Local Differential Privacy for Mixtures of Experts

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce a new approach to the mixture of experts model that consists in imposing local differential privacy on the gating mechanism. This is theoretically justified by statistical learning theory. Notably, we provide generalization bounds specifically tailored for mixtures of experts, leveraging the one-out-of-n gating mechanism rather than the more common n-out-of-n mechanism. Moreover, through experiments, we show that our approach improves the generalization ability of mixtures of experts.

1 Introduction

Mixtures of experts, initially introduced by Jacobs et al. [1991], have found widespread use in modeling sequential data, including applications in classification, regression, pattern recognition and feature selection tasks (Städler et al. [2010] and Khalili and Lin [2013]). One of the fundamental motivations behind mixtures of experts is their ability to break down complex problems into more manageable sub-problems, potentially simplifying the overall task. The structure of these models is well suited to capturing unobservable heterogeneity in the data generation process, dealing with this problem by splitting the data into homogeneous subsets (with the gating network) and associating each subset with an expert. This intuitive architecture has led to significant interest in mixture of experts models, resulting in a wealth of research (Yuksel et al. [2012]), ranging from simple mixtures of experts (Jacobs et al. [1991], Jordan and Jacobs [1993]) to sparsely gated models (Shazeer et al. [2017]). Moreover, this architecture has inspired the development of various other models, such as switch transformers (Fedus et al. [2022]). However, despite the considerable attention mixtures of experts have received, advancements in their theoretical analysis have been relatively limited. Azran and Meir [2004] proved data-dependent risk bounds for mixtures of experts (with the n-out-of-n gating mechanism) using Rademacher complexity, but they exhibit a dependence on the complexity of the class of gating networks and the sum of the complexities of the expert classes, which reflects the complex structure of mixtures of experts but unfortunately leads to potentially large bounds. We are not aware of other work proving generalization bounds specifically tailored to mixtures of experts.

To make theoretical progress, we utilize a well-known privacy-preserving technique called Local Differential Privacy (LDP). It was initially introduced by Dwork [2006] and has since been widely used to preserve privacy for individual data points as in Kasiviswanathan et al. [2010]. This is achieved by introducing stochasticity in algorithm outputs to control their dependence on specific inputs. This stochasticity is generally quantified by a positive real number ϵ. In this case, we write ϵ-LDP instead of just LDP. The parameter ϵ quantifies the level of privacy protection in the local differential privacy mechanism. A smaller value indicates stronger privacy protection, which requires the addition of more noise.

In this work, we exploit this noise for regularization in our models by imposing the ϵ-LDP condition on their gating networks. This method allows us to leverage the numerous benefits of the most complex architectures, such as neural networks, without compromising theoretical guarantees on risk. By relying on LDP, we offer tight theoretical guarantees on the risk of mixtures of experts models,

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

provided with the one-out-of-n gating mechanism. Unlike the very few existing guarantees, these bounds depend only logarithmically on the number of experts we have, and the complexity of the gating network only appears in our bounds through the parameter ϵ of the LDP condition.

2 Preliminaries

Let $\mathcal{X}$ be the instance space, $\mathcal{Y}$ the label space, and $\mathcal{Y}'$ the output space (which can be different from $\mathcal{Y}$). As is usual in supervised learning, we assume that data $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are generated independently from an unknown probability distribution $\mathcal{D}$. We consider a training set of m examples $S = ((x_1, y_1), \ldots, (x_m, y_m)) \sim \mathcal{D}^m$ and a bounded loss function $\ell \colon \mathcal{Y}' \times \mathcal{Y} \to [0, 1]$.

2.1 Mixtures of experts

We consider classes $\mathcal{H}_i$ of experts $h_i \colon \mathcal{X} \to \mathcal{Y}'$ for $i = 1, \ldots, n$. Let $\mathcal{G}$ be a set of gating functions $\mathbf{g} \colon \mathcal{X} \to [0, 1]^n$ such that, given any $x \in \mathcal{X}$, we have that $\sum_{i=1}^n g_i(x) = 1$, where $g_i(x)$ is the i-th component of $\mathbf{g}(x)$. This means that each gating function defines a probability distribution on $[n] = \{1, \ldots, n\}$ for each $x \in \mathcal{X}$, where $g_i(x)$ is the probability of i.

In this work, a mixture of experts consists of n experts, $\mathbf{h} = (h_1, \ldots, h_n) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$, a gating function $\mathbf{g} \in \mathcal{G}$ and a gating mechanism that combines the outputs of the experts and the output of the gating function to produce the final output. Our models use the stochastic one-out-of-n gating mechanism, as described in Jacobs et al. [1991]. It is defined as follows: to make a prediction with $(\mathbf{g}, \mathbf{h}) \in \mathcal{G} \times \prod_{i=1}^n \mathcal{H}_i$ given an instance x, draw $i \sim \mathbf{g}(x)$ and output $h_i(x)$. This stochastic predictor has risk and empirical risk defined by, respectively,

$$R(\mathbf{g}, \mathbf{h}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x)} \ell(h_i(x), y), \quad \text{and} \quad R_S(\mathbf{g}, \mathbf{h}) = \frac{1}{m} \sum_{j=1}^m \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x_j)} \ell(h_i(x_j), y_j).$$

The preference for the one-out-of-n gating mechanism over the n-out-of-n mechanism in mixtures of experts is justified by its ability to induce sparsity and noise, enhancing computational efficiency and robustness to overfitting. This sparsity also offers scalability benefits, particularly in large-scale applications, where activating all experts for each input can lead to increased computational and memory requirements as explained in Shazeer et al. [2017] and Jacobs et al. [1991]. Moreover, the one-out-of-n mechanism is more amenable to certain kinds of theoretical analysis, including ours.

2.2 Local Differential Privacy

Definition 2.1. Let $\mathcal{I}$ be a finite set, consider a mechanism that produces an output $i \in \mathcal{I}$, given an input $x \in \mathcal{X}$, with probability $\mathbb{P}(i \,|\, x)$, and let ϵ be a nonnegative real number. Then, the mechanism satisfies the ϵ-Local Differential Privacy (ϵ-LDP) property if and only if

$$\mathbb{P}(i \,|\, x) \leq e^\epsilon \, \mathbb{P}(i \,|\, x') \quad \text{for all } x, x' \in \mathcal{X} \text{ and all } i \in \mathcal{I}.$$

Unless stated otherwise, we assume that each $\mathbf{g} \in \mathcal{G}$ satisfies ϵ-LDP, for some fixed nonnegative real number ϵ. Since we can interpret $\mathbf{g}$ as a random mechanism that, given $x \in \mathcal{X}$, selects $i \in [n]$ with probability $g_i(x)$, the condition of ϵ-LDP amounts to the following:

$$g_i(x) \leq e^\epsilon g_i(x') \quad \text{for all } x, x' \in \mathcal{X} \text{ and all } i \in [n].$$

Since ϵ-LDP is an important condition for all of our theoretical results, we provide a practical way of obtaining gating functions satisfying ϵ-LDP from an arbitrary set $\mathcal{F}$ of bounded functions, in the form of the following theorem.

Theorem 2.2. *Let $b > 0$ and $\beta \geq 0$ be real numbers, and suppose that $\mathcal{F}$ is a set of functions $\mathbf{f} \colon \mathcal{X} \to [-b, b]^n$. Let $\mathcal{G}$ be the set of functions $\mathbf{g} : \mathcal{X} \to [0, 1]^n$ defined by*

$$g_i(x) \;=\; \frac{\exp(\beta f_i(x) + c_i)}{\sum_{k=1}^n \exp(\beta f_k(x) + c_k)}, \quad \textit{where } \mathbf{f} = (f_1, \ldots, f_n) \in \mathcal{F} \textit{ and } (c_1, \ldots, c_n) \in \mathbb{R}^n.$$

Then, each $\mathbf{g} \in \mathcal{G}$ satisfies $4\beta b$-LDP.

Proof. The proof is obtained by performing simple calculations, bounding the ratio $g_i(x)/g_i(x')$, for all $x, x' \in \mathcal{X}$ and all $i \in [n]$. The detailed proof is given in Appendix A. □

3 PAC-Bayesian bounds for mixtures of experts

To apply the PAC-Bayes theory, we need to add a level of stochasticity to our predictors: instead of training experts h_i, we train probability measures Q_i on each expert set $\mathcal{H}_i$. For convenience, we write $Q = Q_1 \otimes \cdots \otimes Q_n$. Now, putting everything together, a mixture of experts $(\mathbf{g}, Q)$ makes predictions as follows: given $x \in \mathcal{X}$, draw $i \sim \mathbf{g}(x)$, then draw $h \sim Q_i$, and finally output $h(x)$. Such a predictor has risk and empirical risk defined by, respectively,

$$R(\mathbf{g}, Q) = \mathop{\mathbb{E}}_{\mathbf{h}\sim Q} R(\mathbf{g}, \mathbf{h}) \quad \text{and} \quad R_S(\mathbf{g}, Q) = \mathop{\mathbb{E}}_{\mathbf{h}\sim Q} R_S(\mathbf{g}, \mathbf{h}).$$

Notice that, though probability distributions have replaced the individual experts, there is no need to define a probability distribution on the gating functions to get a PAC-Bayesian bound. Training a single gating function will do, and, remarkably, Lemma 3.1 below shows that it can be obtained from a very complicated function, such as a neural network, provided we impose ϵ-LDP (for example, with Theorem 2.2).

Finally, let us recall the notion of *Kullback-Leibler (KL) divergence*. Given probability distributions Q_i and P_i on $\mathcal{H}_i$, it is defined by

$$\mathrm{KL}(Q_i \,\|\, P_i) = \begin{cases} \displaystyle\mathop{\mathbb{E}}_{h\sim Q_i} \ln \frac{dQ_i}{dP_i}(h) & \text{if } Q_i \ll P_i \\ \infty & \text{otherwise,} \end{cases}$$

where dQ_i/dP_i is a Radon-Nikodym derivative.

Lemma 3.1. *We consider mixtures of experts as defined in section 2.1 and provided with the one-out-of-n routing mechanism. Let $\Delta : \mathbb{R}^2 \to \mathbb{R}$ be a convex function that is decreasing in its first argument and increasing in its second argument, and let ϵ be a nonnegative real number. Then, for any $\mathbf{g} \in \mathcal{G}$ that satisfies the ϵ-LDP property, for any $Q = Q_1 \otimes \cdots \otimes Q_n$ on $\mathcal{H}_1 \times \cdots \times \mathcal{H}_n$, and for any $x' \in \mathcal{X}$:*

$$\Delta\big(e^{\epsilon} R_S(\mathbf{g}, Q), e^{-\epsilon} R(\mathbf{g}, Q)\big) \leq \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \Delta\big(R_S(Q_i), R(Q_i)\big)$$

where $R(Q_i) = \mathbb{E}_{x\sim\mathcal{D}} \mathbb{E}_{h\sim Q_i} \ell(h(x), y)$ and $R_S(Q_i) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{h\sim Q_i} \ell(h(x_j), y_j)$.

Proof. Since the gating function satisfies ϵ-LDP, we have that $e^{-\epsilon} g_i(x') \leq g_i(x) \leq e^{\epsilon} g_i(x')$ for all $x, x' \in \mathcal{X}$ and all $i \in [n]$. It follows that $e^{\epsilon} R_S(\mathbf{g}, Q) \geq \mathbb{E}_{i\sim\mathbf{g}(x')} R_S(Q_i)$ and $e^{-\epsilon} R(\mathbf{g}, Q) \leq \mathbb{E}_{i\sim\mathbf{g}(x')} R(Q_i)$. Given that Δ is decreasing in its first argument and increasing in its second argument, we find that

$$\Delta\big(e^{\epsilon} R_S(\mathbf{g}, Q), e^{-\epsilon} R(\mathbf{g}, Q)\big) \leq \Delta\Big(\mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} R_S(Q_i), \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} R(Q_i)\Big)$$

Since Δ is a convex function, we can apply Jensen's inequality to the expression on the right-hand side, yielding the desired result. □

Different choices of function Δ will allow us to obtain different PAC-Bayes bounds:

- Let $\Delta(u, v) = v - u$. This is compatible with typical PAC-Bayes bounds on the difference between the true and empirical risks.
- Given $\lambda > 1/2$, let Δ be defined by $\Delta(u, v) = v - \frac{2\lambda}{2\lambda - 1} u$. This choice is compatible with a Catoni-type bound, as we will see below.
- Let Δ be defined by $\Delta(u, v) = \mathrm{kl}(u \,\|\, v) = u \ln \frac{u}{v} + (1 - u) \ln \frac{1-u}{1-v}$. This choice is compatible with a Langford-Seeger-type bound. However, note that the function Δ defined here does not quite obey the hypotheses of lemma 3.1. Indeed, it is only defined for $(u, v) \in [0, 1]^2$, and only has the right monotonicity properties on the set $\{\, (u, v) \in [0, 1]^2 \mid u \leq v \,\}$. We can remedy those defects through small adjustments to the proof.

We prove a generalization bound of Catoni-type as an illustration of the machinery just described.

Theorem 3.2 (Theorem 2 in McAllester [2013]). *Let $\delta \in (0,1)$ and $\lambda > 1/2$. Fix $i \in [n]$, and let P_i be a probability measure on $\mathcal{H}_i$ (chosen without seeing the training data). Then, with probability at least $1-\delta$ over the draws of S, for all probability measures Q_i on $\mathcal{H}_i$, we have that*

$$R(Q_i) \leq \frac{2\lambda}{2\lambda - 1}\left(R_S(Q_i) + \frac{\lambda}{m}\Big(\mathrm{KL}(Q_i \,\|\, P_i) + \ln\frac{1}{\delta}\Big)\right).$$

Theorem 3.3. *Let $\delta \in (0,1)$, $\epsilon \geq 0$, and $\lambda > 1/2$. For each $i \in [n]$, let P_i be a probability measure on $\mathcal{H}_i$ (chosen without seeing the training data). Then, with probability at least $1-\delta$ over the draws of S, for all probability measures $Q = Q_1 \otimes \cdots \otimes Q_n$ on $\mathcal{H}$, all $\mathbf{g} \in \mathcal{G}$ that satisfy ϵ-LDP, and all $x' \in \mathcal{X}$, we have that*

$$R(\mathbf{g}, Q) \leq \frac{2\lambda e^{\epsilon}}{2\lambda - 1}\left(e^{\epsilon} R_S(\mathbf{g}, Q) + \frac{\lambda}{m}\Big(\mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \mathrm{KL}(Q_i \,\|\, P_i) + \ln\frac{n}{\delta}\Big)\right).$$

Proof. By n applications of Theorem 3.2, we have that, for each $i \in [n]$, with probability at least $1-\delta/n$, for all Q_i,

$$R(Q_i) \leq \frac{2\lambda}{2\lambda - 1}\left(R_S(Q_i) + \frac{\lambda}{m}\Big(\mathrm{KL}(Q_i \,\|\, P_i) + \ln\frac{n}{\delta}\Big)\right).$$

We can make all these inequalities (for each $i \in [n]$) hold simultaneously with a union bound. Now, applying Lemma 3.1 with $\Delta(u,v) = v - \frac{2\lambda}{2\lambda-1}u$, we find that, with probability at least $1-\delta$, for all Q, all $\mathbf{g} \in \mathcal{G}$ and all $x' \in \mathcal{X}$, we have that

$$\begin{aligned} e^{-\epsilon} R(\mathbf{g}, Q) - \frac{2\lambda e^{\epsilon}}{2\lambda - 1} R_S(\mathbf{g}, Q) &\leq \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')}\Big(R(Q_i) - \frac{2\lambda}{2\lambda-1} R_S(Q_i)\Big) \\ &\leq \frac{2\lambda^2}{(2\lambda-1)m}\Big(\mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \mathrm{KL}(Q_i \,\|\, P_i) + \ln\frac{n}{\delta}\Big). \end{aligned}$$

□

We also give a bound of Langford-Seeger type, since they are generally recognized as among the tightest PAC-Bayes bounds available, and to prove the flexibility of our approach.

Theorem 3.4. *Let $\delta \in (0,1)$, $\epsilon \geq 0$, and $m \geq 8$. For each $i \in [n]$, let P_i be a probability measure on $\mathcal{H}_i$ (chosen without seeing the training data). Then, with probability at least $1-\delta$ over the draws of S, for all probability measures $Q = Q_1 \otimes \cdots \otimes Q_n$ on $\mathcal{H}$, all $\mathbf{g} \in \mathcal{G}$ that satisfy ϵ-LDP, and all $x' \in \mathcal{X}$, we have that, either $R(\mathbf{g}, Q) < e^{2\epsilon} R_S(\mathbf{g}, Q)$, or*

$$\mathrm{kl}(e^{\epsilon} R_S(\mathbf{g}, Q) \,\|\, e^{-\epsilon} R(\mathbf{g}, Q)) \leq \frac{1}{m}\Big(\mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \mathrm{KL}(Q_i \,\|\, P_i) + \ln\frac{2n\sqrt{m}}{\delta}\Big).$$

Proof. The proof, which is similar to that of Theorem 3.3, is available in Appendix A. □

3.1 Comparison with other bounds

Very few generalizations bound tailored specifically to mixtures of experts appear in the literature, and those we could find do not apply to mixtures of experts with the one-out-of-n gating mechanism. We can, however, compare our bounds to those obtained by naively applying generic PAC-Bayes generalization bounds to mixtures of experts. In this case, we need to consider classifiers of the form $(Q_{\mathcal{G}}, Q)$, where $Q_{\mathcal{G}}$ is a probability measure on $\mathcal{G}$, and $Q = Q_1 \otimes \cdots \otimes Q_n$ is a probability measure on $\mathcal{H}_1 \times \cdots \times \mathcal{H}_n$ as before. Then, note that

$$\mathrm{KL}(Q_{\mathcal{G}} \otimes Q_1 \otimes \cdots \otimes Q_n \,\|\, P_{\mathcal{G}} \otimes P_1 \otimes \cdots \otimes P_n) = \mathrm{KL}(Q_{\mathcal{G}} \,\|\, P_{\mathcal{G}}) + \sum_{i=1}^{n} \mathrm{KL}(Q_i \,\|\, P_i).$$

This means that a generic PAC-Bayes bound applied to mixtures of experts will depend on the sum of the KL divergences corresponding to the gating functions and each of the experts. Obviously, this sum could be very large. By imposing ϵ-LDP to the gating functions as in our approach, we can eliminate the stochasticity associated to the gating functions, and rid our bounds of the (potentially very large) $\mathrm{KL}(Q_{\mathcal{G}} \,\|\, P_{\mathcal{G}})$ term. Instead, it is ϵ-LDP which controls our gating functions to ensure generalization. Furthermore, our bounds replace the sum of the KL divergences of the experts by a $\mathbf{g}(x')$-weighted average, which means we can have many more experts with almost no penalty from the theoretical point of view. Indeed, our bounds only depend on the number n of experts logarithmically, through the use of the union bound.

4 Rademacher bounds for mixtures of experts

Let us start with a slight modification of Lemma 3.1.

Lemma 4.1. *We consider mixtures of experts as defined in section 2.1 and provided with the one-out-of-n routing mechanism. Let $\Delta : \mathbb{R}^2 \to \mathbb{R}$ be a convex function that is decreasing in its first argument and increasing in its second argument, and let ϵ be a nonnegative real number. Then, for any $\mathbf{g} \in \mathcal{G}$ that satisfies the ϵ-LDP property, for any $\mathbf{h} \in \mathcal{H}$, and for any $x' \in \mathcal{X}$:*

$$\Delta\big(e^{\epsilon} R_S(\mathbf{g},\mathbf{h}), e^{-\epsilon} R(\mathbf{g},\mathbf{h})\big) \leq \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \Delta\big(R_S(h_i), R(h_i)\big)$$

where $R(h_i) = \mathbb{E}_{x\sim\mathcal{D}}\, \ell(h_i(x), y)$ and $R_S(h_i) = \frac{1}{m}\sum_{j=1}^{m} \ell(h_i(x_j), y_j)$.

Proof. The proof is similar to that of Lemma 3.1 and is provided in Appendix A. □

Let us now recall the following definition.

Definition 4.2 (Rademacher complexity)**.** Given a space $\mathcal{H}$ of predictors, a loss function ℓ, and a data generating distribution $\mathcal{D}$, the Rademacher complexity $\mathcal{R}(\ell \circ \mathcal{H})$ is defined by

$$\mathcal{R}(\ell \circ \mathcal{H}) \;=\; \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{j=1}^{m} \sigma_j \ell(h(x_j), y_j),$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ is distributed uniformly on $\{-1, 1\}^m$.

Our main theorem will make use of the following well-known risk bound.

Theorem 4.3 (Basic Rademacher risk bound)**.** *Given a $[0, 1]$-valued loss function ℓ, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$, we have that*

$$R(h) \;\leq R_S(h) + 2\mathcal{R}(\ell \circ \mathcal{H}) + \sqrt{\frac{2\ln(2/\delta)}{m}}\,.$$

Theorem 4.4. *Let $\delta \in (0, 1)$ and $\epsilon \geq 0$. Given a $[0, 1]$-valued loss function ℓ, then, with probability at least $1 - \delta$ over the draws of S, for all $\mathbf{h} \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$, for all $\mathbf{g} \in \mathcal{G}$ that satisfy ϵ-LDP, and all $x' \in \mathcal{X}$, we have that*

$$R(\mathbf{g},\mathbf{h}) \leq e^{\epsilon}\left(e^{\epsilon} R_S(\mathbf{g},\mathbf{h}) + 2 \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \mathcal{R}(\ell \circ \mathcal{H}_i) + \sqrt{\frac{2\ln(2n/\delta)}{m}}\right).$$

Proof. By n applications of Theorem 4.3, we have that, for each $i \in [n]$, with probability at least $1 - \delta/n$, for all $h_i \in \mathcal{H}_i$,

$$R(h_i) \leq R_S(h_i) + 2\mathcal{R}(\ell \circ \mathcal{H}_i) + \sqrt{\frac{2\ln(2n/\delta)}{m}}.$$

We can make all these inequalities (for each $i \in [n]$) hold simultaneously with a union bound. Now, applying Lemma 4.1 with $\Delta(u, v) = v - u$, we find that, with probability at least $1 - \delta$, for all $\mathbf{h} \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$, all $\mathbf{g} \in \mathcal{G}$ and all $x' \in \mathcal{X}$, we have that

$$\begin{aligned} e^{-\epsilon} R(\mathbf{g},\mathbf{h}) - e^{\epsilon} R_S(\mathbf{g},\mathbf{h}) &\leq \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \big(R(h_i) - R_S(h_i)\big) \\ &\leq \mathop{\mathbb{E}}_{i\sim\mathbf{g}(x')} \left(2\mathcal{R}(\ell \circ \mathcal{H}_i) + \sqrt{\frac{2\ln(2n/\delta)}{m}}\right). \end{aligned}$$

□

Note, that the risk bound of Theorem 4.4 depends only on the average Rademacher complexity of the classes of experts instead of the sum of their Rademacher complexities. Note also that, as in the previous section, the complexity of $\mathcal{G}$ does not affect the risk bound. Finally, the risk bound does not hold uniformly for all values of ϵ. However, by the union bound, the theorem holds for any fixed set $\{\epsilon_1, \ldots, \epsilon_k\}$ if we replace δ by δ/k. Consequently, this suggests a learning algorithm that minimizes $R_S(\mathbf{g},\mathbf{h})$ for $\epsilon \in \{\epsilon_1, \ldots, \epsilon_k\}$.

Also note that Lemma 4.1 allows us to obtain risk bounds for mixtures of experts as long as we have bounds on $\Delta\big(R_S(h_i), R(h_i)\big)$ which hold with high probability, whether they are based on Rademacher complexity, margins, VC dimension, or algorithmic stability.

4.1 The need to use adaptive experts

Following these theoretical results, we may be tempted to use a gating network satisfying ϵ-LDP to accomplish a learning task all by itself using non-adaptive experts, that is, experts h_i each taking a constant value, no matter the input: $h_i(x) = i$ for all $x \in \mathcal{X}$. In that case, each Rademacher complexity $\mathcal{R}(\ell \circ \mathcal{H}_i)$ is zero and we can show that Theorem 4.4 can become vacuous under reasonable circumstances.

Consider, for example, the binary classification case with the 0-1 loss. In that case, we have two experts h_{+1} and h_{-1} such that $h_{+1}(x) = +1$ and $h_{-1}(x) = -1$ for all $x \in \mathcal{X}$, and a gating network $\mathbf{g} = (g_{+1}, g_{-1})$. Then, the following holds:

$$\begin{aligned}
R_S(\mathbf{g}, \mathbf{h}) &= \frac{1}{m} \sum_{j=1}^{m} \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x_j)} \ell_{0\text{-}1}(h_i(x_j), y_j) \\
&= \frac{1}{m} \sum_{j=1}^{m} \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x_j)} \mathbf{1}(h_i(x_j) \neq y_j) \\
&\geq \frac{1}{m} \sum_{j=1}^{m} \sum_{i \in \mathcal{I}} e^{-\epsilon} \max_{x' \in \mathcal{X}} g_i(x') \mathbf{1}(h_i(x_j) \neq y_j), \quad \text{with } \mathcal{I} = \{+1, -1\} \\
&= e^{-\epsilon} \frac{1}{m} \sum_{j=1}^{m} \max_{x' \in \mathcal{X}} g_{-y_j}(x').
\end{aligned}$$

Under the assumption that the classes are balanced, meaning that the (marginal) probability of a positive label is equal to the (marginal) probability of a negative label, we have the following:

$$\begin{aligned}
\lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} \max_{x' \in \mathcal{X}} g_{-y_j}(x') &= \frac{1}{2} \Big(\max_{x' \in \mathcal{X}} g_{-1}(x') + \max_{x' \in \mathcal{X}} g_{+1}(x') \Big) \\
&\geq \frac{1}{2} \max_{x' \in \mathcal{X}} \big(g_{-1}(x') + g_{+1}(x') \big) = \frac{1}{2}.
\end{aligned}$$

It follows that, in the limit $m \to \infty$, the risk bound of Theorem 4.4 for any $\mathbf{g}$ has a value of at least $e^{\epsilon}/2 \geq 1/2$. Consequently, the risk bound becomes large or even vacuous in this regime, highlighting the importance of having adaptive experts of finite complexity that can drive the empirical risk to zero when they are selected by the gating network.

5 Experiments and results

In what follows, we consider mixtures of n linear experts in binary classification tasks. Let $\mathcal{X} = \mathbb{R}^d$ for some positive integer d. Let S be a training set of m examples. Each expert, denoted by h_i, where i ranges from 1 to n, is characterized by a weight vector $\mathbf{w}_i$. Given an input $\mathbf{x} \in \mathcal{X}$, the output of the expert h_i is given by $h_i(\mathbf{x}) = \mathbf{w}_i \cdot \mathbf{x}$. We use the probit loss function $\ell = \Phi$, which can be seen as a smooth surrogate to the 0-1 loss function, when it is used with an argument of the form $\frac{y \mathbf{w}_i \cdot \mathbf{x}}{\|\mathbf{x}\|}$. In this case, $R(\mathbf{g}, Q)$ and $R_S(\mathbf{g}, Q)$ are given by:

$$R(\mathbf{g}, Q) = \mathop{\mathbb{E}}_{(\mathbf{x}, y) \sim \mathcal{D}} \mathop{\mathbb{E}}_{i \sim \mathbf{g}(\mathbf{x})} \Phi\Big(\frac{y \mathbf{w}_i \cdot \mathbf{x}}{\|\mathbf{x}\|} \Big)$$

and

$$R_S(\mathbf{g}, Q) = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} g_i(\mathbf{x}_j) \Phi\Big(\frac{y_j \mathbf{w}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_j\|} \Big), \tag{1}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-t^2/2}\, dt$ provides the probability that a standard normal random variable is greater than a given value x.

To illustrate the regularizing effect of the LDP condition, we carried out several experiments, on different datasets, by minimizing the empirical risk as defined in Equation 1. For all experiments, our models consist of mixtures of $n = 100$ linear experts and a gating network. The gating network is a

neural network having 2 hidden layers. It is parameterized by weights $\mathbf{W}_1 \in \mathbb{R}^{64\times d}$, where d is the dimension of input vectors, $\mathbf{W}_2 \in \mathbb{R}^{64\times 64}$, and $\mathbf{W}_3 \in \mathbb{R}^{n\times 64}$, and biases $\mathbf{b}_1 \in \mathbb{R}^{64}$, $\mathbf{b}_2 \in \mathbb{R}^{64}$ and $\mathbf{b}_3 \in \mathbb{R}^n$. Given an input $\mathbf{x} \in \mathbb{R}^d$, the output of the gating network $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}))$ is computed as follows: first, we compute $\mathbf{f}_0(\mathbf{x}) = \tanh(\mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)$. Then, when we want the ϵ-LDP condition to be satisfied, we ensure that the outputs are between $-\epsilon/4$ and $\epsilon/4$:

$$\mathbf{f}(\mathbf{x}) = \begin{cases} \frac{\epsilon \mathbf{W}_3 \mathbf{f}_0(\mathbf{x})}{4\|\mathbf{f}_0(\mathbf{x})\|\|\mathbf{W}_3\|_F} & \text{if the gating network must satisfy } \epsilon\text{-LDP} \\ \mathbf{W}_3\mathbf{f}_0(\mathbf{x}) & \text{otherwise.} \end{cases}$$

Note that $\tanh$ is the hyperbolic tangent activation function, ReLU the Rectified Linear Unit function, $\|\mathbf{W}_3\|_F$ the Frobenius norm of the matrix $\mathbf{W}_3$, and $\|\mathbf{f}_0(\mathbf{x})\|$ the euclidean norm of the vector $\mathbf{f}_0(\mathbf{x})$. Indeed, if we let $\mathbf{W}_3^i$ denote the i-th row of $\mathbf{W}_3$, then the i-th component of $\mathbf{W}_3\mathbf{f}_0(\mathbf{x})$ is

$$\mathbf{W}_3^i \cdot \mathbf{f}_0(\mathbf{x}) \leq \|\mathbf{W}_3^i\|\|\mathbf{f}_0(\mathbf{x})\| \leq \|\mathbf{W}_3\|_F\|\mathbf{f}_0(\mathbf{x})\|,$$

by the Cauchy-Schwarz inequality and the definition of the Frobenius norm. The reason we use the Frobenius norm instead of directly using $\|\mathbf{W}_3^i\|$ is to preserve the proportions between the components of $\mathbf{W}_3\mathbf{f}_0(\mathbf{x})$ when setting up ϵ-LDP.

The final output of the gating network is given by

$$g_i(\mathbf{x}) \;= \frac{\exp(f_i(\mathbf{x}) + (\mathbf{b}_3)_i)}{\sum_{k=1}^n \exp(f_k(\mathbf{x}) + (\mathbf{b}_3)_k)} \quad \text{for all} \quad i \in [n].$$

In our experiments, we ran the Stochastic Gradient Descent algorithm 10 times with a learning rate fixed to 0.1. In each experiment, we trained the model for 1000 epochs, except for the MNIST dataset, where the training duration was shortened to 300 epochs due to dataset size. We allocated approximately 75% of the data to the training set and the remaining 25% to the test set. At the outset of each experiment, the weights of our neural networks were reinitialized to ensure a fresh starting point. After each training run, we computed both the training and test loss values to evaluate the model's performance. We first ran the training without imposing any constraints on the gating network, except for the architecture. Then, we ran several experiments with a gating mechanism satisfying ϵ-LDP, with $\epsilon \in \{0.5, 2, 4, 5, 10\}$. A summary of the results is shown in Table 1. One can observe that regularization with ϵ-LDP improves results in practice, and this regularization is even more evident when the models employing a gating network not satisfying LDP overfit heavily, as in the Breast Cancer and Heart experiments. The regularization effect is slightly less pronounced on MNIST, where the overfitting is not as severe as with the previous datasets. We can also observe the importance of choosing the right hyperparameter ϵ. Indeed, if the value is too small, the output of the gating network becomes insufficiently dependent on the input $\mathbf{x}$. In this case, the experts have to do all the work, and the gating network does not allow them to specialize in well-defined subsets of the instance space. This makes our model closer to a weighted sum of linear classifiers and significantly reduces its performance. Conversely, if ϵ is overly large, our model tends towards a situation where the LDP condition does not hold, making it prone to overfitting.

Note that our experiments are executed on GPUs in order to parallelize computations and take advantage of the sparsity of our model, but they can also be performed without GPUs. The duration of experiments can range from a few minutes for small datasets such as Breast Cancer to around 3 hours for large datasets like MNIST.

6 Conclusion

In this work, we introduce a new way to regularize mixtures of experts. We provide both theoretical and algorithmic contributions in this regard. Our approach offers a significant advantage in that it allows us to harness the remarkable performances of neural networks by using them as gating networks, without being constrained by their architecture or their complexity from the theoretical point of view. By imposing LDP, we obtain nonvacuous bounds on the mixture of experts' risk. Our bounds can become significantly tighter than those presented in section 3.1 and those presented in

[2] If N denotes the number of runs, R_k denotes the training or test empirical risk during the k-th run, and $\bar{R}$ denotes the average, then standard deviation is given by $\sqrt{\frac{1}{N}\sum_{k=1}^N (R_k - \bar{R})^2}$.

Table 1: Experiment results for mixtures of 100 linear models applied to binary classification tasks: Ads, Breast Cancer [Zwitter and Soklic, 1988], Heart [Janosi et al., 1988] and MNIST [Deng, 2012]. The objective is to minimize the empirical risk as defined in Equation 1. We report the mean training loss (R_S) and mean test loss (R_T), averaged over ten runs, along with their associated standard deviations.[2]

			MoE with a gating network satisfying ϵ-LDP				
Dataset	Risk	No LDP	$\epsilon = 0.5$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 5$	$\epsilon = 10$
Ads	R_S	0.02425	0.13854	0.01829	0.05288	0.06459	0.02811
	$\pm$	0.00499	0.00261	0.00216	0.05543	0.05821	0.03648
	R_T	0.03822	0.13051	**0.03206**	0.06693	0.07757	0.04384
	$\pm$	0.00696	0.01138	**0.00564**	0.05276	0.05822	0.03501
Breast	R_S	0.00780	0.04520	0.01252	0.01062	0.01089	0.01207
Cancer	$\pm$	0.00347	0.00426	0.00182	0.00286	0.00193	0.00181
	R_T	0.03617	0.04930	0.03238	0.03297	0.02942	**0.02604**
	$\pm$	0.01505	0.01244	0.01349	0.01379	0.00948	**0.01277**
Heart	R_S	0.00001	0.03524	0.00015	0.00010	0.00009	0.00013
	$\pm$	0.00000	0.00487	0.00002	0.00001	0.00001	0.00006
	R_T	0.00029	0.03962	**0.00026**	0.00026	0.00032	0.00032
	$\pm$	0.00065	0.01013	**0.00014**	0.00033	0.00030	0.00032
MNIST	R_S	0.00525	0.00558	0.00529	0.00504	0.00536	0.00523
0 vs 8	$\pm$	0.00029	0.00059	0.00044	0.00031	0.00031	0.00032
	R_T	0.00844	0.00869	0.00815	0.00864	**0.00769**	0.00802
	$\pm$	0.00103	0.00109	0.00131	0.00165	**0.00144**	0.00067
MNIST	R_S	0.00287	0.00330	0.00289	0.00285	0.00298	0.00286
1 vs 7	$\pm$	0.00024	0.00033	0.00028	0.00025	0.00023	0.00013
	R_T	0.00501	0.00485	0.00501	0.00518	**0.00450**	0.00526
	$\pm$	0.00042	0.00101	0.00093	0.00098	**0.00101**	0.00066
MNIST	R_S	0.01419	0.01509	0.01388	0.01396	0.01440	0.01154
5 vs 6	$\pm$	0.00046	0.00057	0.00038	0.00051	0.00056	0.00336
	R_T	0.02195	0.02131	0.02206	0.02236	0.02072	**0.01852**
	$\pm$	0.00111	0.00160	0.00185	0.00269	0.00229	**0.00518**

Azran and Meir [2004], especially in cases where the empirical risk is close to zero and $\epsilon < \ln n$. However, as the empirical risk is multiplied by e^ϵ, the bounds can become loose when ϵ is large and the empirical risk is significant.

Even though the ϵ-LDP condition is easy to set up, a challenge arises in striking a balance between the parameter ϵ and the KL divergence or the Rademacher complexity of our experts. Our method introduces an extra hyperparameter ϵ to optimize but does not provide theoretical guidance on configuring it. This forces us to navigate a trade-off between the value of ϵ, which measures the extent to which the output of the gating network can depend on a given $x \in \mathcal{X}$, and the complexity of our experts, which reflects how well our model captures the data distribution. Finding the right balance requires empirical testing and careful consideration and can open up new avenues of study in the future.

References

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Nicolas Städler, Peter Bühlmann, and Sara van de Geer. l1-penalization for mixture regression models (with discussion). *TEST*, 19(2):209–285, 2010. doi: 10.1007/s11749-010-0197-z.

Abbas Khalili and Shili Lin. Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics*, 69, 04 2013. doi: 10.1111/biom.12020.

Seniha Yuksel, Joseph Wilson, and Paul Gader. Twenty years of mixture of experts, 08 2012.

M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pages 1339–1344 vol.2, 1993. doi: 10.1109/IJCNN.1993.716791.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022.

Arik Azran and Ron Meir. Data dependent risk bounds for hierarchical mixture of experts classifiers. In John Shawe-Taylor and Yoram Singer, editors, *Learning Theory*, pages 427–441, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-27819-1.

Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-35908-1.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately?, 2010.

David McAllester. A PAC-bayesian tutorial with a dropout bound, 2013.

Matjaz Zwitter and Milan Soklic. Breast Cancer. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C51P4M.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Michael D. Perlman. Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *Journal of Multivariate Analysis*, 4(1):52–65, 1974. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(74)90005-0. URL https://www.sciencedirect.com/science/article/pii/0047259X74900050.

Andreas Maurer. A note on the PAC bayesian theorem, 2004.

A Proofs and auxiliary results

Proof of theorem 2.2. Given $x \in \mathcal{X}$, let $Z(x) = \sum_{i=1}^n \exp(\beta f_i(x) + c_i)$, for convenience. For all $x, x' \in \mathcal{X}$ and all $i \in [n]$, we have that

$$\begin{aligned}
\frac{g_i(x)}{g_i(x')} &= \exp\big(\beta(f_i(x) - f_i(x'))\big) \frac{1}{Z(x)} \sum_{k=1}^n \exp(\beta f_k(x') + c_k) \\
&= \exp\big(\beta(f_i(x) - f_i(x'))\big) \frac{1}{Z(x)} \sum_{k=1}^n \exp(\beta f_k(x) + c_k) \exp\big(\beta(f_k(x') - f_k(x))\big) \\
&\leq \max_{i \in [n];\, x_1, x_2 \in \mathcal{X}} \exp\big(2\beta(f_i(x_1) - f_i(x_2))\big) \frac{1}{Z(x)} \sum_{k=1}^n \exp(\beta f_k(x) + c_k) \\
&\leq \exp(4\beta b). \qquad \square
\end{aligned}$$

Theorem A.1 (Jensen's inequality, proposition 1.1 in Perlman [1974])**.** *Let Ω be a probability space, let A be a convex subset of $\mathbb{R}^k$, let $X \colon \Omega \to A$ be an integrable vector-valued random variable, and let $\phi : A \to \mathbb{R}$ be a convex function. Then, $\mathbb{E}\, X \in A$, and $\phi(\mathbb{E}\, X) \leq \mathbb{E}\, \phi(X)$ (in particular, the right-hand side of this inequality exists, though it may be infinite).*

Theorem A.2 (Theorem 5 in Maurer [2004])**.** *Let $\delta \in (0, 1)$ and $m \geq 8$. Fix $i \in [n]$, and let P_i be a probability measure on $\mathcal{H}_i$ (chosen without seeing the training data). Then, with probability at least $1 - \delta$ over the draws of S, for all probability measures Q_i on $\mathcal{H}_i$, we have that*

$$\mathrm{kl}(R_S(Q_i) \,\|\, R(Q_i)) \leq \frac{1}{m}\Big(\mathrm{KL}(Q_i \,\|\, P_i) + \ln \frac{2\sqrt{m}}{\delta}\Big).$$

Proof of theorem 3.4. As remarked earlier, the function $(u, v) \to \mathrm{kl}(u \,\|\, v) : [0, 1]^2 \to \mathbb{R}$ does not exactly satisfy the hypotheses of lemma 3.1, but it is convex. Moreover, on $\{\, (u, v) \in [0, 1]^2 \,|\, u \leq v \,\}$, it is decreasing in its first argument and increasing in its second argument. Also note that, assuming that $R(\mathbf{g}, Q) \geq e^{2\epsilon} R_S(\mathbf{g}, Q)$, then we also have the following inequalities:

$$0 \leq \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R_S(Q_i) \leq e^{\epsilon} R_S(\mathbf{g}, Q) \leq e^{-\epsilon} R(\mathbf{g}, Q) \leq \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R(Q_i) \leq 1.$$

It follows that

$$\begin{aligned}
\mathrm{kl}(e^{\epsilon} R_S(\mathbf{g}, Q) \,\|\, e^{-\epsilon} R(\mathbf{g}, Q)) &\leq \mathrm{kl}\Big(\mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R_S(Q_i) \,\Big\|\, e^{-\epsilon} R(\mathbf{g}, Q)\Big) \\
&\leq \mathrm{kl}\Big(\mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R_S(Q_i) \,\Big\|\, \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R(Q_i)\Big),
\end{aligned}$$

and therefore

$$\mathrm{kl}(e^{\epsilon} R_S(\mathbf{g}, Q) \,\|\, e^{-\epsilon} R(\mathbf{g}, Q)) \leq \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} \mathrm{kl}\big(R_S(Q_i) \,\|\, R(Q_i)\big)$$

by Jensen's inequality. Now, by theorem A.2, for a fixed i, with probability at least $1 - \delta/n$, we have that

$$\mathrm{kl}(R_S(Q_i) \,\|\, R(Q_i)) \leq \frac{1}{m}\Big(\mathrm{KL}(Q_i \,\|\, P_i) + \ln \frac{2n\sqrt{m}}{\delta}\Big).$$

We can make the above inequality hold for all $i \in [n]$ simultaneously with the union bound. Then, with probability at least $1 - \delta$, for all $(\mathbf{g}, Q)$, given that $R(\mathbf{g}, Q) \geq e^{2\epsilon} R_S(\mathbf{g}, Q)$, we have that

$$\mathrm{kl}(e^{\epsilon} R_S(\mathbf{g}, Q) \,\|\, e^{-\epsilon} R(\mathbf{g}, Q)) \leq \frac{1}{m}\Big(\mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} \mathrm{KL}(Q_i \,\|\, P_i) + \ln \frac{2n\sqrt{m}}{\delta}\Big). \qquad \square$$

Proof of Lemma 4.1. Since the gating function satisfies ϵ-LDP, we have that $e^{-\epsilon} g_i(x') \leq g_i(x) \leq e^{\epsilon} g_i(x')$ for all $x, x' \in \mathcal{X}$ and all $i \in [n]$. It follows that $e^{\epsilon} R_S(\mathbf{g}, \mathbf{h}) \geq \mathbb{E}_{i \sim \mathbf{g}(x')} R_S(h_i)$ and $e^{-\epsilon} R(\mathbf{g}, \mathbf{h}) \leq \mathbb{E}_{i \sim \mathbf{g}(x')} R(h_i)$. Given that Δ is decreasing in its first argument and increasing in its second argument, we find that

$$\Delta\big(e^{\epsilon} R_S(\mathbf{g}, \mathbf{h}), e^{-\epsilon} R(\mathbf{g}, \mathbf{h})\big) \leq \Delta\Big(\mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R_S(h_i), \mathop{\mathbb{E}}_{i \sim \mathbf{g}(x')} R(h_i)\Big)$$

Since Δ is a convex function, we can apply Jensen's inequality to the expression on the right-hand side, yielding the desired result. $\square$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims are supported by theorems, which we prove, and by experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results are proved, either in the main paper or in the appendix. Moreover, all assumptions necessary for our theorems to hold are mentioned in their statement.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All datasets used are freely available, and we describe the architecture of the model and the hyperparameters of our experiments in detail. We also provide the code used to run our experiments as supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in our experiments are freely available. The code used to run our experiments is provided as supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In experiments, we had multiple runs and reported averages along with standard deviations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the type of compute workers (GPU) and time of execution in section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: In this work, we provide theoretical results that do not have any direct negative impact on society and do not pose any obvious ethical problem. We have ensured to cite our sources and adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our paper has a theoretical orientation and no clear negative impacts.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our models are trained on freely available datasets as a way to prove the efficacy of our approach, but they pose no risk as such.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use openly accessible datasets and ensure to cite them properly when necessary.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is provided as supplemental material and the details are given in the README file.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve research with human subjects nor crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.