

Consistency results for contrastive learning via Rademacher complexity

Dennis Christensen^{1,2}

¹Department of Mathematics, University of Oslo, Oslo, 1053, Norway

²Defence Systems Division, Norwegian Defence Research Establishment, Kjeller, 2007, Norway

Correspondence

Dennis Christensen, Defence Systems Division, Norwegian Defence Research Establishment, Kjeller, Norway
Email: dennis.christensen@ffi.no

Funding information

Nothing to declare

Contrastive learning (CL) has become one of the key tools for statistical inference in problems where the likelihood function is intractable. Its fundamental idea is to infer a distribution p by introducing a simpler yet similar distribution q , and training a discriminator d to discriminate between samples from p and q , thus “learning by contrasting”. Similar to M-estimators, the consistency of CL rests on the uniform convergence of the empirical risk function. As real applications of CL commonly involve complex parametrisations of p , or complicated discriminators such as neural networks, this condition is too difficult to verify in practice. In this paper, introduce a technique for verifying the uniform convergence condition based on bounding the Rademacher complexity of the discriminator model class. This new technique is applied for models where d is a neural network, and a recently introduced model where p is a mixture of exponential families.

KEYWORDS

Consistency, contrastive learning, neural networks, Rademacher complexity

1 | INTRODUCTION

Let $\mathcal{X} \subseteq \mathbb{R}^m$ be a sample space and suppose that we wish to estimate an unknown probability density p on $\mathcal{X}$ from simulated data. If p cannot be calculated efficiently, for example due to an intractable normalisation constant, then standard methods such as maximum likelihood theory fail. Over the course of the preceding decade, the principle of

contrastive learning has grown to become an extensively popular tool for solving such problems. It has been used for parameter estimation in non-normalised models (Gutmann and Hyvärinen, 2010, 2012; Ceylan and Gutman, 2018; Rhodes and Gutmann, 2019; Matsuda and Hyvärinen, 2019; Gutmann et al., 2022), likelihood-free Bayesian inference (Gutmann et al., 2018; Rogers-Smith et al., 2018; Dinev and Gutman, 2018; Hermans et al., 2020; Thomas et al., 2022) and deep generative models in machine learning (Goodfellow et al., 2014; Dumoulin et al., 2016). The basic idea of contrastive learning is to introduce a separate distribution q and to train a discriminator $d : \mathcal{X} \rightarrow [0, 1]$, which takes a sample X as input, to return 1 if $X \sim p$ and 0 if $X \sim q$. As we shall see shortly, this yields an implicit estimation of p , bypassing explicit evaluations of the intractable likelihood function.

Formulated as a supervised learning problem, we have training data $X_1, \dots, X_n \sim \frac{1}{2}p + \frac{1}{2}q$, and labels $Y_1, \dots, Y_n \in \{0, 1\}$, where $Y_i = 1$ if $X_i \sim p$ and $Y_i = 0$ if $X_i \sim q$. We then train the discriminator $d(x)$ on this data using the binary cross-entropy loss function

$$L(y, d(x)) = -y \log d(x) - (1 - y) \log(1 - d(x)). \quad (1)$$

That is, we aim to minimise the the risk, or the expected loss, given by the functional

$$J[d] = \mathbb{E}[L(Y, d(X))] = -\frac{1}{2} \mathbb{E}_p \log(d(X)) - \frac{1}{2} \mathbb{E}_q \log(1 - d(X)).$$

We write

$$d^*(x) = \arg \min_{d: \mathcal{X} \rightarrow [0, 1]} J[d]$$

for the optimal discriminator. Given no further restrictions on which functions $d : \mathcal{X} \rightarrow [0, 1]$ are allowed, d^* can be written explicitly as

$$d^*(x) = \frac{p(x)}{p(x) + q(x)}.$$

To see why this holds, note that

$$J[d] = -\frac{1}{2} \int \{\log d(x)p(x) + \log(1 - d(x))q(x)\} dx,$$

and this integrand is pointwise optimised by $d^*(x) = p(x)/(p(x) + q(x))$.

The infinite-dimensional space of all functions from $\mathcal{X}$ to $[0, 1]$ is too large to optimise over directly, so instead, we parametrise the discriminator $d(x; \theta)$ by some parameter $\theta \in \Theta$, and minimise the risk as a function of θ , declaring $J(\theta) = J[d(x; \theta)]$. It is assumed that the optimal discriminator d^* is obtainable within this parametrisation (Gutmann and Hyvärinen, 2012). That is, writing

$$\theta^* = \arg \min_{\theta} J(\theta),$$

we assume that $d^*(x) = d(x; \theta^*)$.

In practice, the true risk is an intractable expectation, so the empirical risk, given by

$$J_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \{y_i \log d(x_i; \theta) + (1 - y_i) \log(1 - d(x_i; \theta))\}$$

is minimised instead. We write

$$\hat{\theta}_n = \arg \min_{\theta} J_n(\theta)$$

for the empirical risk minimiser.

It is useful to introduce the notation $a(x; \theta) = \text{logit}(d(x; \theta))$, where the logit function, $\text{logit}(d) = \log(d/(1-d))$. This means that $d(x; \theta) = \sigma(a(x; \theta))$, where σ is the logistic sigmoid function, $\sigma(a) = 1/(1 + \exp(-a))$. Now, estimating θ^* by $\hat{\theta}_n$ yields

$$d(x; \hat{\theta}_n) \approx d(x; \theta^*) = \frac{p(x)}{p(x) + q(x)},$$

and so

$$a(x; \hat{\theta}_n) \approx a(x; \theta^*) = \log p(x) - \log q(x),$$

thus yielding the estimate $p(x) \approx q(x) \exp(a(x; \hat{\theta}_n))$ for the unknown density p . This summarises the fundamental idea of contrastive learning.

Remark In practice, the training data may not contain as many samples from p as from q . This would mean that $X \sim \alpha p + (1 - \alpha)q$, where $\alpha \in (0, 1)$ is a fixed proportion. Although this is a common consideration in the contrastive learning literature (Gutmann and Hyvärinen, 2012; Thomas et al., 2022), it makes to considerations of this paper and so we only consider the case $\alpha = 1/2$ for ease of notation.

Remark The interpretations of the function $a(x; \theta)$ are somewhat different across the contrastive learning literature. Here, we view $a(x; \theta)$ as a function to be parametrised explicitly as e.g. a neural network or a logistic regression model, akin to Hermans et al. (2020); Thomas et al. (2022). On the other hand, several authors (Gutmann and Hyvärinen, 2012; Matsuda and Hyvärinen, 2019) impose an explicit model for the distributions $p(x)$ and $q(x)$ instead, say $p_m(x; \theta)$ and $q_n(x; \theta)$, which yields an implicit parametrisation of a , namely $a(x; \theta) = \log p_m(x; \theta) - \log q_n(x; \theta)$. Note that our assumption that $d^*(x) = d(x; \theta^*)$ now translates into the assumption that $p(x) = p_m(x; \theta^*)$, which indeed is what is assumed by Gutmann and Hyvärinen (2012).

A natural question to ask at this point is whether $\hat{\theta}_n$ is a consistent estimator. That is, whether $\hat{\theta}_n \xrightarrow{P} \theta^*$, where $\xrightarrow{P}$ denotes convergence in probability. Gutmann and Hyvärinen (2012) provided sufficient conditions for consistency, summarised in the following theorem.

Theorem 1 *Suppose that*

1. q is nonzero whenever p is nonzero,
2. $\sup_{\theta \in \Theta} |J_n(\theta) - J(\theta)| \xrightarrow{P} 0$,

3. the matrix

$$\mathcal{I} = \int \mathbf{g}(x) \mathbf{g}(x)^\top \frac{p(x)q(x)}{p(x) + q(x)} dx$$

has full rank, where $\mathbf{g}(x) = \nabla_\theta a(x; \theta)|_{\theta=\theta^*}$.

Then $\hat{\theta}_n \xrightarrow{P} \theta^*$.

Proof See Gutmann and Hyvärinen (2012, Appendix A. Proof A.3).

As pointed out by Gutmann and Hyvärinen (2012), conditions 2 and 3 have analogous counterparts in the theory of maximum likelihood estimation, which is a special case of that of M -estimators. See for example Van der Vaart (1998, Chapter 5. Theorem 5.7).

The aim of this paper is to simplify condition 2, which is too difficult to verify in practice. Indeed, it is rarely verified for applications considered in the existing literature. Henceforth, we shall refer to condition 2 as the uniform convergence condition.¹ Perhaps the most standard approach for proving uniform convergence of this kind is to employ the uniform law of large numbers, which, in the case of contrastive learning, asserts that the uniform convergence condition holds if the three following conditions are satisfied:

- Θ is compact,
- For all $\theta \in \Theta$, the function $(x, y) \mapsto L(y, d(x; \theta))$ is continuous and measurable,
- There exists a dominating function $D(x, y)$ such that $E[D(X, Y)] < \infty$, and

$$|L(y, d(x; \theta))| \leq D(x, y) \text{ for all } \theta \in \Theta.$$

Although the first two of these conditions are commonly satisfied, constructing a dominating function for the third condition is usually too challenging, especially if the parametrisation of the discriminator $d(x; \theta)$ is complicated. It is therefore necessary to look for other methods of establishing uniform convergence.

Riou-Durand and Chopin (2018) verified the uniform convergence condition for the case where $p(x)$ is an exponential family. This result was extended by Bell (2021), who verified the condition for all cases where the empirical risk function J_n is convex. However, neither of these assumptions hold in common cases such as when $a(x; \theta)$ is a neural network, or when $p(x)$ is a mixture model.

In the present paper, we verify the uniform convergence condition in contrastive learning for a wide variety of models by bounding the Rademacher complexity of their model classes. In the machine learning literature, Rademacher complexity has become a highly useful tool for providing non-asymptotic bounds in expectation for supervised learning problems, particularly since it is able to successfully bound statistical complexity even for complicated models such as decision trees, deep neural networks and support vector machines (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002; Gao and Zhou, 2016). When combined with Markov's inequality, we will see that these non-asymptotic bounds establish uniform convergence of $J_n(\theta)$.

The remainder of the paper is structured as follows. In Section 2, we introduce the notion of Rademacher complexity and show how it can be utilised to verify the uniform convergence condition. In Section 3, we consider the case where $a(x; \theta)$ is parametrised as a neural network, and in Section 4, we look at non-normalised mixtures of exponential families. We conclude and discuss further work in Section 5.

¹Not to be confused with the uniform convergence property, see Shalev-Shwartz and Ben-David (2014, Chapter 4. Definition 4.3).

2 | RADEMACHER COMPLEXITY AND UNIFORM CONVERGENCE

In this section, we introduce the notion of Rademacher complexity and connect it with the uniform convergence condition.

Definition The Rademacher complexity of a subset $\mathcal{T} \subseteq \mathbb{R}^n$ is defined as

$$\text{Rad}(\mathcal{T}) = \mathbb{E} \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \Omega_i t_i,$$

where $\Omega_1, \dots, \Omega_n$ are independent Rademacher random variables. That is, $\mathbb{P}(\Omega_i = 1) = \mathbb{P}(\Omega_i = -1) = 1/2$, independently.

We shall be particularly interested in the (random) sets

$$\begin{aligned} \mathcal{A} \circ X &= \{(a(X_1; \theta), \dots, a(X_n; \theta)) \in \mathbb{R}^n : \theta \in \Theta\} \\ \mathcal{L} \circ (X, Y) &= \{(L(Y_1, d(X_1; \theta)), \dots, L(Y_n, d(X_n; \theta))) \in \mathbb{R}^n : \theta \in \Theta\} \\ &= \{(L(Y_1, \sigma(a(X_1; \theta))), \dots, L(Y_n, \sigma(a(X_n; \theta)))) \in \mathbb{R}^n : \theta \in \Theta\}. \end{aligned}$$

By bounding the Rademacher complexities of these sets, we will be able to provide a bound in expectation for $\sup_{\theta \in \Theta} |J_n(\theta) - J(\theta)|$. This is done via a series of standard results from non-asymptotic learning theory, which we now present.

Lemma 2 We have

$$\mathbb{E} \left[\sup_{\theta \in \Theta} |J_n(\theta) - J(\theta)| \right] \leq 4 \mathbb{E} \text{Rad}(\mathcal{L} \circ (X, Y)).$$

Proof See for example Shalev-Shwartz and Ben-David (2014, Chapter 26. Lemma 26.2). Our upper bound contains a factor of 4 rather than 2 to account for the absolute value in the expectation.

Next, we state the contraction lemma, originally due to Ledoux and Talagrand (1991).

Lemma 3 (Contraction lemma) Let $\mathcal{T} \subseteq \mathbb{R}^n$ be a set and let $f_i: \mathbb{R} \rightarrow \mathbb{R}$ be γ -Lipschitz functions, for $i = 1, \dots, n$. Write

$$(f_1, \dots, f_n) \circ \mathcal{T} = \{(f_1(t_1), \dots, f_n(t_n)) : (t_1, \dots, t_n) \in \mathcal{T}\}.$$

Then

$$\text{Rad}((f_1, \dots, f_n) \circ \mathcal{T}) \leq \gamma \text{Rad}(\mathcal{T}).$$

Proof See for example Shalev-Shwartz and Ben-David (2014, Chapter 26. Lemma 26.9).

We are particularly interested in the case where f plays the role of the loss function.

Lemma 4 Suppose that for any $y \in \{0, 1\}$, the function $a \mapsto L(y, \sigma(a))$ is γ -Lipschitz. Then

$$\text{Rad}(\mathcal{L} \circ (X, Y)) \leq \gamma \text{Rad}(\mathcal{A} \circ X).$$

Proof By the contraction lemma,

$$\text{Rad}(\mathcal{L} \circ (X, Y)) = \text{Rad}((L(Y_1, \sigma(\cdot)), \dots, L(Y_n, \sigma(\cdot))) \circ \mathcal{A} \circ X) \leq \text{Rad}(\mathcal{A} \circ X),$$

as required.

We now show that Lemma 4 applies to the binary cross-entropy loss function (1).

Lemma 5 Given a fixed $y \in \{0, 1\}$, let $\tau_y : \mathbb{R} \rightarrow \mathbb{R}$ be the map $\tau_y(a) = L(y, \sigma(a))$, where L is the binary cross-entropy loss function (1). Then τ_y is 1-Lipschitz.

Proof Differentiating, we have

$$\left| \frac{d}{da} \tau_y(a) \right| = \left| -y \frac{1}{\sigma(a)} \sigma(a) (1 - \sigma(a)) + (1 - y) \frac{1}{1 - \sigma(a)} \sigma(a) (1 - \sigma(a)) \right| = |\sigma(a) - y| \leq 1,$$

which implies that τ_y is 1-Lipschitz.

We are now ready to connect the uniform convergence condition with the Rademacher complexity of the set $\mathcal{A} \circ X$.

Proposition 6 The uniform convergence condition is satisfied if $\mathbb{E} \text{Rad}(\mathcal{A} \circ X) \rightarrow 0$ as $n \rightarrow \infty$.

Proof Let $\varepsilon > 0$. Combining Markov's inequality with Lemmas 2 and 4, we have

$$\mathbb{P} \left(\sup_{\theta \in \Theta} |J_n(\theta) - J(\theta)| > \varepsilon \right) \leq \frac{1}{\varepsilon} \mathbb{E} \left[\sup_{\theta \in \Theta} |J_n(\theta) - J(\theta)| \right] \leq \frac{4}{\varepsilon} \mathbb{E} \text{Rad}(\mathcal{L} \circ (X, Y)) \leq \frac{4}{\varepsilon} \mathbb{E} \text{Rad}(\mathcal{A} \circ X) \rightarrow 0,$$

proving that $\sup_{\theta \in \Theta} |J_n(\theta) - J(\theta)| \xrightarrow{P} 0$.

Proposition 6 thus simplifies the problem of verifying the uniform convergence condition for $\hat{\theta}_n$ to showing that $\mathbb{E} \text{Rad}(\mathcal{A} \circ X) \rightarrow 0$. We shall now prove this latter statement in two different model classes, beginning with the case where $a(x; \theta)$ is parametrised as a neural network.

3 | NEURAL NETWORKS

Parametrising $a(x; \theta)$ as a neural network in order to estimate the (log) ratio $\log p(x) - \log q(x)$ has become a popular technique in the contrastive learning literature (Cranmer et al., 2015; Rogers-Smith et al., 2018; Hermans et al., 2020), mostly due to neural networks' ability to replicate complex mathematical functions (see Raghu et al. (2016) for a useful survey). Throughout this section, we let $a(x; \theta)$ be an L -layered dense neural network with λ -Lipschitz activation functions. That is, $a(x; \theta)$ is the composition of L functions of the form $a_l(x) = \sigma_l(W_l x + b_l)$, where the activations

σ_l are λ -Lipschitz, for $l = 1, \dots, L$. The parameters θ thus comprise the weights W_l and the biases b_l , for $l = 1, \dots, L$. Optimisation of $J_n(\theta)$ will take place over the compact space

$$\Theta = \{\theta : \|W_l\|_\infty \leq \omega, |b_l| \leq \beta, l = 1, \dots, L\}, \quad (2)$$

where $\omega, \beta > 0$ are fixed constants and $\|\cdot\|_\infty$ denotes the maximum norm. There exist multiple bounds for the Rademacher complexity of this neural network class (Bartlett and Mendelson, 2002; Neyshabur et al., 2015; Golowich et al., 2018). We shall use the following bound, taken from Rebeschini (2018).

Proposition 7 *There exist constants $\alpha_1 = \alpha_1(L, \omega, \beta, \lambda)$ and $\alpha_2 = \alpha_2(L, \omega, \beta, \lambda, m)$ not depending on n , such that*

$$\text{Rad}(\mathcal{A} \circ X) \leq \frac{1}{\sqrt{n}} (\alpha_1 + \alpha_2 M_n).$$

Proof See for example Rebeschini (2018, Proposition 3.6).

We are now able to easily verifiable sufficient conditions for the uniform convergence condition to hold.

Proposition 8 *Let Θ be as in (2) and suppose that there exists $\eta > 0$ such that*

$$\mathbb{E}_p[\|X\|_\infty^{2+\eta}] < \infty \text{ and } \mathbb{E}_q[\|X\|_\infty^{2+\eta}] < \infty.$$

Then the uniform convergence condition holds for $\hat{\theta}_n$.

Proof From our assumption, it follows that

$$\mathbb{E}_{\frac{1}{2}p + \frac{1}{2}q}[\|X\|_\infty^{2+\eta}] = M < \infty.$$

As the map $x \mapsto x^{2+\eta}$ is convex on the domain $[0, \infty)$, Jensen's inequality gives

$$\mathbb{E}[M_n]^{2+\eta} \leq \mathbb{E}[M_n^{2+\eta}] \leq \sum_{i=1}^n \mathbb{E}_{\frac{1}{2}p + \frac{1}{2}q}[\|X_i\|_\infty^{2+\eta}] \leq nM. \quad (3)$$

Thus, $\mathbb{E}[M_n] \leq (nM)^{1/(2+\eta)}$, and so crucially, $\mathbb{E}[M_n]/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$. Now apply Propositions 6 and 7.

4 | MIXTURES OF EXPONENTIAL FAMILIES

We now verify the uniform convergence condition for non-normalised mixtures of exponential families, studied by Matsuda and Hyvärinen (2019). In this model, $a(x; \theta)$ is parametrised implicitly through explicit parametrisations $p_m(x; \theta)$ and $q_n(x; \theta)$ for $p(x)$ and $q(x)$, respectively. Both p_m and q_n are modelled as non-normalised mixtures of exponential families, namely

$$p_m(x; \theta) = h(x) \sum_{k=1}^K \exp(\phi_k^\top f(x)), \quad q_n(x; \theta) = h(x) \sum_{l=1}^L \exp(w_l^\top f(x)),$$

where $f : \mathcal{X} \rightarrow \mathbb{R}^{d+1}$ is a pretrained neural network and $h : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown function. Compared with the notation of Matsuda and Hyvärinen (2019, Equations (26) and (28)), we have augmented the vector $f(x)$ to include the constant term 1 in order to absorb the normalisation constants c_k and $\exp(M_1/M_l)$ into the inner products $\phi_k^\top f(x)$ and $w_l^\top f(x)$, respectively. We consider optimisation of $J_n(\theta)$ over the compact space

$$\Theta = \{\theta : \|\phi_k\|_2 \leq a, \|w_l\|_2 \leq b, k = 1, \dots, K, l = 1, \dots, L\}, \quad (4)$$

where $a, b > 0$ are fixed constants and $\|\cdot\|_2$ denotes the Euclidean norm. We are now ready to prove an analogous result to Proposition 8 which requires no further assumptions about the moments of $\mathcal{X}$.

Proposition 9 *Let Θ be as in (4). Then the uniform convergence condition holds for $\hat{\theta}_n$.*

Proof The proof will follow a similar strategy to that of Proposition 8. First, we consider the variable $M_n = \|f(X)\|_\infty$, where $X \sim \frac{1}{2}p + \frac{1}{2}q$. Note that $\mathbb{E}[M_n^4] = M < \infty$, as the fourth moment of the sufficient statistics of an exponential family just equal the fourth derivative of its cumulant function. Thus, applying Jensen's inequality like in (3) with $\eta = 2$, we get that $\mathbb{E}[M_n] \leq (nM)^{1/4}$.

Next, we find a suitable bound on the Rademacher complexity $\text{Rad}(\mathcal{A} \circ \mathcal{X})$. Note that

$$\begin{aligned} n \text{Rad}(\mathcal{A} \circ \mathcal{X}) &= \mathbb{E}_\Omega \sup_{\theta \in \Theta} \sum_{i=1}^n \Omega_i \left\{ \log \left(h(X_i) \sum_{k=1}^K \exp(\phi_k^\top f(X_i)) \right) - \log \left(h(X_i) \sum_{l=1}^L \exp(w_l^\top f(X_i)) \right) \right\} \\ &= \mathbb{E}_\Omega \sup_{\theta \in \Theta} \sum_{i=1}^n \Omega_i \left\{ \log \left(\sum_{k=1}^K \exp(\phi_k^\top f(X_i)) \right) - \log \left(\sum_{l=1}^L \exp(w_l^\top f(X_i)) \right) \right\} \\ &\leq \mathbb{E}_\Omega \sup_{\|\phi_k\| \leq a} \sum_{i=1}^n \Omega_i \log \left(\sum_{k=1}^K \exp(\phi_k^\top f(X_i)) \right) + \mathbb{E}_\Omega \sup_{\|w_l\| \leq b} \sum_{i=1}^n \Omega_i \log \left(\sum_{l=1}^L \exp(w_l^\top f(X_i)) \right) \end{aligned} \quad (5)$$

Consider the first of these two terms. To simplify notation, let $a_{ik} = \phi_k^\top f(X_i)$, for $i = 1, \dots, n$, $k = 1, \dots, K$. Also, write

$$\text{LogSumExp}(a_{i1}, \dots, a_{iK}) = \log \left(\sum_{k=1}^K \exp(a_{ik}) \right),$$

for $i = 1, \dots, n$. The LogSumExp function has the softmax function as its gradient, and is therefore 1-Lipschitz with respect to the maximum norm $\|\cdot\|_\infty$. Consequently, it is $\sqrt{K}$ -Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$, and so by the vector-contraction inequality by Maurer (2016, Corollary 4), we have

$$\begin{aligned} \mathbb{E}_\Omega \sup_{\|\phi_k\| \leq a} \sum_{i=1}^n \Omega_i \log \left(\sum_{k=1}^K \exp(a_{ik}) \right) &= \mathbb{E}_\Omega \sup_{\|\phi_k\| \leq a} \sum_{i=1}^n \Omega_i \text{LogSumExp}(a_{i1}, \dots, a_{iK}) \\ &\leq \sqrt{2K} \mathbb{E}_\Omega \sup_{\|\phi_k\| \leq a} \sum_{i=1}^n \sum_{k=1}^K \Omega_{ik} a_{ik}, \end{aligned}$$

where Ω_{ik} are independent Rademacher random variables for $i = 1, \dots, n$ and $k = 1, \dots, K$. By the Cauchy-Schwarz inequality, we have

$$\sum_{i=1}^n \sum_{k=1}^K \Omega_{ik} a_{ik} = \sum_{k=1}^K \phi_k^\top \left(\sum_{i=1}^n \Omega_{ik} f(X_i) \right) \leq \sum_{k=1}^K \|\phi_k\|_2 \left\| \sum_{i=1}^n \Omega_{ik} f(X_i) \right\|_2 \leq aK \left\| \sum_{i=1}^n \Omega_i f(X_i) \right\|_2,$$

and so

$$\mathbb{E}_{\Omega} \sup_{\|\phi\| \leq a} \sum_{i=1}^n \sum_{k=1}^K \Omega_{ik} \phi_k^{\top} f(X_i) \leq aK \mathbb{E}_{\Omega} \left\| \sum_{i=1}^n \Omega_i f(X_i) \right\|_2.$$

Next, by Jensen's inequality,

$$\mathbb{E}_{\Omega} \left\| \sum_{i=1}^n \Omega_i f(X_i) \right\|_2 = \mathbb{E}_{\Omega} \sqrt{\left\| \sum_{i=1}^n \Omega_i f(X_i) \right\|_2^2} \leq \sqrt{\mathbb{E}_{\Omega} \left\| \sum_{i=1}^n \Omega_i f(X_i) \right\|_2^2} = \sqrt{\mathbb{E}_{\Omega} \sum_{j=1}^d \left(\sum_{i=1}^n \Omega_i f_j(X_i) \right)^2}.$$

Since the Ω_i are independent and $\mathbb{E} \Omega_i = 0$ for all i , we further have

$$\sqrt{\mathbb{E}_{\Omega} \sum_{j=1}^d \left(\sum_{i=1}^n \Omega_i f_j(X_i) \right)^2} = \sqrt{\mathbb{E}_{\Omega} \sum_{j=1}^d \sum_{i=1}^n (\Omega_i f_j(X_i))^2} = \sqrt{\mathbb{E}_{\Omega} \sum_{i=1}^n \|f(X_i)\|_2^2} \leq \sqrt{dn} \max_{1 \leq i \leq n} \|f(X_i)\|_{\infty} = \sqrt{dn} M_n.$$

Hence, putting everything together, we have

$$\mathbb{E}_{\Omega} \sup_{\|\phi_k\| \leq a} \sum_{i=1}^n \Omega_i \log \left(\sum_{k=1}^K \exp(\phi_k^{\top} f(X_i)) \right) \leq a\sqrt{2dn} K^{3/2} M_n.$$

Similarly, for the latter term in (5), we get

$$\mathbb{E}_{\Omega} \sup_{\|w_l\| \leq b} \sum_{i=1}^n \Omega_i \log \left(\sum_{l=1}^L \exp(w_l^{\top} f(X_i)) \right) \leq b\sqrt{2dn} L^{3/2} M_n,$$

and so finally

$$\mathbb{E} \text{Rad}(\mathcal{A} \circ X) \leq \sqrt{2d} (aK^{3/2} + bL^{3/2}) \frac{1}{\sqrt{n}} \mathbb{E} M_n \leq \sqrt{2d} (aK^{3/2} + bL^{3/2}) M^{1/4} \frac{1}{n^{1/4}} \rightarrow 0$$

as $n \rightarrow \infty$. Now apply Proposition 6.

5 | DISCUSSION

We have introduced a technique for verifying the uniform convergence condition for contrastive learning via bounding the Rademacher complexity of the discriminator model class. We applied this technique in two settings. In the former, the non-activated discriminator a was parametrised as a neural network, which required the finiteness of the $(2+\eta)$ th moment of X . In the latter setting we considered parameter estimation for non-normalised mixtures of exponential families, and verified the uniform convergence condition without any further assumptions about X . Like the uniform law of large numbers, the new approach requires the parameter space to be compact. However, it obviates the need to find a dominating function, which is too complicated to do in practice. Future investigations into Rademacher complexity bounds which do not require compactness are encouraged.

Since many other estimation methods other than contrastive learning also require a uniform convergence condition for consistency (see M -estimators, see e.g. Van der Vaart (1998, Chapter 5)), it is worth exploring whether

the technique presented here may be applicable elsewhere. This may be of particular interest in estimation problems which involve neural networks or other model classes for which the Rademacher complexity can be bounded appropriately.

acknowledgements

Parts of this result first appeared in my Master's dissertation (Christensen, 2021). I therefore wish to thank Professor Geoff Nicholls, my supervisor at the time, for his guidance. I also want to thank Ingrid Dæhlen and Cameron Bell for useful discussions.

references

- Bartlett, P. L. and Mendelson, S. (2002) Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.
- Bell, C. (2021) *Contrastive learning for exponential random graph models*. Unpublished master's dissertation, Oxford University.
- Ceylan, C. and Gutman, M. U. (2018) Conditional noise-contrastive estimation of unnormalised models. In *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 726–734. PMLR.
- Christensen, D. (2021) *Contrastive learning for doubly intractable problems in Bayesian nonparametrics with censored data*. Unpublished master's dissertation, Oxford University.
- Cranmer, K., Pavez, J. and Louppe, G. (2015) Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv:1506.02169*.
- Dinev, T. and Gutman, M. U. (2018) Dynamic likelihood-free inference via ratio estimation (DIRE). *arXiv:1810.09899*.
- Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M. and Courville, A. (2016) Adversarially learned inference. *arXiv:1606.00704*.
- Gao, W. and Zhou, Z.-H. (2016) Dropout Rademacher complexity of deep neural networks. *Science China Information Sciences*, **59**.
- Golowich, N., Rakhlin, A. and Shamir, O. (2018) Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference On Learning Theory*, vol. 75, 297–299. PMLR.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative adversarial nets. *Advances in neural information processing systems*, **27**.
- Gutmann, M. and Hyvärinen, A. (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, 297–304. PMLR.
- Gutmann, M. U., Dutta, R., Kaski, S. and Corander, J. (2018) Likelihood-free inference via classification. *Statistics and Computing*, **28**, 411–425.
- Gutmann, M. U. and Hyvärinen, A. (2012) Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, **13**, 307–361.
- Gutmann, M. U., Kleinegesse, S. and Rhodes, B. (2022) Statistical applications of contrastive learning. *Behaviormetrika*, **49**, 277–301.

- Hermans, J., Volodimir, B. and Louppe, G. (2020) Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 4239–4248. PMLR.
- Koltchinskii, V. and Panchenko, D. (2000) Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, vol. 47 of *Progress in Probability*, 443–457. Springer.
- Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces: Isoperimetry and Processes*, vol. 23 of *Classics in Mathematics*. Springer Berlin, Heidelberg.
- Matsuda, T. and Hyvärinen, A. (2019) Estimation of non-normalized mixture models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2555–2563. PMLR.
- Maurer, A. (2016) A vector-contraction inequality for Rademacher complexities. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory*, 3–17. Springer.
- Neysshabur, B., Tomioka, R. and Srebro, N. (2015) Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory*, vol. 40, 1376–1401. PMLR.
- Raghu, M., Poole, B., Kleinberg, J., Ganguli, S. and Solh-Dickstein, J. (2016) Survey of expressivity in deep neural networks. *arXiv:1611.08083*.
- Rebeschini, P. (2018) Lecture notes in algorithmic foundations of learning.
- Rhodes, B. and Gutmann, M. U. (2019) Variational noise-contrastive estimation. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, vol. 89, 2741–2750. PMLR.
- Riou-Durand, L. and Chopin, N. (2018) Noise contrastive estimation: asymptotics, comparison with MC-MLE. *arXiv:1801.10381*.
- Rogers-Smith, C., Pesonen, H. and Kaski, S. (2018) Approximate Bayesian computation via population Monte Carlo and classification. *arXiv:1810.12233*.
- Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
- Thomas, O., Dutta, R., Corander, J., Kaski, S. and Gutmann, M. U. (2022) Likelihood-free inference by ratio estimation. *Bayesian Analysis*, **17**, 1–31.
- Van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge University Press.