf{I} - \mu\mathbf{K})^{-1}\mathbf{K} + \mathbf{K}(\mathbf{I} - \mu\mathbf{K})^{-1}(\mathbf{I} - \mu\mathbf{K}) \\
&= \mathbf{K}(\mathbf{I} - \mu\mathbf{K})^{-1},
\end{aligned}$$

showing that the expression does indeed satisfy the recurrence. Clearly, by the symmetry of the definition the expression for $\hat{\mathbf{G}}$ also satisfies the recurrence. □

**Proposition A.6** *Let* $\bar{\mathbf{K}}(\mu) = \mathbf{K}\exp(\mu\mathbf{K})$*. Then* $\bar{\mathbf{K}}(\mu)$ *corresponds to a semantic proximity matrix*

$$\exp\left(\frac{\mu}{2}\mathbf{G}\right).$$

*Proof* Let $\mathbf{D}' = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'$ be the singular value decomposition of $\mathbf{D}'$, so that $\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}'$ is the eigenvalue decomposition of $\mathbf{K}$, where $\Lambda = \boldsymbol{\Sigma}'\boldsymbol{\Sigma}$. We can write $\bar{\mathbf{K}}$ as

$$\begin{aligned}
\bar{\mathbf{K}} &= \mathbf{V}\Lambda\exp(\mu\Lambda)\mathbf{V}' = \mathbf{D}\mathbf{U}\boldsymbol{\Sigma}\exp(\mu\Lambda)\boldsymbol{\Sigma}^{-1}\mathbf{U}'\mathbf{D}' \\
&= \mathbf{D}\mathbf{U}\exp(\mu\Lambda)\mathbf{U}'\mathbf{D}' = \mathbf{D}\exp(\mu\mathbf{G})\mathbf{D}',
\end{aligned}$$

as required. □