PAC-Bayesian Bounds on Rate-Efficient Classifiers

Alhabib Abbas [1 2] **Yiannis Andreopoulos** [2]

Abstract

We derive analytic bounds on the noise invariance of majority vote classifiers operating on compressed inputs. Specifically, starting from recent bounds on the true risk of majority vote classifiers, we extend the applicability of PAC-Bayesian theory to quantify the resilience of majority votes to input noise stemming from compression. The derived bounds are intuitive in binary classification settings, where they can be measured as expressions of voter differentials and voter pair agreement. By combining measures of input distortion with analytic guarantees on noise invariance, we prescribe rate-efficient machines to compress inputs without affecting subsequent classification. Our validation shows how bounding noise invariance can inform the compression stage for any majority vote classifier such that worst-case implications of bad input reconstructions are known, and inputs can be compressed to the minimum amount of information needed prior to inference.

1. Introduction

To learn concepts inherently contained in data, recent breakthroughs in deep learning, adversarial robustness, and bounded bayesian inference (Germain et al., 2015; Letarte et al., 2019; Vidot et al., 2021) typically assume models that can observe whole uncompressed volumes of d-dimensional inputs $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ in order to map $\boldsymbol{x}$ to target concepts $y \in \mathcal{Y}$. However, in practice, systems with distributed computing assets (Pradhan et al., 2002; Xiao et al., 2006) employ lossy compression techniques to reduce the load of communication between inference machines and sensors that collect data at test time. In such contexts, whole volumes of $\boldsymbol{x}$ are inaccessible to machines tasked with inference, and only noisy reconstructions of $\boldsymbol{x}$ are available to infer concepts y.

Adapting recent proposals to data scarcity at test time entails the design of models that infer concepts from *compressed* samples of input $\boldsymbol{x}_c = C(\boldsymbol{x})$, where $C : \mathcal{X} \rightarrow \mathcal{E} \rightarrow \mathcal{X}$ encodes and decodes $\boldsymbol{x}$ *with information loss*, $\mathcal{E} \subseteq \mathbb{R}^{d_c}$ is the set describing compressed codes of $\boldsymbol{x}$, and $d_c \leq d$. From a rate-distortion theory perspective (Gray & Neuhoff, 1998), code lengths d_c quantify the rate of bitstreams ingested by arbitrary models $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$, and since $d_c^{-1} \propto |\boldsymbol{x}_c - \boldsymbol{x}|$, rate-efficient classifiers define the family of models that satisfy $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}_c)$ for lower values of d_c. Generally, models that return outputs sensitive to small perturbations on $\boldsymbol{x}$ can be deemed as *greedy* models, due to their requirement of long code lengths d_c to reduce $|\boldsymbol{x}_c - \boldsymbol{x}|$. On the other hand, models that are resilient to input distortion are *rate-efficient*, since they require shorter code lengths d_c to correctly infer concepts y. Thus, rate-efficiency defines a sub-class of models that are resilient to noise afflicted on $\boldsymbol{x}$, and this sets the context of our approach to solve for rate-efficient machines.

We hereon define *noise invariance* as the probability of output change with respect to noisy input perturbations, and in the case of binary classifiers where $y \in \{-1, 1\}$, noise invariance specifically refers to the percentile of outputs *flipped* due to input compression noise. Bounding the noise invariance of classifiers $\mathcal{M}$ can allow for more aggressive compression, since it gives guarantees on the implications of reconstruction loss $|\boldsymbol{x}_c - \boldsymbol{x}|$ on predicted outputs. Hence, wherever noise invariance bounds are defined over probabilistic measures on input noise, empirical measures of $|\boldsymbol{x}_c - \boldsymbol{x}|$ can be inspected to compress inputs up to allowable limits on code lengths d_c without distorting predictions $\mathcal{M}(\boldsymbol{x}_c)$. The latter motivates our proposal, and we summarise our contributions below:

1. We introduce the notion of rate-efficient classifiers to PAC-Bayesian theory, and derive relevant expressions of noise-invariance to define such classifiers.

2. We derive unsupervised general bounds on noise-invariance, and specialize them to binary majority vote classifiers comprising linear kernels.

3. We combine noise-invariance bounds with measures of reconstruction loss to realize rate-efficient machines that bound errors resulting from lossy compression.

[1] Meme Research Ltd., London, UK [2] Dept. of Electronic and Electrical Eng., University College London, London, UK. Correspondence to: A. Abbas <alhabib.abbas.13@ucl.ac.uk>, Y. Andreopoulos <i.andreopoulos@ucl.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Our bounds on noise invariance incorporate two key aspects detrimental to the resilience of majority votes (James, 1998; Lacasse et al., 2006) to input distortion: (i) the agreement across voter pairs on predicted continuous outputs, and (ii) differentials of voter outputs with respect to shifting inputs. Both terms are intuitive, where the first emphasizes how the consensus across voters should exist away from decision boundaries, and the second relates the importance of smooth voters to noise-invariance. Generally, the noise invariance bounds we derive are applicable wherever inputs are reliably modelled as gaussian processes.

In Section 2 we give a concise introduction to relevant definitions in *existing* work on PAC-Bayesian theory. In Section 3 we formalise rate-efficient classifiers and state the problem we solve. In Section 4 we prime definitions on individual voters to qualify Section 5, where we derive noise invariance measures and bounds for majority votes. In Section 6 we discuss recent work on input perturbation bounds. In Section 7 we report results on all relevant theorems to validate our proposal detailed in Sections (3, 4, 5), and Section 8 concludes the paper.

2. An Introduction to PAC-Bayesian Theory

Let $(\boldsymbol{x}, y)$ denote input-output pairs drawn from an unknown domain D over the support $\mathcal{X} \times \mathcal{Y}$, where inputs $\boldsymbol{x}$ are normalised such that $\boldsymbol{x} \in \mathcal{X} \subseteq [-1,1]^d$, and $y \in \mathcal{Y}$ defines targets in a binary classification setting when $\mathcal{Y} = \{-1, 1\}$. Given a set of m examples $\mathcal{S} \sim D^m$ and a hypotheses set $\mathcal{H} = \{h_i(\boldsymbol{x})\}$ where $h_i : \mathcal{X} \to [-1,1]$, we define the true risk $R(h_i)$ and empirical risk $R_{\mathcal{S}}(h_i)$ as the expectation of errors made by classifiers $\text{sgn}(h_i)$ on D and $\mathcal{S}$, respectively.

Definition 1 For any source domain D, and for any set of examples $\mathcal{S}$, the true risk $R(h_i)$ and empirical risk $R_{\mathcal{S}}(h_i)$ measure the expectation:

$$R(h_i) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim D} \mathbf{1}_{\mathbb{R}^-}\Big(y \cdot h_i(\boldsymbol{x})\Big)$$

$$R_{\mathcal{S}}(h_i) := \frac{1}{m} \sum_{(\boldsymbol{x},y)\in\mathcal{S}} \mathbf{1}_{\mathbb{R}^-}\Big(y \cdot h_i(\boldsymbol{x})\Big)$$

where $\mathbf{1}_{\mathbb{R}^-}(c)$ defines the indicator function such that $\mathbf{1}_{\mathbb{R}^-}(c) = 1$ if $c \in \mathbb{R}^-$, and $\mathbf{1}_{\mathbb{R}^-}(c) = 0$ otherwise.

Specifically, the expectations of Definition 1 measure the mean zero-one loss (Bartlett et al., 2006). For any density function $Q(h_i)$ defined over $\mathcal{H}$, PAC-Bayesian theory (Catoni, 2007; Langford & Shawe-Taylor, 2002; McAllester, 2003) bounds the true risk of majority vote (Bayes) classifiers $B_Q(\boldsymbol{x})$. The following formally states the output of kernelised majority votes $B_Q(\boldsymbol{x})$ weighted by $Q(h_i)$ when $\mathcal{H}$ defines a set of m voters $h_i(\boldsymbol{x})$ as functions of kernels $(\boldsymbol{x}'_i, y_i) \in \mathcal{S}$ (Germain et al., 2015).

Definition 2 For any input $\boldsymbol{x} \in \mathcal{X}$, for any posterior Q defined over training examples $(\boldsymbol{x}'_i, y_i) \in \mathcal{S}$, and for any set of kernelised voters $h_i(\boldsymbol{x}) = y_i K(\boldsymbol{x}, \boldsymbol{x}'_i)$:

$$B_Q(\boldsymbol{x}) = \text{sgn}\Big(\mathop{\mathbb{E}}_{(\boldsymbol{x}'_i,y_i)\sim Q} y_i K(\boldsymbol{x}, \boldsymbol{x}'_i)\Big)$$

Interestingly, for arbitrary hypotheses $h_i(\boldsymbol{x})$, many binary classification techniques implicitly define majority vote classifiers. For example, kernelised support vector machines (Gholami & Fakhari, 2017), boosting methods (Sagi & Rokach, 2018), and mixtures of experts (Shazeer et al., 2017) can all be construed as special cases of Definition 2. Learning accurate majority vote classifiers entails learning posteriors $Q(h_i|\mathcal{S})$ over $\mathcal{H}$ such that the true risk $R(B_Q)$ of the Q-weighted majority vote $B_Q(\boldsymbol{x})$ is minimized.

Bounds on the true risk of $B_Q(\boldsymbol{x})$ are commonly derived from intermediary bounds on the risk of Gibbs classifiers $G_Q(\boldsymbol{x})$ (Lacasse et al., 2006) that randomly sample hypotheses $h_i \sim Q$ to classify $\boldsymbol{x}$. Indeed, the output of Gibbs classifiers can be different even if the same input is passed twice, and the following defines risk measures for $G_Q(\boldsymbol{x})$.

Definition 3 For any source domain D, and for any posterior Q on $\mathcal{H}$, the true Gibbs risk $R(G_Q)$ and empirical Gibbs risk $R_{\mathcal{S}}(G_Q)$ measure the expectation of drawing an erroneous classifier $\text{sgn}(h_i)$ from Q:

$$R(G_Q) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim D} \mathop{\mathbb{E}}_{h_i\sim Q} \mathbf{1}_{\mathbb{R}^-}\Big(y \cdot h_i(\boldsymbol{x})\Big)$$

$$R_{\mathcal{S}}(G_Q) := \frac{1}{m} \sum_{(\boldsymbol{x},y)\in\mathcal{S}} \mathop{\mathbb{E}}_{h_i\sim Q} \mathbf{1}_{\mathbb{R}^-}\Big(y \cdot h_i(\boldsymbol{x})\Big)$$

From Definition 3, a trivial bound on the majority vote classifier can be directly derived as $R(B_Q) \leq 2R(G_Q)$ (Lacasse et al., 2006; Langford & Shawe-Taylor, 2002). This is because, whenever B_Q misclassifies $\boldsymbol{x}$, the probability of drawing a classifier $h_i \sim Q$ that misclassifies $\boldsymbol{x}$ is at at least 0.5. Hence, it follows that the true risk of $B_Q(\boldsymbol{x})$ is at most twice the true risk of $G_Q(\boldsymbol{x})$ and $R(B_Q) \leq 2R(G_Q)$.

2.1. PAC-Bayesian Bounds on Measures of Risk

The canonical majority vote bounds of (Catoni, 2007; Germain et al., 2015; Langford & Shawe-Taylor, 2002; McAllester, 2003) on $R(G_Q)$ are typically parameterized by: (i) the number of training examples m constituting a training set $\mathcal{S}$ used to learn the posterior $Q(h_i|\mathcal{S})$, (ii) a prior distribution $P(h_i)$, and (iii) an arbitrary $\delta \in (0, 1]$ that specifies the probability of the bound. The following theorem expresses a common starting point for PAC-Bayesian bounds on $R(G_Q)$, first introduced by McAllester *et al.* in (McAllester, 1999).

Theorem 1 For any source distribution D, for any prior P on the hypothesis set $\mathcal{H}$, for any convex distance function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$, and for any posterior Q learned by observing m examples $\boldsymbol{x} \in \mathcal{S}$:

$$\underset{\mathcal{S}\sim D}{Pr}\Bigg(\mathcal{D}(R_{\mathcal{S}}(G_Q), R(G_Q)) \leq \frac{1}{m}\Bigg[KL(Q||P) + \ln\Big(\frac{1}{\delta}\underset{\mathcal{S}\sim D}{\mathbb{E}}\,\underset{h_i\sim P}{\mathbb{E}}\, e^{m\mathcal{D}(R_{\mathcal{S}}(h_i),R(h_i))}\Big)\Bigg]\Bigg) \geq 1-\delta$$

Proof: Applying Markov's inequality on the positive expression $\mathbb{E}_{h\sim P}e^{m\mathcal{D}(R_{\mathcal{S}}(h_i),R(h_i))}$, and converting $\mathbb{E}_{h_i\sim P}$ to $\mathbb{E}_{h_i\sim Q}$ yields:

$$\underset{\mathcal{S}\sim D}{Pr}\left(\underset{h_i\sim Q}{\mathbb{E}}\frac{P(h_i)}{Q(h_i)}e^{m\mathcal{D}(R_{\mathcal{S}}(h_i),R(h_i))} \leq \frac{1}{\delta}\underset{\mathcal{S}\sim D}{\mathbb{E}}\,\underset{h_i\sim P}{\mathbb{E}}\, e^{m\mathcal{D}(R_{\mathcal{S}}(h_i),R(h_i))}\right) \geq 1-\delta$$

The result follows after taking $\ln(\cdot)$ on each side and applying Jensen's inequality by exploiting the concavity and convexity of $\ln(\cdot)$ and $\mathcal{D}$, respectively. For a step-by-step account of this proof, see (Germain et al., 2009; 2015).

Relevant to our proposal is the specialization of Theorem 1 to convex divergence measures $\mathcal{D}$ on the probability of binary events. Specifically, by setting $\mathcal{D}(R_{\mathcal{S}}(G_Q), R(G_Q))$ as the Kullback-Leibler divergence of *binomial* distributions $\mathrm{kl}(R_{\mathcal{S}}(G_Q)||R(G_Q))$, the following corollary simplifies Theorem 1.

Corollary 1 For any source domain D, for any prior P on the hypothesis set $\mathcal{H}$, for the binomial distance function $\mathrm{kl}(\cdot) : [0,1] \times [0,1] \to \mathbb{R}$, and for any posterior Q learned by observing m examples $\boldsymbol{x} \in \mathcal{S}$:

$$\underset{\mathcal{S}\sim D}{Pr}\left(\mathrm{kl}(R_{\mathcal{S}}(G_Q)||R(G_Q)) \leq \frac{1}{m}\left[KL(Q||P) + \ln\frac{\xi(m)}{\delta}\right]\right) \leq 1-\delta$$

when $\xi(m) := \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$

Proof: Since the Kullback-Leibler divergence $\mathrm{kl}(q||p)$ between two binomial distributions parameterised by p and q is:

$$\mathrm{kl}(q||p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$$

by interpreting $R_{\mathcal{S}}(h_i)$ and $R(h_i)$ as parameters of distinct binomial variates, $\xi(m)$ emerges after breaking the exponential term in $\mathbb{E}_{\mathcal{S}\sim D}\,\mathbb{E}_{h_i\sim P}\, e^{m\mathcal{D}(R_{\mathcal{S}}(h_i),R(h_i))}$ of Theorem 1, where the expectation becomes:

$$\underset{\mathcal{S}\sim D}{\mathbb{E}}\,\underset{h_i\sim P}{\mathbb{E}}\Big(\frac{R_{\mathcal{S}}(h_i)}{R(h_i)}\Big)^{mR_{\mathcal{S}}(h_i)}\Big(\frac{1-R_{\mathcal{S}}(h_i)}{1-R(h_i)}\Big)^{m(1-R_{\mathcal{S}}(h_i))}$$
$$= \underset{h_i\sim P}{\mathbb{E}}\sum_{k=0}^{m}\underset{\mathcal{S}\sim D}{Pr}\Big(R_{\mathcal{S}}(h_i)=\frac{k}{m}\Big)\Big(\frac{\frac{k}{m}}{R(h_i)}\Big)^{mR_{\mathcal{S}}(h_i)}\Big(\frac{1-\frac{k}{m}}{1-R(h_i)}\Big)^{m(1-R_{\mathcal{S}}(h_i))}$$

Since $\mathbb{E}_{\mathcal{S}\sim D}\, R_{\mathcal{S}}(h_i) = R(h_i)$, the last expectation is simplified to $\xi(m)$. Following the proof of Theorem 1 then yields the result of the corollary. For more details, see (Germain et al., 2009; 2015).

Importantly, even if Corollary 1 is typically defined to bound the empirical risk $R_{\mathcal{S}}(G_Q)$ (Germain et al., 2009; 2015), parameters of $\mathcal{D}(\cdot)$ can be set to bound other binary events defined as functions of $Q(h_i|\mathcal{S})$ and $P(h_i)$. Thus, in order to adapt Corollary 1 to noise invariance measures for majority vote classifiers $B_Q(\boldsymbol{x})$, in Sections 4 and 5 we derive exact expressions of noise invariance as expectations over $Q(h_i|\mathcal{S})$ and $P(h_i)$.

3. Formalising Rate-Efficient Machines

We propose to realize rate-efficient machines by quantifying the resilience of classifiers to input degradation via PAC-Bayesian bounds on noise invariance. Let $\boldsymbol{x}_c = C(\boldsymbol{x}, \theta_c)$ denote any lossy compression of $\boldsymbol{x}$ parameterised by θ_c. To give a probabilistic handle $\sigma_c^2 \in \mathbb{R}$ on reconstruction noise, and letting $I \in \mathbb{R}^{d\times d}$ denote the identity matrix, we fit $\mathcal{N}(\mathbf{0}, \sigma_c^2 I)$ on $\boldsymbol{x}_c - \boldsymbol{x}$ such that σ_c^2 quantifies the extent of distortion inflicted on $\boldsymbol{x}$ by the compressor (see Figure 1). For high-rate compression regiments, gaussians accurately model reconstruction loss $|\boldsymbol{x}_c - \boldsymbol{x}|$ in statistical representations of rate-distortion theory (Gray & Neuhoff, 1998). Capitalising on this, we define η^D as the probability of output change with respect to input perturbation for any model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$.

Definition 4 For any source domain D, for any classification model $\mathcal{M}$, and for any noise vector $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 I)$ modelling perturbations on $\boldsymbol{x}$, noise invariance η^D quantifies the probability of output change due to $\boldsymbol{n}$:

$$\eta^D = \underset{\boldsymbol{x}\sim D}{\mathbb{E}}\,\underset{\boldsymbol{n}\sim\mathcal{N}}{Pr}\Big(\mathcal{M}(\boldsymbol{x}) \neq \mathcal{M}(\boldsymbol{x}+\boldsymbol{n})\Big)$$

In distributed computing settings, inference (classifier) machines $\mathcal{M}$ cannot observe $\boldsymbol{n}$ directly. However, prior measures of σ_c^2 can be combined with knowledge of the functional form of $\mathcal{M}$ to infer probabilities of misclassification when $\mathcal{M}$ observes $\boldsymbol{x}_c = \boldsymbol{x}+\boldsymbol{n}$. We therefore endeavor to design resilient models that give PAC bounds $\mathcal{B}_\eta$ on the noise invariance η^D of $\mathcal{M}$ such that $\eta^D \leq \mathcal{B}_\eta$ with probability δ_η, whenever empirical measures are given for σ_c^2 at test time. Our approach leverages existing PAC-Bayesian theory (Germain et al., 2009; 2015; Langford & Shawe-Taylor, 2002; McAllester, 2003) qualified in Section 2 to derive bounds on η^D for majority vote classifiers when $\mathcal{M}(\boldsymbol{x}) = B_Q(\boldsymbol{x})$. Note that we use the subscripted notation δ_η to distinguish it from δ, which commonly denotes the probability of PAC bounds $\mathcal{B}_R$ on the true risk of majority votes in existing literature on PAC-Bayesian inference (Germain et al., 2009; 2015).

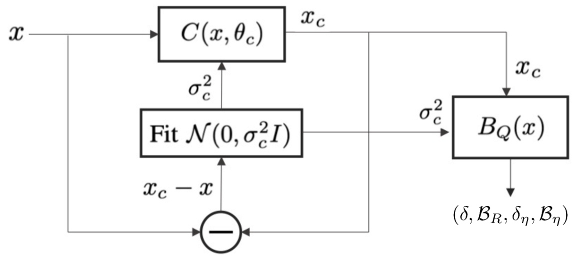
x
$C(x,\theta_c)$
x_c
x_c
σ_c^2
σ_c^2
Fit $\mathcal{N}(0,\sigma_c^2 I)$
$B_Q(x)$
$x_c - x$
$(\delta, \mathcal{B}_R, \delta_\eta, \mathcal{B}_\eta)$

Figure 1. Observation model of η-enabled rate-efficient inference. By combining analytical guarantees with measures of compression noise, we realize rate-efficient machines that bound noise invariance. Note that $\mathcal{M}(x) = B_Q(x)$ returns PAC bounds on risk via $(\delta, \mathcal{B}_R)$ and on compression noise invariance via $(\delta_\eta, \mathcal{B}_\eta)$.

On the relevance of η^D to rate-efficiency: Bounds $\mathcal{B}_\eta$ on noise invariance η^D facilitate for rate-efficient machines, where compressors $C(\boldsymbol{x}, \theta_c)$ can compress inputs to limits that do not distort classification outputs beyond $\mathcal{B}_\eta$ with probabilities $\geq 1 - \delta_\eta$. That is, by estimating the variance of density functions $\mathcal{N}(\boldsymbol{0}, \sigma_c^2 I)$ fitted on empirical measures of $\boldsymbol{x}_c - \boldsymbol{x}$ at test time, compression parameters θ_c can be tuned until σ_c^2 reaches a target value σ_t^2 that guarantees a bound $\mathcal{B}_\eta$ on the noise invariance of $B_Q(\boldsymbol{x})$. For example, Figure 1 illustrates how measures of compression noise can be coupled with analytic PAC guarantees on noise invariance, such that compressors $C(\boldsymbol{x}, \theta_c)$ can refine $\boldsymbol{x}_c$ until realizing the highest σ_c^2 that is suitably bounded by $\mathcal{B}_\eta$ and δ_η. Importantly, compressors in Figure 1 can include any compression model with tunable parameters θ_c.

4. Quantifying Voter Noise Invariance

We are interested in resilient models that give consistent outputs despite noise vectors $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma}_c^2 I)$ inflicted on compressed inputs $\boldsymbol{x}_c = \boldsymbol{x} + \boldsymbol{n}$. To quantify the noise invariance of majority vote classifiers to compression noise, we begin by defining a measure of resilience $r_Q(\boldsymbol{x})$ on individual voters $h_i(\boldsymbol{x})$ of majority vote classifiers $B_Q(\boldsymbol{x})$ as the expectation of voter disagreement due to perturbations $\boldsymbol{n}$ inflicted on compressed inputs.

Definition 5 For any source domain D, for any posterior Q defined over continuous voters $h_i : \mathcal{X} \to [-1, 1]$, and for any $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma_c^2 I)$, we define continuous voter resilience $r_Q(\boldsymbol{x})$ as:

$$r_Q(\boldsymbol{x}) \coloneqq \mathop{\mathbb{E}}_{h_i \sim Q} \mathop{\mathbb{E}}_{\boldsymbol{n} \sim \mathcal{N}} h_i(\boldsymbol{x}) \cdot h_i(\boldsymbol{x} + \boldsymbol{n})$$

Notably, the inner product $h_i(\boldsymbol{x}) \cdot h_i(\boldsymbol{x} + \boldsymbol{n})$ in Definition 5 returns values $\in \{-1, 1\}$ when both $h_i(\boldsymbol{x})$ and $h_i(\boldsymbol{x} + \boldsymbol{n})$ yield exact binary values $\in \{-1, 1\}$. However, $h_i(\boldsymbol{x}) \cdot h_i(\boldsymbol{x} + \boldsymbol{n})$ in Definition 5 also gives a rich measure on the resilience of individual voters by returning intermediate values $\in [-1, 1]$ whenever $h_i(\boldsymbol{x}), h_i(\boldsymbol{x} + \boldsymbol{n}) \in [-1, 1]$, to account for the certainty of $h_i(\boldsymbol{x})$ and $h_i(\boldsymbol{x} + \boldsymbol{n})$ in assigning positive or negative classes. To understand the key characteristics that determine the resilience of individual voters to input perturbation, the next lemma decomposes $r_Q(\boldsymbol{x})$ of Definition 5.

Lemma 1 For any input $\boldsymbol{x} \in \mathcal{X}$, for any posterior Q, and for any noise vector $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma_c^2 I)$, the resilience measure $r_Q(\boldsymbol{x})$ can be decomposed to:

$$r_Q(\boldsymbol{x}) = \mathop{\mathbb{E}}_{h_i \sim Q} \mathop{\mathbb{E}}_{\boldsymbol{n} \sim \mathcal{N}} h_i(\boldsymbol{x})^2 + h_i(\boldsymbol{x}) \Big(h_i(\boldsymbol{x} + \boldsymbol{n}) - h_i(\boldsymbol{x}) \Big)$$

Proof: We rearrange Definition 5 and use expectation properties to segregate $r_Q(\boldsymbol{x})$ into the two terms of the lemma:

$$\begin{aligned} r_Q(\boldsymbol{x}) &\coloneqq \mathop{\mathbb{E}}_{h_i \sim Q} \mathop{\mathbb{E}}_{\boldsymbol{n} \sim \mathcal{N}} h_i(\boldsymbol{x}) h_i(\boldsymbol{x} + \boldsymbol{n}) \\ &= \mathop{\mathbb{E}}_{h_i \sim Q} \mathop{\mathbb{E}}_{\boldsymbol{n} \sim \mathcal{N}} h_i(\boldsymbol{x}) \Big[h_i(\boldsymbol{x}) + \Big(h_i(\boldsymbol{x} + n) - h_i(\boldsymbol{x}) \Big) \Big] \end{aligned}$$

and distributing $h_i(\boldsymbol{x})$ into the the sum yields the lemma.

The expression of Lemma 1 highlights the dependence of $r_Q(\boldsymbol{x})$ on two measures: (i) the confidence of individual voters $h_i(\boldsymbol{x})^2 \in [0, 1]$ where $h_i(\boldsymbol{x})^2$ scales quadratically against votes $h_i(\boldsymbol{x})$, and (ii) the differential of voter outputs $h_i(\boldsymbol{x} + \boldsymbol{n}) - h_i(\boldsymbol{x})$, which measures the amount of change on $h_i(\boldsymbol{x})$ as a result of perturbing inputs $\boldsymbol{x}$ with a noise vector $\boldsymbol{n}$. It is also interesting to point out that, from Lemma 1, the term $h_i(\boldsymbol{x} + \boldsymbol{n}) - h_i(\boldsymbol{x})$ correlates with the differentials $\frac{\delta h_i(\boldsymbol{x})}{\delta \boldsymbol{x}}$ specifically in local regions softly demarked by gaussian hyperspheres defined over the multivariate $\mathcal{N}(0, \sigma_c^2 I)$.

To close on properties relating to individual voters, the following defines $a_Q(\boldsymbol{x})$ as the expectation of agreement between continuous voter pairs.

Definition 6 For any input $\boldsymbol{x} \in \mathcal{X}$, for any posterior Q, $a_Q(\boldsymbol{x})$ quantifies the agreement between voters as:

$$a_Q(\boldsymbol{x}) \coloneqq \mathop{\mathbb{E}}_{h_i \sim Q} \mathop{\mathbb{E}}_{h_j \sim Q} h_i(\boldsymbol{x}) \cdot h_j(\boldsymbol{x})$$

Agreement as measured by Definition 6 will become helpful in simplifying expressions relating to the noise invariance of bayesian classifiers $B_Q(\boldsymbol{x})$. It is also worth noting here that $a_Q(\boldsymbol{x})$ correlates inversely with the concept of disagreement in the $\mathcal{C}$-bound of (Germain et al., 2015), which emerges often in derivations of PAC-Bayesian bounds on risk. Specifically, the $\mathcal{C}$-bound suggests that majority votes perform a trade-off between lower Gibbs risk $R(G_Q)$ and higher disagreements $a_Q(\boldsymbol{x})^{-1}$ to achieve better majority vote risks $R(B_Q)$. Bounds defined as functions of voter agreement benefit from the unsupervised nature of $a_Q(\boldsymbol{x})$, since it can be measured with higher precision without needing labels y, and our bound on noise invariance that we derive in Section 5 inherits this unsupervised aspect of $a_Q(\boldsymbol{x})$.

5. Noise Invariance of the Majority Vote

From Definition 4, we define the noise invariance η_Q^D of majority vote classifiers as the probability of disagreement $\mathcal{M}(\boldsymbol{x}) \neq \mathcal{M}(\boldsymbol{x}+\boldsymbol{n})$ when $\mathcal{M}(\boldsymbol{x}) = B_Q(\boldsymbol{x})$.

Definition 7 For any majority vote classifier B_Q, for any posterior Q, and for any noise vector $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 I)$, the resilience η_Q^D of $B_Q(\boldsymbol{x})$ to perturbations on $\boldsymbol{x}$ is:

$$\eta_Q^D \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{x}\sim D} \mathop{Pr}_{\boldsymbol{n}\sim\mathcal{N}} \left(\mathop{\mathbb{E}}_{h_i\sim Q} h_i(\boldsymbol{x}+\boldsymbol{n}) \cdot \mathop{\mathbb{E}}_{h_j\sim Q} h_j(\boldsymbol{x}) \leq 0 \right)$$

Lemma 2 For any posterior Q, for any noise vector $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 I)$, and by extending $h_i(\boldsymbol{x}+\boldsymbol{n})$, the noise invariance measure η_Q^D can be decomposed to:

$$\eta_Q^D = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim D} \mathop{Pr}_{\boldsymbol{n}\sim\mathcal{N}} \left(\mathop{\mathbb{E}}_{h_i\sim Q} \mathop{\mathbb{E}}_{h_j\sim Q} h_i(\boldsymbol{x}) \cdot h_j(\boldsymbol{x}) \leq \right.$$
$$\left. \mathop{\mathbb{E}}_{h_i\sim Q} \mathop{\mathbb{E}}_{h_j\sim Q} h_j(\boldsymbol{x}) \cdot \Big(h_i(\boldsymbol{x}) - h_i(\boldsymbol{x}+\boldsymbol{n}) \Big) \right)$$

Proof: Exploiting the fact that $h_i(\boldsymbol{x}+\boldsymbol{n}) = h_i(\boldsymbol{x}) + \big(h_i(\boldsymbol{x}+\boldsymbol{n}) - h_i(\boldsymbol{x})\big)$, we rearrange Definition 7 and use expectation properties to decompose the inner multiplication:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{h_i\sim Q} h_i(\boldsymbol{x}+\boldsymbol{n}) \cdot \mathop{\mathbb{E}}_{h_j\sim Q} h_j(\boldsymbol{x}) \\
&\quad = \mathop{\mathbb{E}}_{h_i\sim Q} \left(\mathop{\mathbb{E}}_{h_j\sim Q} h_j(\boldsymbol{x}) \cdot h_i(\boldsymbol{x}+\boldsymbol{n}) \right) \\
&\quad = \mathop{\mathbb{E}}_{h_i\sim Q} \mathop{\mathbb{E}}_{h_j\sim Q} h_i(\boldsymbol{x}) \cdot h_j(\boldsymbol{x}) \\
&\quad + \mathop{\mathbb{E}}_{h_i\sim Q} \mathop{\mathbb{E}}_{h_j\sim Q} h_j(\boldsymbol{x}) \cdot \Big(h_i(\boldsymbol{x}+\boldsymbol{n}) - h_i(\boldsymbol{x}) \Big)
\end{aligned}$$

Substituting the last expression in the inequality of Definition 7 yields the result of the lemma.

Notably from Lemma 2, since the L.H.S of its inequality expresses $a_Q(\boldsymbol{x})$, noise invariance η_Q^D becomes a function of the agreement between individual voters expressed in Definition 6. Next, we show how to leverage Lemma 2 in order to tractably compute the resilience of majority vote classifiers $B_Q(\boldsymbol{x})$.

5.1. Exact Expressions of η_Q^D for Linear Kernels

So far, we discussed general Bayes classifiers defined over sets of kernelised voters, where each kernel is parameterized by training examples $(\boldsymbol{x}_i', y_i) \in \mathcal{S}$. Interestingly, some voter kernels yield exact expressions of Lemma 2 without the need to calculate expectations over $\boldsymbol{n}$, and salient among these are linear classifiers $h_i(\boldsymbol{x}) = y_i \boldsymbol{x}_i' \boldsymbol{x}^\top$ that specify constant differentials $\frac{\delta h_i(\boldsymbol{x})}{\delta \boldsymbol{x}} = y_i \boldsymbol{x}_i'$. By exploiting the commutative property of linear products to simplify $h_i(\boldsymbol{x}+\boldsymbol{n}) - h_i(\boldsymbol{x})$, the next lemma collapses the multivariate expression of noise in Lemma 2 to a tractable univariate.

Lemma 3 For any majority vote classifier defining a posterior Q over normalised linear voters $\boldsymbol{x}_i' \in \mathcal{S} \subseteq \mathcal{X}$ where $h_i(\boldsymbol{x}) = y_i \boldsymbol{x}_i' \boldsymbol{x}^\top$, and when $\boldsymbol{\omega} \in \mathbb{R}^d$ is a variate defining $\boldsymbol{\omega} = \mathbb{E}_{x_i'\sim Q} \mathbb{E}_{x_j'\sim Q} \boldsymbol{x}_j' \boldsymbol{x}^\top \boldsymbol{x}_i'$, noise invariance η_Q^D becomes:

$$\eta_Q^D = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim D} Pr_{t\sim\mathcal{N}(\mathbf{0}, \boldsymbol{\omega}^\top \sigma_c^2 I \boldsymbol{\omega})} \Big(a_Q(\boldsymbol{x}) + t < 0 \Big)$$

Proof: From the R.H.S of Lemma 2, and by exploiting the commutative property of linear dot products:

$$Pr_{\boldsymbol{n}\sim\mathcal{N}(\mathbf{0},\sigma_c^2 I)} \left(\mathop{\mathbb{E}}_{h_i\sim Q} \mathop{\mathbb{E}}_{h_j\sim Q} h_j(\boldsymbol{x}) \cdot \Big(h_i(\boldsymbol{x}+\boldsymbol{n}) - h_i(\boldsymbol{x}) \Big) \right)$$
$$= Pr_{\boldsymbol{n}\sim\mathcal{N}(\mathbf{0},\sigma_c^2 I)} \left(\boldsymbol{n} \mathop{\mathbb{E}}_{x_i'\sim Q} \mathop{\mathbb{E}}_{x_j'\sim Q} \boldsymbol{x}_j' \boldsymbol{x}^\top \boldsymbol{x}_i' \right) = \mathcal{N}(\mathbf{0}, \boldsymbol{\omega}^\top \sigma_c^2 I \boldsymbol{\omega})$$

Letting $\boldsymbol{\omega} = \mathbb{E}_{x_i'\sim Q} \mathbb{E}_{x_j'\sim Q} \boldsymbol{x}_j' \boldsymbol{x}^\top \boldsymbol{x}_i'$ and exploiting the multiplication property of multivariates yields the last step, and the lemma is obtained by substituting $t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\omega} \sigma_c^2 I \boldsymbol{\omega}^\top)$ in Lemma 2.

From Lemma 3, we derive an analytic expression of η_Q^D for majority vote classifiers comprising linear kernels.

Theorem 2 For any majority vote classifier defining a posterior Q over normalised linear voters $\boldsymbol{x}_i' \in \mathcal{S} \subseteq \mathcal{X}$ where $h_i(\boldsymbol{x}) = y_i \boldsymbol{x}_i' \boldsymbol{x}^\top$, and when $\boldsymbol{\omega} = \mathbb{E}_{\boldsymbol{x}_i'\sim Q} \mathbb{E}_{\boldsymbol{x}_j'\sim Q} \boldsymbol{x}_j' \boldsymbol{x}^\top \boldsymbol{x}_i'$, invariance coefficients η_Q^D are simplified to:

$$\eta_Q^D = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim D} \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{a_Q(\boldsymbol{x})}{\sqrt{\boldsymbol{\omega} \sigma_c^2 I \boldsymbol{\omega}^\top} \sqrt{2}} \right) \right]$$

Proof: From the inequality of Lemma 3, moving t to the R.H.S gives:

$$\eta_Q^D = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim D} Pr_{t\sim\mathcal{N}(\mathbf{0}, \boldsymbol{\omega}\sigma_c^2 I \boldsymbol{\omega}^\top)} \Big(-a_Q(\boldsymbol{x}) \geq t \Big)$$

When $t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\omega}\sigma_c^2 I \boldsymbol{\omega}^\top)$, the last expression simplifies to the CDF: $\Phi(\frac{a_Q(\boldsymbol{x})-\mu}{\sigma}) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{a_Q(\boldsymbol{x})-\mu}{\sigma\sqrt{2}}\right)\right]$ where $\operatorname{erf}(z) \coloneqq \frac{2}{\sqrt{\pi}} \int e^{-z^2} dz$ defines the gaussian error function.

The result of Theorem 2 calculates η_Q^D knowing only $B_Q(\boldsymbol{x})$ and σ_c^2, and is tractable via importance sampling due to the collapsed univariate t. Hence, any non-vacuous bound on η_Q^D of Theorem 2 would satisfy all the requirements described in Section 3 characterising good indicators of over-compression. Also interesting to note here is that, whenever $\mathcal{N}(\boldsymbol{x} - \boldsymbol{x}_c | \mathbf{0}, \sigma_c^2 I)$ expresses an isotropic gaussian symmetric on $\mathbf{0}$, η_Q^D of Theorem 2 can never exceed 0.5. This is because $a_Q(\boldsymbol{x}) \geq 0$ such that Theorem 2 always returns the integration of $\mathcal{N}(0, \boldsymbol{\omega}\sigma_c^2 I \boldsymbol{\omega}^\top)$ up to the center of the gaussian, thereby ensuring $\eta_Q^D \leq 0.5$. This is also intuitive, since binary linear classifiers define demarcation hyperplanes in $\mathcal{X}$, and examples are always equally likely to move *towards* or *away* from classification hyperplanes.

5.2. Upper-Bounding η_Q^D and Distortion Inflicted Risk

To account for the double expectations η_Q^D defines over voter pairs (h_i, h_j), and in order to bound η_Q^D via Corollary 1, similar to (Germain et al., 2015) we consider the posteriors Q^2 defined over $\mathcal{H}^2 : \mathcal{H} \times \mathcal{H}$ denoting hypothesis sets comprising voter pairs $h_{ij} = (h_i, h_j)$. The product rule gives $Q^2(h_{ij}) = Q(h_i)Q(h_j)$, and the next theorem adapts Corollary 1 to η_Q^D.

Theorem 3 For any source distribution D, for any prior P^2 on the hypothesis set $\mathcal{H}^2$, for any posterior Q^2 learned by observing $\mathcal{S} \sim D^m$, and for any arbitrary probability $\delta_\eta \in (0, 1]$:

$$\Pr_{\mathcal{S}\sim D}\left(\mathrm{kl}(\eta_Q^{\mathcal{S}}||\eta_Q^D) \leq \frac{1}{m}\left[2KL(Q||P) + \ln\frac{\xi(m)}{\delta_\eta}\right]\right) \geq 1 - \delta_\eta$$

Proof: Because η_Q^D is a binary event defined over $Q^2(h_{ij}|\mathcal{S})$, the KL divergence term becomes:

$$\begin{aligned} KL(Q^2||P^2) &= \mathop{\mathbb{E}}_{h_{ij}\sim Q^2} \ln\left(\frac{Q^2(h_{ij})}{P^2(h_{ij})}\right) \\ &= \mathop{\mathbb{E}}_{h_{ij}\sim Q^2}\left(\ln\frac{Q(h_i)}{P(h_i)} + \ln\frac{Q(h_j)}{P(h_j)}\right) \end{aligned}$$

the last expression simplifies to $2KL(Q||P)$ and we then apply Corollary 1 to get the bound of the theorem. For a detailed account on paired voters and how they relate to Corollary 1, see (Germain et al., 2015).

Corollary 2 For any arbitrary probability $\delta_\eta \in (0, 1]$, there exists an upper bound $\eta_Q^D \leq \mathcal{B}_\eta$ with probabilty $\geq 1 - \delta_\eta$ when $\mathcal{B}_\eta$ defines the supremum:

$$\mathcal{B}_\eta \coloneqq \sup\left\{b_\eta : \mathrm{kl}(\eta_Q^{\mathcal{S}}||b_\eta) \leq \frac{1}{m}\left[2KL(Q||P) + \ln\frac{\xi(m)}{\delta_\eta}\right]\right\}$$

Proof: Upper bounds $\mathcal{B}_\eta$ on η_Q^D define the highest value of η_Q^D that satisfies Theorem 3, and this yields the corollary.

5.3. Bounding the Effect of η_Q^D on True Risk $R(B_Q)$

Bounds $\mathcal{B}_\eta$ on η_Q^D actually carry additional implicit bounds $\mathcal{B}_R$ on the true risk of $B_Q(\boldsymbol{x} + \boldsymbol{n})$ *after* noise distorts the input of majority votes $B_Q(\boldsymbol{x})$. Specifically, whenever the probability of output change η_Q^D is bounded by Theorem 3, and when $R(B_Q)$ measures the true risk of $B_Q(\boldsymbol{x})$ for non-perturbed inputs $\boldsymbol{x}$, the next corollary expresses $\mathcal{B}_R$ as a closed-form expression of $R(B_Q)$ and $\mathcal{B}_\eta$.

Corollary 3 For any source domain D, for any bound $\mathcal{B}_\eta$ on η_Q^D, there exists a bound $R(B_Q(\boldsymbol{x} + \boldsymbol{n})) \leq \mathcal{B}_R$ after noise $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma_c^2 I)$ is applied on $\boldsymbol{x}$, where $\mathcal{B}_R$ measures:

$$\mathcal{B}_R \coloneqq R(B_Q) + \mathcal{B}_\eta - R(B_Q) \cdot \mathcal{B}_\eta$$

Proof: When $B_Q(\boldsymbol{x} + \boldsymbol{n})$ infers targets from m perturbed examples $\boldsymbol{x}$, the total number of correct outputs becomes $\big[1 - R(B_Q)\big]m - \big[1 - R(B_Q)\big]\mathcal{B}_\eta m$, and $\mathcal{B}_R$ measures:

$$\begin{aligned} \mathcal{B}_R &\coloneqq 1 - \frac{m\Big(\big[1 - R(B_Q)\big] - \big[1 - R(B_Q)\big]\mathcal{B}_\eta\Big)}{m} \\ &= 1 - \Big(\big[1 - R(B_Q)\big]\big[1 - \mathcal{B}_\eta\big]\Big) \end{aligned}$$

The corollary follows by simplifying the last expression.

Closing remarks on $\mathcal{B}_\eta$: From Corollary 2 we observe that upper bounds $\mathcal{B}_\eta$ on η_Q^D are actually functions of: (i) the posterior $Q(h_i|\mathcal{S})$, (ii) the prior $P(h_i)$, and (iii) the variance σ_c^2 of input compression noise $\boldsymbol{n}$. Therefore, by combining the bound of Theorem 3 with empirical estimates of σ_c^2 at test time, implications of compression are known before outputs of majority vote classifiers $B_Q(\boldsymbol{x})$ are observed. Thus, using Theorem 3 and following the prescription of Section 3, we are able to realize rate-efficient machines that compress inputs up to known limits that guarantee the bound $\eta_Q^D \leq \mathcal{B}_\eta$ with probability $\geq 1 - \delta_\eta$.

5.4. Final Observations on Voter Kernels and η_Q^D

On regularizing kernel selection: When classifiers $B_Q(\boldsymbol{x})$ are defined per Definition 2 as majority votes over a set of training kernels $\boldsymbol{x}_i' \in \mathcal{S}$, Definition 7 expresses noise invariance coefficients η_Q^D as functions of: individual voters $h_i(\boldsymbol{x})$, training kernels $\boldsymbol{x}_i'$, and parameters that weight the importance of each voter $Q(h_i|\mathcal{S})$. Thus, noise invariant bayes classifiers $B_Q(\boldsymbol{x})$ specify posteriors $Q(h_i|\mathcal{S})$ that prioritise smooth voters with kernels minimising $\frac{\delta h_i(\boldsymbol{x})}{\delta \boldsymbol{x}}$. Regulating training to learn such posteriors is possible via the MinCq learning algorithm (Germain et al., 2013), where some voters can be favoured to others by setting higher probabilities $Q(h_i|\mathcal{S})$ for kernels $\boldsymbol{x}_i'$ that return lower differentials $\frac{\delta h_i(\boldsymbol{x})}{\delta \boldsymbol{x}}$. Further exploration of this is outside the scope of our proposal, and we leave implementations of noise invariance regulation to future work.

On viable kernel types: Generally, the bound of Theorem 3 is applicable for any majority vote whenever voters $h_i(\boldsymbol{x})$ are continuously differentiable. For example, this is the case for Multi-Layer Perceptrons (MLP), which follow linear transformations by sigmoid non-linearities such that $K(\boldsymbol{x}, \boldsymbol{x}_i') = [1 + \exp(-\boldsymbol{x}_i'\boldsymbol{x}^\top)]^{-1}$. Incidentally, MLPs can also be construed as majority vote classifiers, because individual weight vectors make decisions on $\boldsymbol{x}$ that are subsequently pooled. Moreover, and since bounds of Theorem 2 can be calculated prior to compression, the tractability of the compressive loop in Figure 1 is not affected by the complexity of kernels. We leave to future work the derivation of non-vacuous bounds on η_Q^D for majority vote classifiers with non-linear kernels.

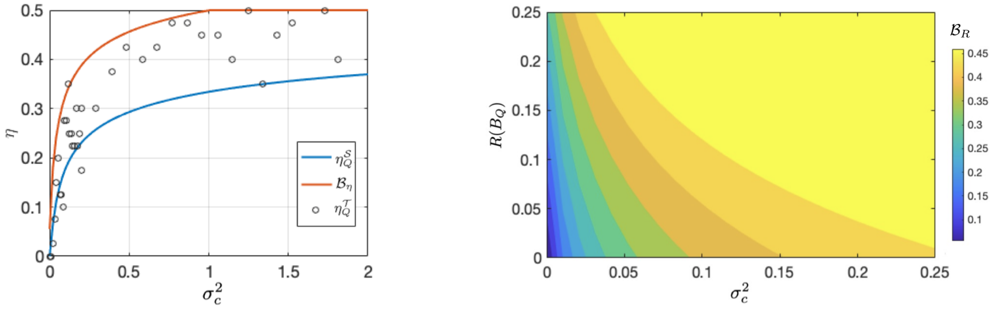

Figure 2. (Left) Noise invariance bound $\mathcal{B}_\eta$, training noise invariance $\eta_Q^{\mathcal{S}}$, and testing noise invariance $\eta_Q^{\mathcal{T}}$ for increasing values of input degradation as quantified by σ_c^2. (Right) Contours of the true risk bound $\mathcal{B}_R$ after outputs are distorted (flipped) with probability η_Q^D ; yellower hues indicate higher values of $\mathcal{B}_R(B_Q)$.

6. Related Work

The notion of noise invariance is closely related - but not equivalent - to adversarial robustness (Goodfellow et al., 2015), which measures the probability $P(\mathcal{M}(\boldsymbol{x}+\boldsymbol{n}^*) \neq y)$ of misclassifying ground truth y due to learnt perturbations $\boldsymbol{n}^*(\boldsymbol{x}, y, \mathcal{M})$, where $\boldsymbol{n}^* = \arg\max_{\boldsymbol{n}} P(\mathcal{M}(\boldsymbol{x}+\boldsymbol{n}) \neq y)$. To study the adverse effect of input perturbations, recent proposals (Cohen et al., 2019; Montasser et al., 2020; Rahimian, 2019; Vidot et al., 2021) introduced various bounds on adversarial robustness and decision making under uncertainty. For instance, (Cohen et al., 2019) provide bounds that estimate minimum perturbation norms required to yield adversarial examples, and (Salman et al., 2019) extend the method of (Cohen et al., 2019) to yield tighter bounds via adversarial training. More recently, (Vidot et al., 2021) provide bounds on adversarial robustness in white-box settings to understand when differentiable decision trees fail against adversarial attacks with high probabilities $P(\mathcal{M}(\boldsymbol{x}+\boldsymbol{n}^*) \neq y)$. Importantly, the contribution of (Vidot et al., 2021) gives *supervised* bounds on risk averaged over sets of *learnt* perturbations $\boldsymbol{n}^*$, and is not applicable to the iterative compression setting detailed in Section 3.

While the works cited above all study different aspects of perturbation invariance, our method contrasts (Cohen et al., 2019; Montasser et al., 2020; Vidot et al., 2021) in that it is the first to: (i) yield unsupervised bounds on probabilities $P(\mathcal{M}(\boldsymbol{x}+\boldsymbol{n}) \neq \mathcal{M}(\boldsymbol{x}))$ outside the context of adversarial robustness that assumes knowledge of ground truth y and posteriors $P(\boldsymbol{n}^*|\boldsymbol{x}, \mathcal{M})$ that maximise $P(\mathcal{M}(\boldsymbol{x}+\boldsymbol{n}^*) \neq y)$, (ii) define tractable bounds as functions of $(\mathcal{M}, \sigma_c^2)$ that can be measured offline prior to compression, and (iii) give bounds on η_Q^D that are not averaged over perturbation sets, where η_Q^D measures the probability of flipped outputs for any input $\boldsymbol{x} \in \mathcal{X}$. The unique set of properties detailed in (i)-(iii) allow us to integrate PAC-Bayesian bounds into the compression loop of Figure 1 to derive the rate-efficient PAC-Bayesian classifiers detailed in Section 3.

7. Evaluation

Test setting: To address a class of problems where inputs are typically costly to compress and stream prior to inference, and to overlap our evaluation with vision applications where inputs undergo lossy compression, we focus our experimental validation on joint image compression and classification. We evaluate the measures and bounds of Theorem 3 and Corollary 3 on the handwritten digits dataset MNIST (Deng, 2012) in two distinct settings:

1. Controlled test conditions where noise is drawn directly from gaussian densities $\mathcal{N}(\mathbf{0}, \sigma_c^2)$ to perturb inputs $\boldsymbol{x}$, and σ_c^2 is directly specified.
2. A distributed visual inference setting, where inputs are compressed via JPEG2000 (Rabbani, 2002) prior to inference, and compression is tuned via a quality parameter q. Compression noise here is induced “naturally” when the compressor fails to accurately reconstruct $\boldsymbol{x}$.

Adapting MNIST to binary classification: Following established practice (Germain et al., 2015; Letarte et al., 2019), we split MNIST into 45 binary classification tasks[1], where each task is exclusive to a unique pair of MNIST classes. For each task, a training set $\mathcal{S}$ with $m = 500$ is randomly sampled, and remaining examples go to a test set $\mathcal{T}$ used for validation. Each task uses a unique set of examples, such that any pair of datasets returns the empty set, and combining all datasets returns all examples in MNIST. All results are averaged after validating on all 45 unique tasks.

Measuring σ_c^2: For tests on naturally occurring noise, we derive measures on σ_c^2 by iterating over JPEG2000 compression rounds until a specified target σ_t^2 is met (as per Figure 1). When JPEG$(\boldsymbol{x}, \theta_c)$ denotes JPEG compression (Rabbani, 2002) with parameters θ_c, and $\sigma^2(\cdot)$ is the variance of fitted isotropic gaussians $\mathcal{N}(\mathbf{0}, \sigma^2)$, we perform:

$$\sigma_c^2 = \max_{\theta_c} \ \sigma^2\Big(\boldsymbol{x} - \text{JPEG}(\boldsymbol{x}, \theta_c)\Big) \ \text{ s.t. } \ \sigma_c^2 \leq \sigma_t^2$$

[1] MNIST classes give $10? - 10 = 45$ binary classification tasks, where $k?$ denotes the k^{th} triangle sum (Knuth, 2014).

Table 1. Training and testing noise invariance $(\eta_Q^{\mathcal{S}}, \eta_Q^{\mathcal{T}})$, noise invariance bounds $\mathcal{B}_\eta$, and test risk $R_{\mathcal{T}}(B_Q)$ as measured on MNIST. Lower values of $(\eta_Q^{\mathcal{S}}, \eta_Q^{\mathcal{T}}, \mathcal{B}_\eta, R_{\mathcal{T}})$ indicate higher resilience to noise and correspond to higher rate-efficiency. For MNIST-JPEG we report d_c as the average number of bytes per image as encoded by JPEG (Rabbani, 2002), and $\boldsymbol{n} \sim C(\boldsymbol{x}, \theta_c)$ denotes compression noise.

σ_c^2	MNIST $(\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_c^2))$					σ_c^2	MNIST-JPEG $(\boldsymbol{n} \sim C(\boldsymbol{x}, \theta_c))$						
	$\eta_Q^{\mathcal{S}}$	$\eta_Q^{\mathcal{T}}$	$\lvert\eta_Q^{\mathcal{T}} - \mathcal{B}_\eta\rvert$	$\mathcal{B}_\eta$	$R_{\mathcal{T}}$		q	d_c	$\eta_Q^{\mathcal{S}}$	$\eta_Q^{\mathcal{T}}$	$\lvert\eta_Q^{\mathcal{T}} - \mathcal{B}_\eta\rvert$	$\mathcal{B}_\eta$	$R_{\mathcal{T}}$
0.010	0.046	0.013	0.149	0.162	0.24	0.016	90	571	0.001	0.000	0.060	0.060	0.19
0.100	0.180	0.275	0.058	0.333	0.29	0.022	70	293	0.002	0.008	0.056	0.064	0.23
0.250	0.235	0.348	0.052	0.400	0.32	0.023	50	254	0.002	0.007	0.053	0.060	0.26
0.500	0.294	0.392	0.063	0.455	0.36	0.023	30	208	0.002	0.026	0.043	0.069	0.26
0.750	0.315	0.411	0.077	0.488	0.39	0.026	10	194	0.003	0.064	0.003	0.067	0.27
1.000	0.338	0.470	0.031	0.501	0.44	0.029	5	188	0.004	0.069	0.002	0.071	0.28

Used Classifiers: We construct majority votes on finite self-complemented hypotheses sets $\mathcal{H}$ (Germain et al., 2015) of real-valued voters $h_i(\boldsymbol{x})$ to learn $Q(h_i|\mathcal{S})$ via MinCq (Jean, 2019) when posteriors $Q(h_i|\mathcal{S})$ align with priors $P(h_i)$. Specifically, our implementation[2] assigns training sets $\mathcal{S}$ of m training examples $\boldsymbol{x}'_i \in \mathcal{S}$ to constitute kernels of linear voters $h_i(\boldsymbol{x}) = \boldsymbol{x}'_i \boldsymbol{x}^\top$ that define a majority vote classifier $B_Q(\boldsymbol{x})$, and we use the quadratic program of (Jean, 2019) to solve for $Q(h_i|\mathcal{S})$. To validate the unsupervised bounds of Theorem 3 and Corollary 3, Table 1 and Figure 2 report relevant empirical measures on noise invariance when $\delta = 0.05$ and $\delta_\eta = 0.05$. For all parameters not explicitly mentioned in our discussion, we retain specifications of the GRAAL-Research MinCq implementation (Jean, 2019).

7.1. Evaluation on Controlled Noise

Table 1 (left) validates the results of applying Theorem 3 and Corollary 2 on MNIST examples when artificial noise is applied to $\boldsymbol{x}$ such that $\boldsymbol{x}_c = \boldsymbol{x} + \boldsymbol{n}$ and $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 I)$. We observe from Table 1 (left) that training noise invariance coefficients $\eta_Q^{\mathcal{S}}$ are around 0.3 when $\sigma_c^2 \geq 0.5$, and are at most 0.235 when $\sigma_c^2 \leq 0.25$, giving a baseline on η_Q^D for source domains similar to those of MNIST. Moreover, we note that the difference $|\eta_Q^{\mathcal{T}} - \mathcal{B}_\eta|$ never exceeds 0.08 when $\sigma_c^2 \geq 0.1$, showing that $\mathcal{B}_\eta$ is actually very informative of true noise invariance coefficients for higher values of σ_c^2, where $\eta_Q^{\mathcal{T}}$ never deviates substantially from $\mathcal{B}_\eta$. This is also reflected in Figure 2 (left), where test set measures of $\eta_Q^{\mathcal{T}}$ neatly track $\eta_Q^{\mathcal{S}}$ for $\sigma_c^2 \leq 0.1$, and mostly fall in the mid-point between $\eta_Q^{\mathcal{S}}$ and $\mathcal{B}_\eta$ for $\sigma_c^2 \geq 0.2$.

7.2. Evaluation on JPEG2000 Compression Noise

Table 1 (right) provides results for noise stemming from JPEG compression, the most common in image compression practice. We emulate a typical distributed visual inference setting as per Figure 1, where inputs are compressed prior to inference to match specified constraints on bitrate. To meet any PAC guarantee on noise invariance, we calculate the target σ_t^2 that yields a specified $\eta_Q^{\mathcal{S}}$ from Theorem 3, and vary JPEG compression parameters θ_c to fit σ_c^2 as prescribed by Equation 7. Table 1 details relevant results under MNIST-JPEG, and additionally reports the dimensions of the compression latent space d_c as the average number of bytes resulting from JPEG encoding.

Similar to our observations on synthetically perturbed MNIST inputs, the right half of Table 1 reports low absolute differences $|\eta_Q^{\mathcal{T}} - \mathcal{B}_\eta|$, indicating that bounds on $\eta_Q^{\mathcal{T}}$ are valid even when noise is drawn from complex structures (i.e., the non-linear block models of JPEG in this case). Interestingly, we see that noise variance σ_c^2 induced by JPEG compression is always $\in [0.016, 0.026]$, and so only the bluer left extremes of Figure 2 (right) become relevant for JPEG compression. By inspecting Figure 2 (right) when $\sigma_c^2 \in [0.016, 0.026]$, we note that the discrepancy between $R(B_Q)$ and $\mathcal{B}_R(B_Q)$ is relatively low compared to exponentially increasing differences $\mathcal{B}_R - R(B_Q)$ when approaching the right extremities of Figure 2 (right), indicated by yellower hues. Thus, whenever σ_c^2 is sufficiently low, bounds on $\mathcal{B}_R$ give value to applications that require guarantees on risk degradation induced by lossy input compression.

8. Conclusion

We introduce the notion of rate-efficient classifiers to PAC-Bayesian theory. The unsupervised noise invariance bounds we derive are intuitive and highlight the importance of voters that change gradually with respect to perturbations on input. By inspecting the variance of gaussians fitted on compression noise, our evaluation shows that the proposed bounds are non-vacuous and well-suited to manage compression noise prior to inference, and we demonstrated this on JPEG2000 compression. Notably, the presented noise invariance bounds are general, and applicable wherever inputs are sourced from gaussian processes (e.g., in models of disease progression, physics, and signal processing). Future work may investigate more applications, derivations of η_Q^D for non-linear kernels, and regularization techniques that give tighter bounds on η_Q^D.

[2] https://github.com/git-alhabib/pacb-ni

9. Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council [Grant ID: EP/P02243X/1, EP/R025290/1, EP/R035342/1, EP/S028455/1].

References

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Catoni, O. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *Lecture Notes-Monograph Series*, 56, 2007.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

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

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 353–360, 2009.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pp. 738–746. PMLR, 2013.

Germain, P., Lacasse, A., Laviolette, F., Marchand, M., and Roy, J. F. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

Gholami, R. and Fakhari, N. Support vector machine: principles, parameters, and applications. In *Handbook of Neural Computation*, pp. 515–535. Elsevier, 2017.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Gray, R. M. and Neuhoff, D. L. Quantization. *IEEE transactions on information theory*, 44(6):2325–2383, 1998.

James, G. M. Majority vote classifiers: theory and applications. Stanford University, 1998.

Jean, R. https://github.com/graal-research/mincq. *GRAAL-Research*, 2019.

Knuth, D. *The Art of Computer Programming*. Addison-Wesley Professional, 2014.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. PAC-Bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In *NIPS*, pp. 769–776, 2006.

Langford, J. and Shawe-Taylor, J. PAC-Bayes & margins. *Advances in neural information processing systems*, 15: 439–446, 2002.

Letarte, G., Germain, P., Guedj, B., and Laviolette, F. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 6872–6882, 2019.

McAllester, D. A. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

McAllester, D. A. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Montasser, O., Hanneke, S., and Srebro, N. Reducing adversarially robust learning to non-robust PAC learning. 2020.

Pradhan, S. S., Kusuma, J., and Ramchandran, K. Distributed compression in a dense microsensor network. *IEEE Signal Processing Magazine*, 19(2):51–60, 2002.

Rabbani, M. JPEG2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11 (2):286, 2002.

Rahimian, H. Distributionally robust optimization: A review. 2019.

Sagi, O. and Rokach, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32:11292–11303, 2019.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Vidot, E., Viallard, P., Guillaume, Habrard, A., and Morvant, E. A PAC-Bayes analysis of adversarial robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiao, J., Ribeiro, A., Luo, Z.-Q., and Giannakis, G. B. Distributed compression-estimation using wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):27–41, 2006.