Supplementary Material

A Convergence Theorems for Multi-Categorical Conditional Mean Embeddings

In this section we provide theorems and derivations that establish convergence properties of multi-categorical conditional embeddings (MCEs). Most of the convergence results hold due to MCEs being special cases of conditional mean embeddings (CMEs), whose empirical estimates are known to converge. We include this section for completeness.

Suppose $\{X_i, Y_i\} \sim \mathbb{P}_{XY}$ are *iid* for all $i \in \mathbb{N}_n$, with $X_i : \Omega \to \mathcal{X}$ and $Y_i : \Omega \to \mathcal{Y}$. We wish to estimate some target function $f : \mathcal{X} \to \mathbb{R}$ by $\hat{f} : \mathcal{X} \to \mathbb{R}$ empirically with a dataset $\{X_i, Y_i\}_{i=1}^n$ of size $n \in \mathbb{N}_+$. Since $\hat{f}$ is empirically estimated, it is a random function over the possible data observation events $\omega \in \Omega$. The aim is to provide a sense of the stochastic convergence of $\hat{f}$ to f by providing an upper bound of their absolute pointwise difference $|\hat{f}(x) - f(x)|$, and show that such an upper bound converges to zero at some stochastic rate. Such an upper bound is provided by the convergence properties of CMEs. In particular, the empirical CME stochastically converges to the CME at rate $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$, under the assumption that $k(x, \cdot) \in \text{image}(C_{XX})$ (Song et al., 2009, Theorem 6). That is,

$$\forall x \in \mathcal{X},\ \epsilon > 0,\ \exists M_\epsilon > 0 \quad s.t. \quad \mathbb{P}\Big[\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l} > M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \tag{A.1}$$

In practice, the assumption that $k(x, \cdot) \in \text{image}(C_{XX})$ can be relaxed by replacing $\mathcal{U}_{Y|X} = C_{YX}C_{XX}^{-1}$ with $\mathcal{U}_{Y|X} = C_{YX}(C_{XX} + \lambda I)^{-1}$ (Song et al., 2013). This will apply to all the subsequent theorems in this section.

Theorem A.1 (Pointwise and Uniform Convergence of Estimators based on Conditional Embeddings)**.** *Suppose that $k(x, \cdot)$ is in the image of C_{XX} and that there exists $0 \leq \gamma(x) < \infty$ such that for some estimator function $\hat{f} : \mathcal{X} \to \mathbb{R}$ and target function $f : \mathcal{X} \to \mathbb{R}$,*

$$|\hat{f}(x) - f(x)| \leq \gamma(x)\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l}, \forall x \in \mathcal{X}, \tag{A.2}$$

then the estimator $\hat{f}$ converges pointwise to the target f at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. Further, if $\gamma(x) = \gamma$ is independent of $x \in \mathcal{X}$, then this convergence is uniform.

Proof. Suppose that there exists $0 \leq \gamma(x) < \infty$ such that (A.2) is satisfied. That is, the inequality (A.2) holds for all possible data observations $\{X_i, Y_i\}_{i=1}^n$ where $X_i : \Omega \to \mathcal{X}$, $Y_i : \Omega \to \mathcal{Y}$ for all $i \in \mathbb{N}_n$. For any constant C, the implication statement $\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_\delta} \leq C \implies |\hat{f}(x) - f(x)| \leq C\gamma(x)$ holds for all possible observation events $\omega \in \Omega$. Writing this explicitly in event space translates this to a probability statement,

$$\begin{aligned} &\{\omega \in \Omega : \big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l} \leq C\} \subseteq \{\omega \in \Omega : |\hat{f}(x) - f(x)| \leq C\gamma(x)\} \\ &\implies \mathbb{P}\Big[\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l} \leq C\Big] \leq \mathbb{P}\Big[|\hat{f}(x) - f(x)| \leq C\gamma(x)\Big]. \end{aligned} \tag{A.3}$$

Since we assume that $k(x, \cdot) \in \text{image}(C_{XX})$, statement (A.1) is valid. By letting $C = M_\epsilon((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ in (A.3), we immediately have that the probability inequality in statement (A.1) is also true if we

replace $\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|$ with $|\hat{f}(x) - f(x)|$ and M_ϵ with $\gamma(x)M_\epsilon$,

$$\begin{aligned}
&\mathbb{P}\Big[\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l} > M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon \\
\implies &1 - \mathbb{P}\Big[\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l} \leq M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon \\
\implies &\mathbb{P}\Big[\big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_l} \leq M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] > 1 - \epsilon \\
\implies &\mathbb{P}\Big[|\hat{f}(x) - f(x)| \leq \gamma(x)M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] > 1 - \epsilon \\
\implies &1 - \mathbb{P}\Big[|\hat{f}(x) - f(x)| \leq \gamma(x)M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon \\
\implies &\mathbb{P}\Big[|\hat{f}(x) - f(x)| > \gamma(x)M_\epsilon\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon.
\end{aligned} \tag{A.4}$$

where we employed statement (A.3) between the third and fourth line for $C = M_\epsilon((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. Therefore, since M_ϵ is arbitrary, define $\tilde{M}_\epsilon(x) := \gamma(x)M_\epsilon$ so that, with the above result, the statement (A.1) implies the following,

$$\forall x \in \mathcal{X},\ \epsilon > 0,\ \exists \tilde{M}_\epsilon(x) > 0 \quad s.t. \quad \mathbb{P}\Big[\big|\hat{f}(x) - f(x)\big| > \tilde{M}_\epsilon(x)\Big((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \tag{A.5}$$

In other words, the function $\hat{f}$ stochastically converges pointwise to f with a rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. The convergence is pointwise as the constant $\tilde{M}_\epsilon(x)$ may be different for each point $x \in \mathcal{X}$. If $\gamma(x) = \gamma$ such that $\tilde{M}_\epsilon(x) = \tilde{M}_\epsilon$ does not depend on $x \in \mathcal{X}$, then this stochastic convergence is uniform in its domain $\mathcal{X}$. □

With theorem A.1, we can now show the convergence of various estimators based on the conditional embedding, as long as we can show that their estimator error is upper bounded by a multiple of the conditional embedding error in the reproducing kernel Hilbert space (RKHS) norm. As such, we turn to the convergence of the empirical decision probability function (4) below.

Theorem A.2 (Uniform Convergence of Empirical Decision Probability Function)**.** *Assuming that $k(x, \cdot)$ is in the image of C_{XX}, the empirical decision probability function $\hat{p}_c : \mathcal{X} \to \mathbb{R}$ (4) converges uniformly to the true decision probability $p_c : \mathcal{X} \to [0, 1]$ (3) at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ for all $c \in \mathcal{Y} = \mathbb{N}_m$.*

Proof. Consider the pointwise absolute difference between the decision probability and its empirical estimate,

$$\begin{aligned}
|\hat{p}_c(x) - p_c(x)| &= |\langle \hat{\mu}_{Y|X=x}, \mathbb{1}_c\rangle - \langle \mu_{Y|X=x}, \mathbb{1}_c\rangle| \\
&= |\langle \hat{\mu}_{Y|X=x} - \mu_{Y|X=x}, \mathbb{1}_c\rangle| \\
&\leq \big\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\big\|_{\mathcal{H}_\delta} \big\|\mathbb{1}_c\big\|_{\mathcal{H}_\delta},
\end{aligned} \tag{A.6}$$

where the last inequality follows from the Cauchy Schwarz inequality in a Hilbert space.

Since $\mathbb{1}_c = \delta(c, \cdot)$ and using the fact that δ is a reproducing kernel, we have that for all $c \in \mathcal{Y} = \mathbb{N}_m$.

$$\big\|\mathbb{1}_c\big\|^2_{\mathcal{H}_\delta} = \langle \mathbb{1}_c, \mathbb{1}_c\rangle = \langle \delta(c, \cdot), \delta(c, \cdot)\rangle = \delta(c, c) = 1. \tag{A.7}$$

Therefore, by theorem A.1 with $\gamma(x) = 1$ independent of $x \in \mathcal{X}$, $\hat{p}_c$ converges uniformly to p_c at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ for all $c \in \mathcal{Y} = \mathbb{N}_m$.

□

The above proof is for uniform convergence over all $x \in \mathcal{X}$ at the stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. Intuitively, however, for stationary zero-centered kernels like the Gaussian kernel, the convergence rate may be higher at regions of high data density, since the kernel effects, being centered around the training data, are stronger at these regions. The worse case convergence rate described here in the theorem would be a tight lower bound for regions in $\mathcal{X}$ with lower data density,

where the kernel effects have decayed and most empirical probabilities are smaller and further from summing up to one.

Because the label space $\mathcal{Y} = \mathbb{N}_m$ is discrete and finite, *bounded* functions $g \in \mathcal{H}_\delta$ in the RKHS are equivalent to their vector representations $\mathbf{g} := \{g(c)\}_{c=1}^m$, because one can always write $g = \sum_{c=1}^m g(c)\delta(c,\cdot)$. In other words, there is an isomorphism between $\mathbb{H}_\delta$ and $\mathbb{R}^m$. A convenient consequence is that inner products in the RKHS are simply the usual dot products in a Euclidean space, since

$$\begin{aligned}\langle g_1, g_2\rangle_{\mathcal{H}_\delta} &= \left\langle \sum_{c=1}^m g_1(c)\delta(c,\cdot), \sum_{c'=1}^m g_2(c')\delta(c',\cdot)\right\rangle_{\mathcal{H}_\delta} \\ &= \sum_{c=1}^m \sum_{c'=1}^m g_1(c)g_2(c')\langle \delta(c,\cdot), \delta(c',\cdot)\rangle_{\mathcal{H}_\delta} \\ &= \sum_{c=1}^m g_1(c)g_2(c) \\ &= \mathbf{g}_1 \cdot \mathbf{g}_2.\end{aligned} \tag{A.8}$$

Consequently, the RKHS norm for bounded functions $g \in \mathcal{H}_\delta$ is simply the ℓ_2-norm of its vector representation $\mathbf{g}$,

$$\|g\|_{\mathcal{H}_\delta} = \|\mathbf{g}\|_{\ell_2}. \tag{A.9}$$

A special and convenient result that arises due to this discrete and finite label space is that the decision probabilities and its empirical estimate are simply the conditional embeddings and its empirical estimate.

Lemma A.3 (Decision Probabilities are Conditional Embeddings)**.** *The decision probability for class $c \in \mathbb{N}_m$ given an example $x \in \mathcal{X}$ is the conditional embedding with $l = \delta$ conditioned at example x evaluated at label c,*

$$p_c(x) := \mathbb{P}[Y = c|X = x] = \mu_{Y|X=x}(c). \tag{A.10}$$

Therefore, $\mathbf{p}(x) \equiv \mu_{Y|X=x}$.

Proof. Since indicator functions are the canonical features of the label RKHS $\mathcal{H}_\delta$, we employ the fact that expectations of indicator functions are probabilities to prove this claim,

$$\begin{aligned}\mu_{Y|X=x}(c) :=& \mathbb{E}[l(Y,c)|X=x] = \mathbb{E}[\delta(Y,c)|X=x] \\ =& \mathbb{E}[\mathbb{1}_c(Y)|X=x] = \mathbb{P}[Y \in \{c\}|X=x] \\ =& \mathbb{P}[Y=c|X=x] =: p_c(x).\end{aligned} \tag{A.11}$$

□

Lemma A.4 (Empirical Decision Probabilities are Empirical Conditional Embeddings)**.** *The empirical decision probability* (4) *for class $c \in \mathbb{N}_m$ given an example $x \in \mathcal{X}$ is the empirical conditional embedding with $l = \delta$ conditioned at example x evaluated at label c,*

$$\hat{p}_c(x) = \hat{\mu}_{Y|X=x}(c). \tag{A.12}$$

Therefore, $\hat{\mathbf{p}}(x) \equiv \hat{\mu}_{Y|X=x}$.

Proof. Let the canonical feature maps of $\mathcal{X}$ and $\mathcal{Y}$ be $\phi(x) = k(x,\cdot)$ and $\psi(y) = l(y,\cdot) = \delta(y,\cdot)$, then the empirical conditional embedding is defined by

$$\hat{\mu}_{Y|X=x} := \hat{\mathcal{U}}_{Y|X}\phi(x). \tag{A.13}$$

By the reproducing property, the evaluation of $\hat{\mu}_{Y|X=x} \in \mathcal{H}_l$ is given by a dot product,

$$\begin{aligned}\hat{\mu}_{Y|X=x}(c) &= \langle l(c,\cdot), \hat{\mu}_{Y|X=x} \rangle \\ &= \langle \psi(c), \hat{\mu}_{Y|X=x} \rangle \\ &= \psi(c)^T \hat{\mu}_{Y|X=x} \\ &= \psi(c)^T \hat{\mathcal{U}}_{Y|X} \phi(x) \\ &= \psi(c)^T \Psi (K + n\lambda I)^{-1} \Phi^T \phi(x) \\ &= \mathbf{l}_c^T (K + n\lambda I)^{-1} \mathbf{k}(x),\end{aligned} \tag{A.14}$$

where $\mathbf{l}_c := \{l(y_i, c)\}_{i=1}^n$ and $\mathbf{k}_x := \{k(x_i, x)\}_{i=1}^n$. While the notation $\mathbf{l}_c$ is usually avoided due do its similarity to $\mathbf{1}_c$, in this context they happen to represent equal quantities,

$$\mathbf{l}_c := \{l(y_i, c)\}_{i=1}^n = \{\delta(y_i, c)\}_{i=1}^n = \{\mathbb{1}_c(y_i)\}_{i=1}^n =: \mathbf{1}_c. \tag{A.15}$$

The claim then immediately follows by the definition of our decision probability estimator,

$$\hat{\mu}_{Y|X=x}(c) = \mathbf{1}_c^T (K + n\lambda I)^{-1} \mathbf{k}(x) =: \hat{p}_c(x). \tag{A.16}$$

□

Lemma A.4 shows that the decision function $\mathbf{f}(x)$ (5) of a MCE is no more than the empirical conditional embedding estimated from the data.

Since we have identified the equivalence of decision probabilities and the conditional embedding, we can now also show that the empirical decision probability vector also converges to the true decision probability vector.

Lemma A.5 (Uniform Convergence of Empirical Decision Probability Vector Function in ℓ_1 and ℓ_2)**.** *Assuming that $k(x,\cdot)$ is in the image of C_{XX}, the empirical decision probability vector function $\hat{\mathbf{p}} : \mathcal{X} \to \mathbb{R}^m$ (5) converges uniformly to the true decision probability vector function $\mathbf{p} : \mathcal{X} \to [0,1]^m$ in the ℓ_1-norm and ℓ_2-norm, where $\mathbf{p}(x) := \{p_c(x)\}_{c=1}^m$, at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ for all $c \in \mathcal{Y} = \mathbb{N}_m$.*

Proof. For convergence in ℓ_1, we simply extend theorem A.2, which proved that each entry of $\hat{\mathbf{p}}(x)$ converges pointwise uniformly at a rate of $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ to the corresponding entry of $\mathbf{p}(x)$. Since each entry converges stochastically at a rate of $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$, then so does the entire vector. More formally, from (A.6) and (A.7), the ℓ_1-norm of the difference can be bounded,

$$\begin{aligned}\|\hat{\mathbf{p}}(x) - \mathbf{p}(x)\|_{\ell_1} :=& \sum_{c=1}^m |\hat{p}_c(x) - p_c(x)| \\ \leq& \sum_{c=1}^m \left\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\right\|_{\mathcal{H}_\delta} \\ =& m \left\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\right\|_{\mathcal{H}_\delta}\end{aligned} \tag{A.17}$$

Therefore, by theorem A.1 with $\gamma(x) = m$ independent of $x \in \mathcal{X}$, we have uniform convergence in ℓ_1 where we replace all instances of $|\hat{f}(x) - f(x)|$ in the proof of theorem A.1 with $\|\hat{\mathbf{p}}(x) - \mathbf{p}(x)\|_{\ell_1}$.

For convergence in ℓ_2, we show that the ℓ_2-norm of the difference between the true and empirical decision probability vector functions is the same as the RKHS norm of the difference between the true and empirical conditional embedding, which converges to zero at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ for all $x \in \mathcal{X}$ and $c \in \mathcal{Y} = \mathbb{N}_m$ by (A.1). To this end, we use lemma A.3 and lemma A.4 and write

$$
\begin{aligned}
\|\hat{\mathbf{p}}(x) - \mathbf{p}(x)\|_{\ell_2} &= \|\{\hat{p}_c(x)\}_{c=1}^m - \{p_c(x)\}_{c=1}^m\|_{\ell_2} \\
&= \|\{\hat{p}_c(x) - p_c(x)\}_{c=1}^m\|_{\ell_2} \\
&= \|\{\hat{\mu}_{Y|X=x}(c) - \mu_{Y|X=x}(c)\}_{c=1}^m\|_{\ell_2} \\
&= \|\hat{\boldsymbol{\mu}}_{Y|X=x} - \boldsymbol{\mu}_{Y|X=x}\|_{\ell_2} \\
&= \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta},
\end{aligned}
\tag{A.18}
$$

where the last equality comes from (A.9) and the fact that the empirical and true conditional embeddings are bounded functions in the RKHS. Again, by theorem A.1 with $\gamma(x) = 1$ independent of $x \in \mathcal{X}$, we have uniform convergence in ℓ_2. □

A.1 Information Entropy of MCEs

The MCE provides decision probabilities instead of just a single label prediction. Such a probabilistic classifier allows us to quantify the uncertainty of its predictions for any given example $x \in \mathcal{X}$ through the information entropy (Shannon, 1951; Jaynes, 1957). This is ideal for detecting the decision boundaries of the classifier and areas of low data density.

We present two main approaches for inferring the information entropy from the classifier. Specifically, we would like to infer estimates for $h(x) := \mathbb{H}[Y|X = x] = -\sum_{c=1}^m p_c(x) \log p_c(x)$, the information entropy of the possible labels Y for a given example $X = x$.

The first approach is straight forward, which involves simply computing the information entropy with the clip normalized probabilities (6), at the query point $x \in \mathcal{X}$,

$$
\tilde{h}(x) := -\sum_{c=1}^m \tilde{p}_c(x) \log \tilde{p}_c(x). \tag{A.19}
$$

We call (A.19) the *clip-normalized information entropy*. Since $\tilde{p}_c(x)$ converges pointwise to $p_c(x)$ with increasing data, $\tilde{h}(x)$ also converges pointwise to $h(x)$.

Just as decision probabilities can be expressed as an expectation of indicator functions, information entropy can be expressed as expected information,

$$
\begin{aligned}
\mathbb{H}[Y|X = x] &= -\sum_{c=1}^m \mathbb{P}[Y = c|X = x] \log \mathbb{P}[Y = c|X = x] \\
&= \mathbb{E}[-\log \mathbb{P}[Y|X = x]|X = x] \\
&= \mathbb{E}[u_x(Y)|X = x],
\end{aligned}
\tag{A.20}
$$

where $u_x(y) := -\log \mathbb{P}[Y = y|X = x]$ is the *information* (in nats) we would gain when we discover that example x actually has label y. Note that while $\mathbb{P}[Y = c|X = x]$ is a constant, we employ the shorthand notation $\mathbb{P}[Y|X = x]$ for the random variable $g(Y)$ where $g(y) := \mathbb{P}[Y = y|X = x]$. If $u_x : \mathbb{N}_m \to \mathbb{R}$ is in the RKHS $\mathcal{H}_\delta$, then we know that this expectation can also be approximated by $\langle \hat{\mu}_{Y|X=x}, u_x \rangle$. This is the basis of our second approach.

Assuming that $\mathbb{P}[Y = y|X = x]$ is never exactly zero for all labels $y \in \mathcal{Y}$ and examples $x \in \mathcal{X}$, then $u_x(y)$ is bounded on its discrete domain $\mathbb{N}_m$. We can thus write $u_x = \sum_{c=1}^m -\log \mathbb{P}[Y = c|X = x]\delta(c, \cdot)$ which shows that u_x is in the span of the canonical kernel features and is thus in the RKHS. Hence, similar to the case with decision probabilities, with $u_x \in \mathcal{H}_\delta$ and $\mathbf{u}_x := \{u_x(y_i)\}_{i=1}^n$ we let $g = u_x$ in (2) and estimate $h(x)$ by

$$
\langle \hat{\mu}_{Y|X=x}, u_x \rangle = \mathbf{u}_x^T (K + n\lambda I)^{-1} \mathbf{k}(x). \tag{A.21}
$$

Unfortunately, u_x is not known exactly, since $\mathbb{P}[Y = y|X = x]$ is not known exactly. Instead, since $\hat{p}_c(x)$ is a consistent estimate for $\mathbb{P}[Y = c|X = x]$ by theorem A.2, we propose to replace $u_x(y)$ with the information of $\hat{p}_y(x)$. However, we cannot simply take the log of this estimator, as $\hat{p}_y(x)$ may produce non-positive estimates to the prediction probabilities. The straight forward way to

mitigate this problem is to clip $\hat{p}_y(x)$ from the bottom by a very small number, before taking the log. However, experiments show that this produces non-smooth estimates over $\mathcal{X}$ and the degree of smoothness varies drastically between different choices of that small number. Instead, in virtue of the fact that $\lim_{p\to 0} -p\log p = 0$ even though $\lim_{p\to 0} -\log p = \infty$, we simply define the information estimate $\hat{u}_x(y)$ as zero if the empirical decision probability is non-positive,

$$\hat{u}_x(y) := \begin{cases} -\log \hat{p}_y(x) & \text{if} \quad \hat{p}_y(x) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{A.22}$$

It remains to show that $\hat{u}_x \in \mathcal{H}_\delta$. Indeed, the identity $\hat{u}_x = \sum_{c=1}^m \hat{u}_x(c)\delta(c,\cdot)$ holds and thus $\hat{u}_x$ is in the span of the kernel canonical features. We then arrive at the following estimate for $h(x)$,

$$\hat{h}(x) := \langle \hat{\mu}_{Y|X=x}, \hat{u}_x \rangle = \hat{\mathbf{u}}_x^T (K + n\lambda I)^{-1} \mathbf{k}(x), \tag{A.23}$$

where $\hat{\mathbf{u}}_x := \{\hat{u}_x(y_i)\}_{i=1}^n$. Similar to the case with decision probabilities (3), the information entropy estimate (A.23) is not guaranteed to be non-negative. However, in practice these negative values are close to zero. Furthermore, negative estimated information entropy implies that the model is very confident about its prediction, and it suffices to simply clip the entropy at zero if strict information entropy is required.

Since this estimator is now based on the inner product between the empirical conditional embedding and another empirically estimate function, instead of between the empirical conditional embedding and a known function like the decision probability estimate, it is not immediately clear that such an estimator converges. Nevertheless, intuition tells us that the inner product between two converging quantities should converge. We proceed to show that this intuition is correct.

Theorem A.6 (Convergence of Empirical Information Entropy Function)**.** *Assuming that $k(x,\cdot)$ is in the image of C_{XX}, the empirical information entropy function $\hat{h} : \mathcal{X} \to \mathbb{R}$* (A.23) *converges pointwise to the true information entropy function $h : \mathcal{X} \to [0,\infty)$ at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$.*

Proof. Since we are interested in the asymptotic properties of our estimators when $n \to \infty$, and we have proved that the empirical decision probabilities converges to the true probabilities (theorem A.2), the condition $\hat{p}_c(x) > 0$ holds for large n such that we simply have $\hat{u}_x(c) = -\log \hat{p}_c(x)$. That is, the effects of clipping for the information estimate (A.22) vanishes.

Consider the pointwise absolute difference between the empirical and true information entropy,

$$\begin{aligned} |\hat{h}(x) - h(x)| &= |\langle \hat{\mu}_{Y|X=x}, \hat{u}_x \rangle_{\mathcal{H}_\delta} - \langle \mu_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta}| \\ &= |\langle \hat{\mu}_{Y|X=x}, \hat{u}_x \rangle_{\mathcal{H}_\delta} - \langle \hat{\mu}_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta} + \langle \hat{\mu}_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta} - \langle \mu_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta}| \\ &\leq |\langle \hat{\mu}_{Y|X=x}, \hat{u}_x \rangle_{\mathcal{H}_\delta} - \langle \hat{\mu}_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta}| + |\langle \hat{\mu}_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta} - \langle \mu_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta}| \\ &= |\langle \hat{\mu}_{Y|X=x}, \hat{u}_x - u_x \rangle_{\mathcal{H}_\delta}| + |\langle \hat{\mu}_{Y|X=x} - \mu_{Y|X=x}, u_x \rangle_{\mathcal{H}_\delta}| \\ &\leq \|\hat{\mu}_{Y|X=x}\|_{\mathcal{H}_\delta} \|\hat{u}_x - u_x\|_{\mathcal{H}_\delta} + \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta} \|u_x\|_{\mathcal{H}_\delta}, \end{aligned} \tag{A.24}$$

where the we used the triangle inequality and Cauchy Schwarz inequality in a Hilbert space respectively. Since $l = \delta$ is bounded, so is $\hat{\mu}_{Y|X=x}(c) = \sum_{i=1}^n w_i \delta(y_i, c)$ for some embedding weights w_i and all $c \in \mathbb{N}_m$, and thus its RKHS norm is finite for all $n \in \mathbb{N}_n$. Similarly, assuming that $p_c(x)$ is never exactly zero, $u_x(c)$ is also finite for all $c \in \mathbb{N}_m$ and thus so is its RKHS norm. We already know that $\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta}$ stochastically converges to zero at the rate $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ (A.1). Thus, it remains to bound $\|\hat{u}_x - u_x\|_{\mathcal{H}_\delta}$ by a multiple of $\|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta}$.

To this end, we first use lemma A.3 and lemma A.4 and to express the theoretical and empirical information as the negative log of the embedding, so that it is explicitly written as a function of $c \in \mathcal{Y}$ in $\mathcal{H}_\delta$ indexed by $x \in \mathcal{X}$,

$$\begin{aligned} u_x(c) &= -\log p_c(x) = -\log \mu_{Y|X=x}(c), \\ \hat{u}_x(c) &= -\log \hat{p}_c(x) = -\log \hat{\mu}_{Y|X=x}(c). \end{aligned} \tag{A.25}$$

Since $\log$ is a concave function, we have the property that $\log a - \log b \leq \frac{1}{b}(a-b)$. This allows us to bound $|\hat{u}_x(c) - u_x(c)|$ by $|\hat{\mu}_{Y|X=x}(c) - \mu_{Y|X=x}(c)|$ for all $c \in \mathbb{N}_m$,

$$\begin{aligned} |\hat{u}_x(c) - u_x(c)| &= |\log \hat{\mu}_{Y|X=x}(c) - \log \mu_{Y|X=x}(c)| \\ &\leq \frac{1}{|\mu_{Y|X=x}(c)|} |\hat{\mu}_{Y|X=x}(c) - \mu_{Y|X=x}(c)| \\ &\leq \alpha_x |\hat{\mu}_{Y|X=x}(c) - \mu_{Y|X=x}(c)|, \end{aligned} \tag{A.26}$$

where we define $\alpha_x := \max_{c \in \mathbb{N}_m} \frac{1}{|\mu_{Y|X=x}(c)|}$, which is well defined as the conditional embedding is bounded. Since the RKHS norm of bounded functions in $\mathcal{H}_\delta$ is simply the ℓ_2-norm of their vector representations (A.9), we have

$$\begin{aligned} \|\hat{u}_x - u_x\|_{\mathcal{H}_\delta}^2 &= \|\hat{\mathbf{u}}_x - \mathbf{u}_x\|_{\ell_2}^2 \\ &= \sum_{c=1}^{m} |\hat{u}_x(c) - u_x(c)|^2 \\ &\leq \sum_{c=1}^{m} \alpha_x^2 |\hat{\mu}_{Y|X=x}(c) - \mu_{Y|X=x}(c)|^2 \\ &\leq \alpha_x^2 \sum_{c=1}^{m} |\hat{\mu}_{Y|X=x}(c) - \mu_{Y|X=x}(c)|^2 \\ &\leq \alpha_x^2 \|\hat{\boldsymbol{\mu}}_{Y|X=x} - \boldsymbol{\mu}_{Y|X=x}\|_{\ell_2}^2 \\ &\leq \alpha_x^2 \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta}^2. \end{aligned} \tag{A.27}$$

Therefore, $\|\hat{u}_x - u_x\|_{\mathcal{H}_\delta} \leq \alpha_x \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta}$, and (A.24) becomes

$$\begin{aligned} |\hat{h}(x) - h(x)| &\leq \|\hat{\mu}_{Y|X=x}\|_{\mathcal{H}_\delta} \|\hat{u}_x - u_x\|_{\mathcal{H}_\delta} + \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta} \|u_x\|_{\mathcal{H}_\delta} \\ &= \alpha_x \|\hat{\mu}_{Y|X=x}\|_{\mathcal{H}_\delta} \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta} + \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta} \|u_x\|_{\mathcal{H}_\delta} \\ &= (\alpha_x \|\hat{\mu}_{Y|X=x}\|_{\mathcal{H}_\delta} + \|u_x\|_{\mathcal{H}_\delta}) \|\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}\|_{\mathcal{H}_\delta}. \end{aligned} \tag{A.28}$$

Hence, with $\gamma(x) = \alpha_x \|\hat{\mu}_{Y|X=x}\|_{\mathcal{H}_\delta} + \|u_x\|_{\mathcal{H}_\delta}$, theorem A.1 implies that $\hat{h}$ converges pointwise to h at a stochastic rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. □

B Learning Theoretic Bounds for Multi-Categorical Conditional Mean Embeddings

In this section we derive Rademacher complexity bounds (RCBs) for MCEs, and show that it can be used in conjunction with cross entropy loss to bound the expected risk with high probability.

B.1 Rademacher Complexity Bounds

Suppose a set of training data $\{x_i, y_i\}_{i=1}^n$ is drawn from $\mathbb{P}_{XY}$ in an *iid* fashion. We denote the one hot encoded target labels of $\{y_i\}_{i=1}^n$ by $\mathbf{y}_i := \{\mathbb{1}_c(y_i)\}_{c=1}^m \in \{0,1\}^m$ and $\mathbf{Y} := [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_n]^T \in \{0,1\}^{n\times m}$. Similarly, let $\mathbf{y} \in \{0,1\}^m$ denote the one hot encoded target labels for a generic label $y \in \mathcal{Y}$. Let $k_\theta : \mathcal{X} \times \mathcal{X} \to [0,\infty)$ be a family of positive definite kernels indexed by $\theta \in \Theta$. As before, we define the shorthand notation for the gram matrices $K_\theta := \{k_\theta(x_i, x_j) : i \in \mathbb{N}_n, j \in \mathbb{N}_n\}$ and $\mathbf{k}_\theta(x) := \{k_\theta(x_i, x) : i \in \mathbb{N}_n\}$, and λ denotes the regularization parameter of the conditional embedding (1). The MCE has a predictor form $\hat{\mathbf{p}}(x) = \mathbf{f}_{\theta,\lambda}(x)$ (5) defined by

$$\mathbf{f}_{\theta,\lambda}(x) := \mathbf{Y}^T(K_\theta + n\lambda I)^{-1}\mathbf{k}_\theta(x), \tag{B.1}$$

where each entry of the predictor $\mathbf{f}_{\theta,\lambda}(x)$ is the decision probability estimate for $p_c(x)$. This defines the function class of the predictor over the kernel family and a set of regularization parameters for any set of training observations $\{x_i, y_i\}_{i=1}^n$,

$$F_n(\Theta, \Lambda) := \{\mathbf{f}_{\theta,\lambda}(x) : \theta \in \Theta, \lambda \in \Lambda\}. \tag{B.2}$$

The predictor form (B.1) is linear in the reproducing kernel Hilbert space $\mathcal{H}_{k_\theta}$ induced by k_θ in the sense that

$$\begin{aligned} \mathbf{f}_{\theta,\lambda}(x) &:= W_{\theta,\lambda}^T \phi_\theta(x), \\ W_{\theta,\lambda} &:= \Phi_\theta(K_\theta + n\lambda I)^{-1}\mathbf{Y}, \end{aligned} \tag{B.3}$$

where we decompose $\mathbf{k}_\theta(x) = \Phi_\theta^T \phi_\theta(x)$ by the reproducing property. By lemma A.4, $\mathbf{f}_{\theta,\lambda}(x) = \hat{\mathbf{p}}_{\theta,\lambda}(x) = \hat{\mu}_{Y|X=x}^{(\theta,\lambda)} = \hat{\mathcal{U}}_{Y|X}^{(\theta,\lambda)} \phi_\theta(x)$. Therefore, we have that $\hat{\mathcal{U}}_{Y|X}^{(\theta,\lambda)} \equiv W_{\theta,\lambda}^T$. Throughout this paper, inner products are defined in the Hilbert-Schmidt sense, which induces the Hilbert-Schmidt norm $\|\cdot\|_{HS}$ and generalises the Frobenius inner product with induced norm $\|\cdot\|_{\text{tr}}$ for finite dimensional operators. Nevertheless, while they refer to the same quantity, we will use the standard notations $\|\hat{\mathcal{U}}_{Y|X}^{\theta,\lambda}\|_{HS}$ as per the literature in Hilbert space embeddings and $\|W_{\theta,\lambda}\|_{\text{tr}}$ as per the literature for linear classifiers.

Theorem B.1 (Rademacher Complexity Bound for MCEs)**.** *Suppose that the trace norm $\|W_{\theta,\lambda}\|_{\text{tr}} \leq \rho$ is bounded for all $\theta \in \Theta, \lambda \in \Lambda$. Further suppose that the canonical feature map $\|\phi_\theta(x)\|_{\mathcal{H}_{k_\theta}}^2 = k_\theta(x,x) \leq \alpha^2$, $\alpha > 0$, is bounded in RKHS norm for all $x \in \mathcal{X}, \theta \in \Theta$. For any set of training observations $\{x_i, y_i\}_{i=1}^n$, the Rademacher complexity of the class of MCEs $F_n(\Theta, \Lambda)$ (B.2) defined over $\theta \in \Theta, \lambda \in \Lambda$ is bounded by*

$$\mathcal{R}_n(F_n(\Theta, \Lambda)) \leq 2\alpha\rho. \tag{B.4}$$

Proof. The Rademacher complexity (Bartlett and Mendelson, 2002, Definition 2) of the function class $F_n(\Theta, \Lambda)$ is

$$\mathcal{R}_n(F_n(\Theta, \Lambda)) := \mathbb{E}\left[\sup_{\theta\in\Theta, \lambda\in\Lambda} \left\|\frac{2}{n}\sum_{i=1}^n \sigma_i \mathbf{f}_{\theta,\lambda}(X_i)\right\|\right] = \frac{2}{n}\mathbb{E}\left[\sup_{\theta\in\Theta, \lambda\in\Lambda} \left\|\sum_{i=1}^n \sigma_i \mathbf{f}_{\theta,\lambda}(X_i)\right\|\right], \tag{B.5}$$

where σ_i are *iid* Rademacher random variables, taking values in $\{-1, 1\}$ with equal probability, and X_i are *iid* random variables from the same distribution $\mathbb{P}_X$ as our training data. We further define $\boldsymbol{\sigma} := \{\sigma_i\}_{i=1}^n$.

We first bound the term inside the suprenum using the Cauchy Schwarz inequality,

$$\begin{aligned}
\Big\| \sum_{i=1}^n \sigma_i \mathbf{f}_{\theta,\lambda}(X_i) \Big\| &= \Big\| \sum_{i=1}^n \sigma_i W_{\theta,\lambda}^T \phi_\theta(X_i) \Big\| \\
&= \Big\| W_{\theta,\lambda}^T \boldsymbol{\Phi}_\theta \boldsymbol{\sigma} \Big\| \\
&\leq \|W_{\theta,\lambda}\|_{\mathrm{tr}} \|\boldsymbol{\Phi}_\theta \boldsymbol{\sigma}\| \\
&\leq \|W_{\theta,\lambda}\|_{\mathrm{tr}} \|\boldsymbol{\Phi}_\theta^T\|_{\mathrm{tr}} \|\boldsymbol{\sigma}\| \\
&= \|W_{\theta,\lambda}\|_{\mathrm{tr}} \|\boldsymbol{\Phi}_\theta\|_{\mathrm{tr}} \|\boldsymbol{\sigma}\|,
\end{aligned} \tag{B.6}$$

where we define the random operator $\boldsymbol{\Phi}_\theta := [\phi(X_1) \quad \phi(X_2) \quad \cdots \quad \phi(X_n)]$. Note that this is distinct from Φ_θ, whose columns are the canonical RKHS features at the training observations and is not random. Now, random or not, entries of $\boldsymbol{\sigma} := \{\sigma_i\}_{i=1}^n$ are either -1 or 1, so its norm is simply $\|\boldsymbol{\sigma}\| = \sqrt{n}$. We can then also compute the trace norm of the other random component $\boldsymbol{\Phi}_\theta$,

$$\begin{aligned}
\|\boldsymbol{\Phi}_\theta\|_{\mathrm{tr}} :=& \sqrt{\mathrm{trace}(\boldsymbol{\Phi}_\theta^T \boldsymbol{\Phi}_\theta)} \\
=& \sqrt{\mathrm{trace}(\mathbf{K}_\theta)} \\
=& \sqrt{\sum_{i=1}^n k_\theta(X_i, X_i)} \\
=& \sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^n k_\theta(X_i, X_i)} \\
\leq& \sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^n \alpha^2} \\
=& \sqrt{n} \alpha,
\end{aligned} \tag{B.7}$$

where the inequality comes from the assertion that $k_\theta(x, x) \leq \alpha^2$ for all $x \in \mathcal{X}, \theta \in \Theta$. This bounds all the random components in the expectation by a constant, so that later the expectation can vanish.

Using the assertion that $\|W_{\theta,\lambda}\|_{\mathrm{tr}} \leq \rho$ for all $\theta \in \Theta, \lambda \in \Lambda$, we can now bound the Rademacher complexity,

$$
\begin{aligned}
\mathcal{R}_n(F_n(\Theta,\Lambda)) &= \frac{2}{n}\mathbb{E}\left[\sup_{\theta\in\Theta,\lambda\in\Lambda}\left\|\sum_{i=1}^{n}\sigma_i\mathbf{f}_{\theta,\lambda}(X_i)\right\|\right] \\
&\leq \frac{2}{n}\mathbb{E}\left[\sup_{\theta\in\Theta,\lambda\in\Lambda}\|W_{\theta,\lambda}\|_{\mathrm{tr}}\|\|\mathbf{\Phi}_\theta\|_{\mathrm{tr}}\|\boldsymbol{\sigma}\|\right] \\
&= \frac{2}{n}\sqrt{n}\mathbb{E}\left[\sup_{\theta\in\Theta,\lambda\in\Lambda}\|W_{\theta,\lambda}\|_{\mathrm{tr}}\|\|\mathbf{\Phi}_\theta\|_{\mathrm{tr}}\right] \\
&\leq \frac{2}{n}\sqrt{n}\sqrt{n}\alpha\mathbb{E}\left[\sup_{\theta\in\Theta,\lambda\in\Lambda}\|W_{\theta,\lambda}\|_{\mathrm{tr}}\right] \\
&\leq 2\alpha\mathbb{E}\left[\sup_{\theta\in\Theta,\lambda\in\Lambda}\|W_{\theta,\lambda}\|_{\mathrm{tr}}\right] \\
&= 2\alpha\sup_{\theta\in\Theta,\lambda\in\Lambda}\|W_{\theta,\lambda}\|_{\mathrm{tr}} \\
&\leq 2\alpha\rho.
\end{aligned}
\tag{B.8}
$$

□

Theorem B.1 provides a generic Rademacher complexity bound for any type of MCE with a bounded positive definite kernel and bounded trace norm. One of the most widely used kernels in practice are the family of stationary kernels. We provide a more specific bound for the case of stationary kernels below.

Corollary B.2 (Rademacher Complexity Bound for Stationary Kernels)**.** *Suppose that the trace norm $\|W_{\theta,\lambda}\|_{\mathrm{tr}} \leq \rho$ is bounded for all $\theta \in \Theta, \lambda \in \Lambda$. Suppose that k_θ is a family of positive definite stationary kernels. That is, $k_\theta(x,x') = \tilde{k}_\theta(\|x-x'\|)$ for some real-valued function $\tilde{k} : [0,\infty) \to [0,\infty)$. Select $\tilde{\theta} \in \Theta$ and define $\Theta(\tilde{\theta})$ such that $k_\theta(0,0) \leq k_{\tilde{\theta}}(0,0)$ for all $\theta \in \Theta(\tilde{\theta})$. For any $\tilde{\theta} \in \Theta$ and set of training observations $\{x_i, y_i\}_{i=1}^n$, the Rademacher complexity of the resulting class of MCEs $F_n(\Theta(\tilde{\theta}),\Lambda)$ defined over $\theta \in \Theta(\tilde{\theta}), \lambda \in \Lambda$ is bounded by*

$$
\mathcal{R}_n(F_n(\Theta(\tilde{\theta}),\Lambda)) \leq 2\rho\sqrt{k_{\tilde{\theta}}(0,0)}. \tag{B.9}
$$

Proof. Observe that $k_{\tilde{\theta}}(0,0)$ is an upper bound for $k_\theta(x,x)$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$,

$$
k_\theta(x,x) = \tilde{k}_\theta(\|x-x\|) = \tilde{k}_\theta(\|0\|) = k_\theta(0,0) \leq k_{\tilde{\theta}}(0,0), \tag{B.10}
$$

We simply choose $\alpha^2 = k_{\tilde{\theta}}(0,0)$ in theorem B.1. □

Corollary B.2 motivates the choice $\alpha^2(\theta) = k_\theta(0,0) = \sigma_f^2$ for stationary radial basis type kernels such as the Gaussian or Matérn kernels, where σ_f is the sensitivity (Rasmussen and Williams, 2006) of the stationary kernel, which we employ in our learning algorithm when the kernel is stationary.

B.2 Expected Risk Bounds

In order to quantify the performance of the MCE, we specify a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{A} \to [0,\infty)$, where $\mathcal{L}(y, f(x))$ measures the loss of a decision function $f : \mathcal{X} \to \mathcal{A}$ on a paired example $x \in \mathcal{X}$ and label $y \in \mathcal{Y}$. In the MCE context, the decision function is $\mathbf{f}_{\theta,\lambda} : \mathcal{X} \to \mathbb{R}^m$, with $\mathcal{A} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{N}_m$. The loss function is to capture the desire for $\mathbf{y}^T\mathbf{f}_{\theta,\lambda}(x) = f_y^{(\theta,\lambda)}(x)$ to be high for all likely test points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

A suitable choice of the loss function in the probabilistic multiclass classification context is the cross entropy loss,

$$
\mathcal{L}(y, \mathbf{f}(x)) := -\log \mathbf{y}^T\mathbf{f}(x) = -\log f_y(x), \tag{B.11}
$$

where $\mathbf{f}(x)$ are the inferred decision probability estimates of each class for the example $x \in \mathcal{X}$. Since logarithms explode at zero, in practice the probability estimate is often clipped from below at a predetermined threshold $\epsilon \in (0, 1)$. Furthermore, it is also convenient to clip the probability estimate from above at one to avoid negative losses. Consequently, with the notation $[\ \cdot\ ]_\epsilon^1 := \min\{\max\{\ \cdot\ , \epsilon\}, 1\}$, we define the effective cross entropy loss as

$$\mathcal{L}_\epsilon(y, \mathbf{f}(x)) := -\log\left[\mathbf{y}^T \mathbf{f}(x)\right]_\epsilon^1 = -\log\left[f_y(x)\right]_\epsilon^1, \tag{B.12}$$

In this way, our cross entropy loss (B.12) is both bounded and positive. In our subsequent analysis, we require that our loss function has an image in $[0, 1]$. To do this, we simply rescale the loss function by dividing it by its largest value,

$$\begin{aligned} \bar{\mathcal{L}}_\epsilon(y, \mathbf{f}(x)) &:= \frac{1}{M_\epsilon} \mathcal{L}_\epsilon(y, \mathbf{f}(x)) = -\frac{1}{M_\epsilon} \log\left[f_y(x)\right]_\epsilon^1, \\ M_\epsilon &:= -\log \epsilon. \end{aligned} \tag{B.13}$$

We will refer to (B.13) as the normalized cross entropy loss. We then further define the centered normalized cross entropy loss,

$$\tilde{\mathcal{L}}_\epsilon(y, \mathbf{f}(x)) := \bar{\mathcal{L}}_\epsilon(y, \mathbf{f}(x)) - \bar{\mathcal{L}}_\epsilon(y, \mathbf{0}) = -\frac{1}{M_\epsilon} \log\left[f_y(x)\right]_\epsilon^1 - 1. \tag{B.14}$$

With the normalized cross entropy loss (B.13) as our loss function, we now employ Theorem 8 of Bartlett and Mendelson (2002) for this loss and provide a bound for the expected normalized cross entropy loss for an unseen test example.

Lemma B.3 (Expected Risk Bound)**.** *For any integer $n \in \mathbb{N}_+$ and any set of training observations $\{x_i, y_i\}_{i=1}^n$, with probability $1 - \beta$ over iid samples $\{X_i, Y_i\}_{i=1}^n$ of length n from $\mathbb{P}_{XY}$, every $f \in F_n(\Theta, \Lambda)$ satisfies*

$$\frac{1}{M_\epsilon} \mathbb{E}[\mathcal{L}_\epsilon(Y, f(X))] \leq \frac{1}{n M_\epsilon} \sum_{i=1}^n \mathcal{L}_\epsilon(Y_i, f(X_i)) + \mathcal{R}_n(\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda)) + \sqrt{\frac{8 \log \frac{2}{\beta}}{n}}. \tag{B.15}$$

Proof. Since $\bar{\mathcal{L}}_\epsilon : \mathcal{Y} \times \mathcal{A} \to [0, 1]$ has a unit range and dominates itself, $\bar{\mathcal{L}}_\epsilon(y, f(x)) \leq \bar{\mathcal{L}}_\epsilon(y, f(x))$, the result follows directly from Theorem 8 of Bartlett and Mendelson (2002). We then use the definition (B.13) for the normalized cross entropy loss. □

Equivalently, by definition (B.2), this result holds for $f = \mathbf{f}_{\theta,\lambda}(x)$ for every $\theta \in \Theta, \lambda \in \Lambda$. The bound (B.15) involves the Rademacher complexity $\mathcal{R}_n(\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda))$ of the centered normalized cross entropy loss applied onto the class of functions $F_n(\Theta, \Lambda)$, and not just the Rademacher complexity $\mathcal{R}_n(F_n(\Theta, \Lambda))$ of the class of functions $F_n(\Theta, \Lambda)$ itself. In theorem B.1, we have bounded the latter. We now proceed to bound the former with the latter (B.4), so that the upper bound in lemma B.3 can be written in terms of the latter.

Lemma B.4 (Rademacher Complexity Bound with Cross Entropy Loss)**.** *For any integer $n \in \mathbb{N}_+$ and any set of training observations $\{x_i, y_i\}_{i=1}^n$, the Rademacher complexity of the class of cross entropy loss applied onto the MCE is bounded by*

$$\mathcal{R}_n(\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda)) \leq 2 \frac{1}{\epsilon \log \frac{1}{\epsilon}} \mathcal{R}_n(F_n(\Theta, \Lambda)), \tag{B.16}$$

where $\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda) := \{(x, y) \mapsto \tilde{\mathcal{L}}_\epsilon(y, \mathbf{f}_{\theta,\lambda}(x)) : \theta \in \Theta, \lambda \in \Lambda\}$.

Proof. Let $\tilde{\psi}(z) := -\frac{1}{M_\epsilon} \log[z]_\epsilon^1 - 1$ so that $\tilde{\psi} : \mathbb{R} \to \mathbb{R}$ satisfies $\tilde{\psi}(0) = 0$. Then, the centered normalized cross entropy loss can be written as $\tilde{\mathcal{L}}_\epsilon(y, \mathbf{f}(x)) = \tilde{\psi}(f_y(x))$. In particular, $\tilde{\psi}(z)$ is

piecewise differentiable. We proceed to show that $\tilde{\psi}$ is Lipschitz by showing that the suprenum of its absolute derivative over all piecewise regions is finite, and thus infer its Lipschitz constant.

The real-valued function $\tilde{\psi}$ can be split into three piecewise regions over the real domain,

$$\tilde{\psi}(z) = \begin{cases} 0, & z \in (-\infty, \epsilon], \\ -\frac{1}{M_\epsilon} \log z - 1, & z \in (\epsilon, 1), \\ -1, & z \in [1, \infty). \end{cases} \tag{B.17}$$

The derivative over the regions $z \in (-\infty, \epsilon]$ and $z \in [1, \infty)$ is thus 0 and the local Lipschitz constant over that region is thus 0. We then focus on the other region,

$$\sup_{z \in (\epsilon, 1)} |\tilde{\psi}'(z)| = \sup_{z \in (\epsilon, 1)} \left| -\frac{1}{z M_\epsilon} \right| = \sup_{z \in (\epsilon, 1)} \frac{1}{z M_\epsilon} = \frac{1}{\epsilon M_\epsilon} = \frac{1}{\epsilon \log \frac{1}{\epsilon}}. \tag{B.18}$$

Thus, $\tilde{\psi}$ is Lipschitz with a Lipschitz constant of $L_{\tilde{\psi}} = \frac{1}{\epsilon \log \frac{1}{\epsilon}}$.

For a given general loss function $\mathcal{L}$, Ledoux and Talagrand (2013, Corollary 3.17) proved that if there exists a Lipschitz real-valued function $\psi : \mathbb{R} \to \mathbb{R}$, $\psi(0) = 0$, with constant L_ψ such that $\mathcal{L}(y, f(x)) = \psi(f_y(x))$, then $\mathcal{R}_n(\mathcal{L} \circ F) \leq 2 L_\psi \mathcal{R}_n(F)$ for any class of functions F. This result is also described in Bartlett and Mendelson (2002, Theorem 12.4).

Applying this result to our loss function with $\mathcal{L} = \tilde{\mathcal{L}}_\epsilon$ with $\psi = \tilde{\psi}$ and $F = F_n(\Theta, \Lambda)$, we have $\mathcal{R}_n(\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda)) \leq 2 L_{\tilde{\psi}} \mathcal{R}_n(F_n(\Theta, \Lambda))$, which proves the claim. □

The bound (B.16) in lemma B.4 will be the bridge that relates the expected cross entropy loss over our function class to the Rademacher complexity of our function class. We now proceed to state the main theorem which forms the backbone of our learning algorithm for the MCE.

Lemma B.5 (ϵ-General Expected Risk Bound for MCEs)**.** *Suppose that the trace norm $\|W_{\theta,\lambda}\|_{\mathrm{tr}} \leq \rho$ is bounded for all $\theta \in \Theta, \lambda \in \Lambda$. Further suppose that the canonical feature map $\|\phi_\theta(x)\|^2_{\mathcal{H}_{k_\theta}} = k_\theta(x, x) \leq \alpha^2$, $\alpha > 0$, is bounded in RKHS norm for all $x \in \mathcal{X}, \theta \in \Theta$. For any integer $n \in \mathbb{N}_+$ and any set of training observations $\{x_i, y_i\}_{i=1}^n$, with probability of at least $1 - \beta$ over iid samples $\{X_i, Y_i\}_{i=1}^n$ of length n from $\mathbb{P}_{XY}$, every $f \in F_n(\Theta, \Lambda)$ satisfies*

$$\frac{1}{M_\epsilon} \mathbb{E}[\mathcal{L}_\epsilon(Y, f(X))] \leq \frac{1}{n M_\epsilon} \sum_{i=1}^n \mathcal{L}_\epsilon(Y_i, f(X_i)) + 4 \frac{1}{\epsilon \log \frac{1}{\epsilon}} \alpha \rho + \sqrt{\frac{8 \log \frac{2}{\beta}}{n}}, \tag{B.19}$$

for any $\epsilon \in (0, 1)$. Equivalently, the bound (B.19) *holds for $f = \mathbf{f}_{\theta,\lambda}(x)$ for every $\theta \in \Theta, \lambda \in \Lambda$.*

Proof. From theorem B.1, we have $\mathcal{R}_n(F_n(\Theta, \Lambda)) \leq 2\alpha\rho$. Further, from lemma B.4, we have $\mathcal{R}_n(\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda)) \leq 2 \frac{1}{\epsilon \log \frac{1}{\epsilon}} \mathcal{R}_n(F_n(\Theta, \Lambda))$. These are both deterministic inequalities, leading to $\mathcal{R}_n(\tilde{\mathcal{L}}_\epsilon \circ F_n(\Theta, \Lambda)) \leq 4 \frac{1}{\epsilon \log \frac{1}{\epsilon}} \alpha\rho$. We then apply this inequality to lemma B.3, which proves the claim.

□

Similar to many learning theoretic bounds, the expected risk bound (B.19) is composed of three qualitatively different terms. The first term is a training loss or data fit term, which is a measure of how poorly the decision function f is performing on a given training dataset. The second term is a model complexity or regularization term, which measures how complicated the model is. In this case, the model complexity is measured by the Rademacher complexity, which captures the expressiveness of the function class by quantifying how well the function class is able to shatter noise. The third term is a statistical constant which plays no specific role to the function class.

We will eventually be minimizing the first two terms over some class of functions $f \in F_n(\Theta, \Lambda)$ with some approach, as a proxy to minimizing the actual expected risk. It would be fruitful to develop an

intuition for the tightness of the bound from the contributions of the training loss term and the model complexity term. Since, like the expected loss, the training loss term is always in the unit range $[0, 1]$, we focus on understanding the tightness of the bound contributed from the complexity term.

Consider a clipped cross entropy loss with either a very small clipping factor $\epsilon \approx 0$, or a very large clipping factor $\epsilon \approx 1$. In these scenarios, $\epsilon \log \frac{1}{\epsilon}$ would be very small, so that the coefficient on the complexity term would then be very large, regardless of what the complexity bound factors α and ρ are. As a result, intuitively, this bound is unlikely to be tight due to the large coefficient on the complexity term.

Consequently, it would then be natural to consider a middle-ground choice of the cross entropy loss where this bound is the most tight by varying $\epsilon \in (0, 1)$. Since $\epsilon \log \frac{1}{\epsilon}$ is maximized at $\epsilon = \frac{1}{e}$ for a maximal value of $\frac{1}{e}$, such a choice in the clipping factor would indeed yield the tightest bound for the complexity bound in terms of the bounding slack of the result stated in lemma B.4.

This is great news for the complexity term. What about the training loss term? Intuition tells us that, with a clipping factor of $\epsilon = e^{-1}$ that is slightly more than a third of the way into the interval $(0, 1)$ from zero, the classifier is not being penalised as strongly for assigning probabilities smaller than e^{-1} to observed classes as compared to very small values of ϵ. Furthermore, beyond the clipping point, assigning even lower probabilities to the observations does not result in a higher loss. In practice, the cross entropy loss is renowned for its rapidly growing penalty as the probability assignment gets lower, which is advantageous when using a gradient based optimization scheme. In this case, the gradients are large in magnitude and the classifier can adjust and fix these assignment errors relatively quickly. In other words, by using a slightly larger clipping factor than usual, we have seemingly lost the faster convergence properties from using a cross entropy loss.

Nevertheless, observe that for such a clipping factor $\epsilon = e^{-1}$, the normalization constant becomes $M_{e^{-1}} = -\log \frac{1}{e} = 1$, so that it is effectively removed. Furthermore, we also have the following simple upper bound for the cross entropy loss clipped at $\epsilon = e^{-1}$,

$$\bar{\mathcal{L}}_{e^{-1}}(y, f(x)) = \mathcal{L}_{e^{-1}}(y, f(x)) \leq \mathcal{L}_{\epsilon}(y, f(x)) \quad \forall \epsilon \in (0, e^{-1}), x \in \mathcal{X}, y \in \mathcal{Y}. \tag{B.20}$$

To see why inequality (B.20) holds, note that $[f_y(x)]^1_{\epsilon} \leq [f_y(x)]^1_{e^{-1}}$ holds for all $\epsilon \in (0, e^{-1}), x \in \mathcal{X}, y \in \mathcal{Y}$. Applying negative log to both sides yields the inequality from definition (B.12).

Therefore, we propose to choose $\epsilon = e^{-1}$, and then replace replace $\mathcal{L}_{e^{-1}}$ with $\mathcal{L}_{\epsilon}$ for some new generic $\epsilon \in (0, e^{-1})$ much smaller than e^{-1} on the training loss terms. In this way, we still maintain an upper bound for the training loss term. While this bound would not necessarily be tight for high training losses, the gradients from the high training loss would drive the system to a lower training loss, where the bound would become tight again as equality holds in (B.20) whenever $f_y(x) \geq e^{-1}$.

The above intuition motivates the result in the following theorem.

Theorem B.6 (ϵ-Specific Expected Risk Bound for MCEs)**.** *Suppose that the trace norm $\|W_{\theta,\lambda}\|_{\text{tr}} \leq \rho$ is bounded for all $\theta \in \Theta, \lambda \in \Lambda$. Further suppose that the canonical feature map $\|\phi_\theta(x)\|^2_{\mathcal{H}_{k_\theta}} = k_\theta(x, x) \leq \alpha^2$, $\alpha > 0$, is bounded in RKHS norm for all $x \in \mathcal{X}, \theta \in \Theta$. For any integer $n \in \mathbb{N}_+$ and any set of training observations $\{x_i, y_i\}_{i=1}^n$, with probability of at least $1 - \beta$ over iid samples $\{X_i, Y_i\}_{i=1}^n$ of length n from $\mathbb{P}_{XY}$, every $f \in F_n(\Theta, \Lambda)$ satisfies*

$$\mathbb{E}[\mathcal{L}_{e^{-1}}(Y, f(X))] \leq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\epsilon}(Y_i, f(X_i)) + 4e\ \alpha\rho + \sqrt{\frac{8 \log \frac{2}{\beta}}{n}}, \tag{B.21}$$

for any $\epsilon \in (0, e^{-1})$. Equivalently, the bound (B.21) *holds for $f = \mathbf{f}_{\theta,\lambda}(x)$ for every $\theta \in \Theta, \lambda \in \Lambda$.*

Proof. We first apply lemma B.5 with $\epsilon = e^{-1}$,

$$\mathbb{E}[\mathcal{L}_{e^{-1}}(Y, f(X))] \leq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{e^{-1}}(Y_i, f(X_i)) + 4e\ \alpha\rho + \sqrt{\frac{8 \log \frac{2}{\beta}}{n}}. \tag{B.22}$$

For any $\epsilon \in (0, e^{-1})$, the inequality $\mathcal{L}_{e^{-1}}(Y_i, f(X_i)) \leq \mathcal{L}_\epsilon(Y_i, f(X_i))$ holds almost surely (a.s.) due to the deterministic inequality (B.20). These sets of inequalities together proves the claim. □

B.3 Expected Risk Bounds for Hyperparameter Learning

We are now ready to use the result of theorem B.6 to derive a specific expected risk bound for a given choice of hyperparameters $\theta \in \Theta$ and $\lambda \in \Lambda$ of the MCE, and not just for a general set of hyperparameters. We focus on kernels k_θ that are bounded over the domain $\mathcal{X}$ in the sense that for each $\theta \in \Theta$, $k_\theta(x, x) < \infty$ for all $x \in \mathcal{X}$.

For some kernel parameters $\tilde{\theta} \in \Theta$ and regularization parameter $\tilde{\lambda} \in \Lambda$, we construct a subset of hyperparameters (kernel parameters and regularization parameters) $\Xi(\tilde{\theta}, \tilde{\lambda}) \subseteq \Theta \times \Lambda$ such that

$$\Xi(\tilde{\theta}, \tilde{\lambda}) := \{(\theta, \lambda) \in \Theta \times \Lambda : \|W_{\theta,\lambda}\|_{\mathrm{tr}} \leq \|W_{\tilde{\theta},\tilde{\lambda}}\|_{\mathrm{tr}},\ \sup_{x \in \mathcal{X}} k_\theta(x, x) \leq \alpha^2(\tilde{\theta}) := \sup_{x \in \mathcal{X}} k_{\tilde{\theta}}(x, x)\}. \tag{B.23}$$

Clearly, this subset is non-empty, since $(\tilde{\theta}, \tilde{\lambda}) \in \Xi(\tilde{\theta}, \tilde{\lambda})$ is itself an element of this subset. Note that $\alpha : \Theta \to \mathbb{R}_+$ must necessarily exist as the kernel family k_θ is assumed to be bounded over the domain $\mathcal{X}$. The class of MCEs over this subset of hyperparameters is

$$F_n(\Xi(\tilde{\theta}, \tilde{\lambda})) := \{\mathbf{f}_{\theta,\lambda}(x) : (\theta, \lambda) \in \Xi(\tilde{\theta}, \tilde{\lambda})\}. \tag{B.24}$$

Thus, we can assert that the trace norm $\|W_{\theta,\lambda}\|_{\mathrm{tr}} \leq \rho = \|W_{\tilde{\theta},\tilde{\lambda}}\|_{\mathrm{tr}}$ is bounded for all $(\theta, \lambda) \in \Xi(\tilde{\theta}, \tilde{\lambda})$, and that the canonical feature map $\|\phi_\theta(x)\|^2_{\mathcal{H}_{k_\theta}} = k_\theta(x, x) \leq \alpha^2 = \sup_{x \in \mathcal{X}} k_{\tilde{\theta}}(x, x)$ is bounded in RKHS norm for all $x \in \mathcal{X}$, $(\theta, \lambda) \in \Xi(\tilde{\theta}, \tilde{\lambda})$. By theorem B.6, we can now claim the following.

Lemma B.7 (Expected Risk Bound for a set of MCE Hyperparameters)**.** *For any integer $n \in \mathbb{N}_+$ and any set of training observations $\{x_i, y_i\}_{i=1}^n$, with probability $1 - \beta$ over iid samples $\{X_i, Y_i\}_{i=1}^n$ of length n from $\mathbb{P}_{XY}$, every $(\theta, \lambda) \in \Xi(\tilde{\theta}, \tilde{\lambda})$ satisfies*

$$\mathbb{E}[\mathcal{L}_{e^{-1}}(Y, \mathbf{f}_{\theta,\lambda}(X))] \leq \frac{1}{n} \sum_{i=1}^n \mathcal{L}_\epsilon(Y_i, \mathbf{f}_{\theta,\lambda}(X_i)) + 4e \sqrt{\sup_{x \in \mathcal{X}} k_{\tilde{\theta}}(x, x)} \|W_{\tilde{\theta},\tilde{\lambda}}\|_{\mathrm{tr}} + \sqrt{\frac{8 \log \frac{2}{\beta}}{n}}, \tag{B.25}$$

for every $\epsilon \in (0, e^{-1})$, where

$$\begin{aligned} \mathbf{f}_{\theta,\lambda}(x) :=& \mathbf{Y}^T (K_\theta + n\lambda I)^{-1} \mathbf{k}_\theta(x), \\ \|W_{\tilde{\theta},\tilde{\lambda}}\|_{\mathrm{tr}} =& \sqrt{\mathrm{trace}\Big(\mathbf{Y}^T (K_{\tilde{\theta}} + n\tilde{\lambda} I)^{-1} K_{\tilde{\theta}} (K_{\tilde{\theta}} + n\tilde{\lambda} I)^{-1} \mathbf{Y}\Big)}. \end{aligned} \tag{B.26}$$

Proof. We first apply theorem B.6 with the choice of $\rho = \|W_{\tilde{\theta},\tilde{\lambda}}\|_{\mathrm{tr}}$ and $\alpha^2 = \sup_{x \in \mathcal{X}} k_{\tilde{\theta}}(x, x)$. The inequality (B.21) then only holds for a subset of kernel parameters and regularizations $(\theta, \lambda) \in \Xi(\tilde{\theta}, \tilde{\lambda})$ as defined by (B.23). □

Since inequality (B.25) holds for any $(\theta, \lambda) \in \Xi(\tilde{\theta}, \tilde{\lambda})$ and we know that $(\tilde{\theta}, \tilde{\lambda}) \in \Xi(\tilde{\theta}, \tilde{\lambda})$, we choose $\theta = \tilde{\theta}$ and $\lambda = \tilde{\lambda}$. We now arrive at our final result from which we can bound the expected risk for a specific choice of hyperparameters $\theta \in \Theta$ and $\lambda \in \Lambda$.

Theorem B.8 (Expected Risk Bound for a choice of MCE Hyperparameters)**.** *For any integer $n \in \mathbb{N}_+$ and any set of training observations $\{x_i, y_i\}_{i=1}^n$, with probability $1 - \beta$ over iid samples $\{X_i, Y_i\}_{i=1}^n$ of length n from $\mathbb{P}_{XY}$, every $\theta \in \Theta$ and $\lambda \in \Lambda$ satisfies*

$$\mathbb{E}[\mathcal{L}_{e^{-1}}(Y, \mathbf{f}_{\theta,\lambda}(X))] \leq \frac{1}{n} \sum_{i=1}^n \mathcal{L}_\epsilon(Y_i, \mathbf{f}_{\theta,\lambda}(X_i)) + 4e\, r(\theta, \lambda) + \sqrt{\frac{8 \log \frac{2}{\beta}}{n}}, \tag{B.27}$$

for every $\epsilon \in (0, e^{-1})$*, where*

$$\begin{aligned} \mathbf{f}_{\theta,\lambda}(x) &:= \mathbf{Y}^T (K_\theta + n\lambda I)^{-1} \mathbf{k}_\theta(x), \\ r(\theta, \lambda) &:= \sqrt{\operatorname{trace}\left(\mathbf{Y}^T (K_\theta + n\lambda I)^{-1} K_\theta (K_\theta + n\lambda I)^{-1} \mathbf{Y} \right) \sup_{x \in \mathcal{X}} k_\theta(x, x)}. \end{aligned} \tag{B.28}$$

Proof. We first apply lemma B.7 with the choice of $\theta = \tilde{\theta}$ and $\lambda = \tilde{\lambda}$. We then replace the notation $\tilde{\theta} \to \theta$ and $\tilde{\lambda} \to \lambda$ back to avoid cluttered notation. Note that this should not be confused with the general θ and λ from earlier theorems. □

C Special Cases of Multi-Categorical Conditional Mean Embeddings

For MCEs, the modelling lies in the choice of the kernel family $k_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ over the input space $\mathcal{X}$. The only requirement for the kernel k is that it is symmetric and positive definite, and thus we may construct richer and more expressive kernel families in any way subject to such requirements. Once such a kernel family is constructed, the kernel parameters θ, as well as the regularization parameter λ, can be learned effectively using algorithm 1.

One way to construct richer and more expressive kernels is to compose them from simpler kernels. For example, we can construct new kernels through convex combinations or products of multiple simpler kernels (Genton, 2001). Any new parameters, such as coefficients for linear combinations of simpler kernels, can be included into the kernel parameters θ and learned in the same way as before. Alternatively, there may be domain specific structures or representations within the data that can be exploited. We can then construct the kernel family by incorporating such structural representations into the kernel. Even better, we can construct the kernel family so that it is capable or learning such structural representations by itself, by parameterizing such representations into the kernel.

In this section, we focus on special cases of the MCE where the kernel family is constructed through explicit feature maps. This construction allows the incorporation of trainable domain specific structures and enables scalability to larger datasets. We first begin by introducing the explicit MCE in appendix C.1, where explicit feature maps can be learned while enabling scalability to larger datasets. We then construct the conditional embedding network (CEN) in appendix C.2, where the kernel family is formed from multiple layers of learned representations before a simpler kernel encodes their similarity for inference. Finally, we marry both constructions into the explicit CEN in appendix C.3, which provides a scalable and more applicable version of the deep CEN by placing a linear kernel on the network features.

In essence, we can categorise the MCE using two properties: the model width and the model depth. The model width represents the dimensionality of the feature space used to construct the linear decision boundaries. The model depth represents the number of transformations used to map examples from the input space to the feature space. By implicitly defining a high dimensional feature space through simple transformations, typical nonlinear kernels produce classifiers that have a shallow but wide architecture. In contrast, the three MCE variants to be introduced in this section form other combinations of model architecture in both depth and width. Of course, this characterization of architecture is not mutually exclusive. For example, a polynomial kernel can be seen as a nonlinear kernel where higher order polynomial features are implicitly defined, or as a linear kernel on explicit polynomial features. We summarize those architectures in table C.1.

Table C.1: Properties of MCE architectures

MCE Variant	Width	Depth	Scalability	Flexibility	Typical Datasets
Implicit MCE	Wide	Shallow	Low	High	High or Low d, Low n
Explicit MCE	Narrow	Shallow	High	Low	Low d, High n
Implicit CEN	Wide	Deep	Low	High	Structured d, Low n
Explicit CEN	Narrow	Deep	High	High	Structured d, High n

C.1 Explicit Multi-Categorical Conditional Mean Embedding

The advantage of using a kernel-based classifier is that the kernel k allows us to express nonlinearities in a simple way. It does this by implicitly mapping the input space $\mathcal{X}$ to a high dimensional feature space $\mathcal{H}_k$ of non-linear basis functions such that decision boundaries become linear in that space. For many kernels, such as the Gaussian kernel defined over the Euclidean space, the feature space $\mathcal{H}_k$ has dimensionality that is uncountably infinite. Nevertheless, by virtue of the Representer Theorem (Kimeldorf and Wahba, 1971), the resulting decision functions can be represented by a finite linear combination of kernels centered at the training data, and the MCE is no exception. This elegant and convenient result enables exact inference to be performed while only requiring a finite kernel gram matrix of the size of the dataset ($n \times n$) to be computed. In this way, the capacity of the model grows with the size of the dataset, which makes kernel methods nonparametric and very flexible, as it can adapt to the complexity of a dataset even with relatively simple kernels.

Algorithm C.1 Explicit MCE Hyperparameter Learning with Batch Stochastic Gradient Updates

Input: feature family $\varphi_\theta : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^p$, dataset $\{x_i, y_i\}_{i=1}^n$, initial feature parameters θ_0, initial regularization parameter λ_0, learning rate η, batch size n_b
$\theta \leftarrow \theta_0$, $\lambda \leftarrow \lambda_0$
repeat
  Sample the next batch $\mathcal{I}_b \subseteq \mathbb{N}_n$, $|\mathcal{I}_b| = n_b$ (For gradient descent, $n_b = n$ and $\mathcal{I}_b = \mathbb{N}_n$)
  $Y \leftarrow \{\delta(y_i, c) : i \in \mathcal{I}_b, c \in \mathbb{N}_m\}$ $\in \{0,1\}^{n_b \times m}$
  $Z_\theta \leftarrow \{\varphi_\theta(x_i) : i \in \mathcal{I}_b\}$ $\in \mathbb{R}^{n_b \times p}$
  $L_{\theta,\lambda} \leftarrow \text{cholesky}(Z_\theta^T Z_\theta + n_b \lambda I_p)$ $\in \mathbb{R}^{p \times p}$
  $W_{\theta,\lambda} \leftarrow L_{\theta,\lambda}^T \backslash (L_{\theta,\lambda} \backslash Z_\theta^T Y)$ $\in \mathbb{R}^{p \times m}$
  $P_{\theta,\lambda} \leftarrow Z_\theta W_{\theta,\lambda}$ $\in \mathbb{R}^{n_b \times m}$
  $r(\theta, \lambda) = \alpha(\theta) \sqrt{\sum_{c=1}^m \sum_{j=1}^p (W_{\theta,\lambda})_{j,c}^2}$
  $q(\theta, \lambda) \leftarrow \frac{1}{n_b} \sum_{i=1}^{n_b} \mathcal{L}_\epsilon((Y)_i, (P_{\theta,\lambda})_i) + 4e\, r(\theta, \lambda)$
  $(\theta, \lambda) \leftarrow \text{GradientBasedUpdate}(q, \theta, \lambda; \eta)$
until maximum iterations reached
Output: kernel parameters θ, regularization parameter λ

However, this elegant property is also the very reason that prevents kernel-based methods from scaling to larger datasets, as the size of such a gram matrix grows very quickly by $O(n^2)$. Many kernel-based methods also require the inversion of a regularized gram matrix, which has a time complexity of $O(n^3)$, and cannot be easily parallelized like standard matrix multiplications. As such, inference on datasets beyond tens of thousands of observations quickly becomes impractical to perform with kernel-based techniques.

In order to scale to big datasets, instead of placing a kernel over the input space directly and let it implicitly define the feature space, we explicitly define a finite dimensional feature space $\mathcal{Z} \subseteq \mathbb{R}^p$ of lower dimension p, where $p < n$, and place a linear kernel over it. That is, we specify a family of explicit features maps $\varphi_\theta : \mathcal{X} \to \mathcal{Z}$, and place a linear kernel on top of these explicit features,

$$k_\theta(x, x') = \varphi_\theta(x)^T \varphi_\theta(x'). \tag{C.1}$$

By explicitly defining a finite dimensional feature space, the matrix to be inverted during both learning and inference in the MCE can be reduced from size $n \times n$ to size $p \times p$ by using the Woodbury matrix inversion identity (Higham, 2002). We use this identity to modify algorithm 1 to algorithm C.1 to exploit this computational speed up.

However, with a fixed and finite amount of feature basis, the model becomes parametric and its flexibility is compromised. In other words, the model is narrow in the number of feature representations. We therefore turn to multi-layered feature compositions, where the flexibility of a model comes from the deep architecture instead of implicit high dimensional features.

C.2 Conditional Embedding Network

For many application domains, there are natural structures in the data. For example, in image recognition, pixel dimensions are spatially correlated: nearby pixels are more related, and ordering between the pixel dimensions matter. One would expect convolutional features (LeCun et al., 1998) to be natural in this domain, and provide a performance boost to our classifier should it be included. In this way, we can often benefit by including domain specific structures and features into our model.

In this section, we focus on constructing kernels for which inputs $x, x' \in \mathcal{X}$ is to undergo various stages of feature transformations before such it is passed into a simpler kernel κ that captures the similarity between the representations. Specifically, we pay particular attention to feature transformations in the form of a perceptron, so that the cumulative stages of feature transformation become the (feed-forward) multi-layer perceptron that is familiar within the neural network literature.

Formally, let $\mathcal{F}_0 := \mathcal{X}$ be the original input space. The j^{th} layer of the network $\varphi_{\theta_j}^{(j)} : \mathcal{F}_{j-1} \to \mathcal{F}_j, j = 1, 2, \ldots, L$ is to transform features from the previous layer to features in the current layer,

where L is the total number of such feature transformation layers, and $\theta_j \in \Theta_j$ parametrizes each of those transformations.

For example, in a typical multi-layer perceptron context, each layer can be written as $\varphi^{(j)}_{\theta_j}(x) = \sigma(W_j x + b_j)$, where W_j and b_j are the weight and bias parameters of the layer, and σ is an element-wise activation function, typically the rectified linear unit (ReLU) or the sigmoid. In this case, the layer is parametrized by $\theta_j = \{W_j, b_j\}$.

Let $\kappa_{\theta_0} : \mathcal{F}_p \times \mathcal{F}_p \to \mathbb{R}$ be parametrized by $\theta_0 \in \Theta_0$. We will construct our kernel network k by

$$k_\theta(x, x') := \kappa_{\theta_0}\left(\varphi^{(L)}_{\theta_L}\left(\varphi^{(L-1)}_{\theta_{L-1}}\Big(\dots\varphi^{(2)}_{\theta_2}\big(\varphi^{(1)}_{\theta_1}(x)\big)\Big)\right), \varphi^{(L)}_{\theta_L}\left(\varphi^{(L-1)}_{\theta_{L-1}}\Big(\dots\varphi^{(2)}_{\theta_2}\big(\varphi^{(1)}_{\theta_1}(x')\big)\Big)\right)\right) \tag{C.2}$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_{L-1}, \theta_L, \theta_0) \in \Theta = \Theta_1 \otimes \Theta_2 \otimes \dots \otimes \Theta_{L-1} \otimes \Theta_L \otimes \Theta_0$ are the collection of all parameters of each layer and the kernel κ.

In order to train the multi-layered representations in an end-to-end fashion, we employ algorithm 1. With a deep architecture, the feature representations the CEN can learn are very flexible, and can work very well for structured data by employing suitable network architectures.

If we choose to employ nonlinear kernels κ, the model architecture is also wide in that an even higher dimensional feature space is implicitly defined on top of the feature space of the last network layer. Despite its supreme flexibility, this again prevents the model from being scalable. We therefore turn to the specific case where we employ a linear kernel κ on top of the multi-layered features.

C.3 Explicit Conditional Embedding Network

The explicit CEN is simply a special case at the intersection of the explicit MCE and the CEN. From the explicit MCE perspective, we simply choose $\varphi_\theta(x) = \varphi^{(L)}_{\theta_L}\big(\varphi^{(L-1)}_{\theta_{L-1}}(\dots\varphi^{(2)}_{\theta_2}(\varphi^{(1)}_{\theta_1}(x)))\big)$. From the CEN perspective, we simply choose $\kappa(z, z') = z^T z'$ to be a linear kernel.

This model architecture is a very practical and powerful form of the MCE. By having a deep architecture, the classifier is still capable of learning flexible representations on structured data, while being able to scale to larger datasets due to the linear kernel at the output layer, provided that the dimensionality of the last layer is relatively small compared to the size of the dataset.

As a subclass of explicit MCE, we can employ algorithm C.1 to learn the multi-layered features effectively. In fact, by not mapping the multi-layered features into a nonlinear kernel, the gradients for each network weight and bias are usually more pronounced, and learning is usually faster in comparison. This approach was used to train the neural network features in our experiments.

References

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312.

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*. SIAM.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell Labs Technical Journal*, 30(1):50–64.

Song, L., Fukumizu, K., and Gretton, A. (2013). Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111.

Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM.