Recursive PAC-Bayes: A Frequentist Approach to Sequential Prior Updates with No Information Loss

Yi-Shan Wu
University of South Denmark
yswu@imada.sdu.dk

Yijie Zhang
University of Copenhagen & Novo Nordisk A/S
yizh@di.ku.dk

Badr-Eddine Chérief-Abdellatif
CNRS, LPSM, Sorbonne Université, Université Paris Cité
badr-eddine.cherief-abdellatif@cnrs.fr

Yevgeny Seldin
University of Copenhagen
seldin@di.ku.dk

Abstract

PAC-Bayesian analysis is a frequentist framework for incorporating prior knowledge into learning. It was inspired by Bayesian learning, which allows sequential data processing and naturally turns posteriors from one processing step into priors for the next. However, despite two and a half decades of research, the ability to update priors sequentially without losing confidence information along the way remained elusive for PAC-Bayes. While PAC-Bayes allows construction of data-informed priors, the final confidence intervals depend only on the number of points that were not used for the construction of the prior, whereas confidence information in the prior, which is related to the number of points used to construct the prior, is lost. This limits the possibility and benefit of sequential prior updates, because the final bounds depend only on the size of the final batch.

We present a novel and, in retrospect, surprisingly simple and powerful PAC-Bayesian procedure that allows sequential prior updates with no information loss. The procedure is based on a novel decomposition of the expected loss of randomized classifiers. The decomposition rewrites the loss of the posterior as an excess loss relative to a downscaled loss of the prior plus the downscaled loss of the prior, which is bounded recursively. As a side result, we also present a generalization of the split-kl and PAC-Bayes-split-kl inequalities to discrete random variables, which we use for bounding the excess losses, and which can be of independent interest. In empirical evaluation the new procedure significantly outperforms state-of-the-art.

1 Introduction

PAC-Bayesian analysis was born from an attempt to derive frequentist generalization guarantees for Bayesian-style prediction rules (Shawe-Taylor and Williamson, 1997, McAllester, 1998). The motivation was to provide a way to incorporate prior knowledge into the frequentist analysis of generalization. PAC-Bayesian bounds provide high-probability generalization guarantees for randomized classifiers. A randomized classifier is defined by a distribution ρ on a set of prediction rules $\mathcal{H}$, which is used to sample a prediction rule each time a prediction is to be made. Bayesian posterior is an example of a randomized classifier, whereas PAC-Bayesian bounds hold generally for all randomized classifiers. Prior knowledge is encoded through a prior distribution π on $\mathcal{H}$, and the complexity of a posterior distribution ρ is measured by the Kullback-Leibler (KL) divergence from the prior, $\mathrm{KL}(\rho\|\pi)$. PAC-Bayesian generalization guarantees are optimized by posterior distributions ρ that optimize a trade-off between empirical data fit and divergence from the prior in the KL sense.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0570

Selection of a "good" prior plays an important role in the PAC-Bayesian bounds. If one manages to foresee which prediction rules are likely to produce low prediction error and allocate a higher prior mass for them, then the bounds are tighter, because the posterior only needs to make a small deviation from the prior. But if the prior mass on well-performing prediction rules is small, the bounds are loose. A major technique to design good priors is to use part of the data to estimate a good prior and the rest of the data to compute a PAC-Bayes bound. It is known as data-dependent or data-informed priors (Ambroladze et al., 2007). However, all existing approaches to data-informed priors have three major disadvantages. The first is that the bounds are computed on "the rest of the data" that were not used in construction of the prior. Thus, the sample size in the bounds is only a fraction of the total sample size. Therefore, empirically data-informed priors are not always helpful. In many cases starting with an uninformed prior and using all the data to compute the posterior and the bound turns to be superior to sacrificing part of the data for prior construction (Ambroladze et al., 2007, Mhammedi et al., 2019). The second disadvantage is that all the confidence information about the prior is lost in the process. In particular, a prior trained on a few data points is treated in the same way as a prior trained on a lot of data. And a third related disadvantage is that sequential data processing provides no benefit, because the bounds only depend on the size of the last chunk and all the confidence information from processing earlier chunks is lost in the process.

Our main contribution is a new (and simple) way of decomposing the loss of a randomized classifier defined by the posterior. We write it as an excess loss relative to a downscaled loss of the randomized classifier defined by the prior plus the downscaled loss of the randomized classifier defined by the prior. The excess loss can be bounded using PAC-Bayes-Empirical-Bernstein-style inequalities (Tolstikhin and Seldin, 2013, Mhammedi et al., 2019, Wu et al., 2021, Wu and Seldin, 2022), whereas the loss of the randomized classifier defined by the prior can be bounded recursively. The recursive bound can both use the data used for construction of the prior and "the rest of the data", and thereby preserves confidence information on the prior. Our contribution stands out relative to all prior work on PAC-Bayes, and in fact all prior work on frequentist generalization bounds, because it makes sequential data processing and sequential prior updates meaningful and beneficial.

We note that while several recent papers experimented with sequential posterior updates by using martingale-style analysis, in all these works the prior remained fixed and only the posterior was changing (Chugg et al., 2023, Biggs and Guedj, 2023, Rodríguez-Gálvez et al., 2024). The work on sequential posterior updates is orthogonal to our contribution and can be combined with it. Namely, it is possible to apply sequential posterior updates in-between sequential prior updates. Another line of work used tools from online learning to derive PAC-Bayesian bounds (Jang et al., 2023), and in this context Haddouche and Guedj (2023) have used sequential prior updates, but their bounds hold for a uniform aggregation of sequentially constructed posteriors, which is different from standard posteriors studied in our work. The confidence bounds in their work come primarily from aggregation rather than confidence in individual posteriors in the sequence (the denominator of their bounds depends on the number of aggregated posteriors). The need to construct and maintain a large number of posteriors has a negative impact on the computational efficiency. Our work is the first one allowing sequential prior updates without loss of confidence information.

An additional side contribution of independent interest is a generalization of the split-kl and PAC-Bayes-split-kl inequalities of Wu and Seldin (2022) from ternary to general discrete random variables. It is based on a novel representation of discrete random variables as a superposition of Bernoulli random variables.

The paper is organized in the following way. In Section 2 we briefly survey the evolution of data-informed priors in PAC-Bayes and present our main idea behind Recursive PAC-Bayes; in Section 3 we present our generalization of the split-kl and PAC-Bayes-split-kl inequalities, which are later used to bound the excess losses; in Section 4 we present the Recursive PAC-Bayes bound; in Section 5 we present an empirical evaluation; and in Section 6 we conclude with a discussion.

2 The evolution of data-informed priors and the idea of Recursive PAC-Bayes

In this section we briefly survey the evolution of data-informed priors, and then present our construction of Recursive PAC-Bayes. We consider the standard classification setting, with $\mathcal{X}$ being a sample space, $\mathcal{Y}$ a label space, $\mathcal{H}$ a set of prediction rules $h : \mathcal{X} \to \mathcal{Y}$, and $\ell(h(X), Y) = \mathbb{1}(h(X) \neq Y)$ the zero-one loss function, where $\mathbb{1}(\cdot)$ denotes the indicator function. We let $\mathcal{D}$

denote a distribution on $\mathcal{X} \times \mathcal{Y}$ and $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ an i.i.d. sample from $\mathcal{D}$. Let $L(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h(X), Y)]$ be the expected and $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$ the empirical loss.

Let ρ be a distribution on $\mathcal{H}$. A *randomized classifier* associated with ρ samples a prediction rule h according to ρ for each sample $X \in \mathcal{X}$, and applies it to make a prediction $h(X)$. The expected loss of such randomized classifier, which we call ρ, is $\mathbb{E}_{h\sim\rho}[L(h)]$ and the empirical loss is $\mathbb{E}_{h\sim\rho}[\hat{L}(h, S)]$. For brevity we use $\mathbb{E}_\rho[\cdot]$ to denote $\mathbb{E}_{h\sim\rho}[\cdot]$.

We use $\mathrm{KL}(\rho\|\pi)$ to denote the Kullback-Leibler divergence between two probability distributions, ρ and π (Cover and Thomas, 2006). For $p, q \in [0, 1]$ we further use $\mathrm{kl}(p\|q) = \mathrm{KL}((1-p, p)\|(1-q, q))$ to denote the Kullback-Leibler divergence between two Bernoulli distributions with biases p and q.

The goal of PAC-Bayes is to bound $\mathbb{E}_\rho[L(h)]$. Below we present how the bounds on $\mathbb{E}_\rho[L(h)]$ have evolved. In Appendix A we also provide a graphical illustration of the evolution.

Uninformed priors Early work on PAC-Bayes used *uninformed priors* (McAllester, 1998). An uniformed prior π is a distribution on $\mathcal{H}$ that is independent of the data S. A classical, and still one of the tightest bounds, is the following.

Theorem 1 (PAC-Bayes-kl Inequality, Seeger, 2002, Maurer, 2004)**.** *For any probability distribution π on $\mathcal{H}$ that is independent of S and any $\delta \in (0, 1)$:*

$$\mathbb{P}\left(\exists \rho \in \mathcal{P} : \mathrm{kl}\left(\mathbb{E}_\rho[\hat{L}(h, S)] \middle\| \mathbb{E}_\rho\left[L(h)\right]\right) \geq \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{n}\right) \leq \delta,$$

where $\mathcal{P}$ is the set of all probability distributions on $\mathcal{H}$, including those dependent on S.

A posterior ρ that minimizes $\mathbb{E}_\rho[L(h)]$ has to balance between allocating higher mass to prediction rules h with small $\hat{L}(h, S)$ and staying close to π in the $\mathrm{KL}(\rho\|\pi)$ sense. Since π has to be independent of S, typical uninformed priors aim "to leave maximal options open" for ρ by staying close to uniform.

Data-informed priors Ambroladze et al. (2007) proposed to split the data S into two disjoint sets, $S = S_1 \cup S_2$, and use S_1 to construct a *data-informed prior* π and compute a bound on $\mathbb{E}_\rho[L(h)]$ using π and S_2. Since in this approach π is independent of S_2, Theorem 1 can be applied. The advantage is that π can use S_1 to give higher mass to promising classifiers, thus relaxing the regularization pressure $\mathrm{KL}(\rho\|\pi)$ and making it easier for ρ to allocate even higher mass to well-performing classifiers (those with small $\hat{L}(h, S_2)$). The disadvantage is that the sample size in the bound (the n in the denominator) decreases from the size of S to the size of S_2. Indeed, Ambroladze et al. observed that the sacrifice of S_1 for prior construction does not always pay off.

Data-informed priors + excess loss Mhammedi et al. (2019) observed that if we have already sacrificed S_1 for the construction of π, we could also use it to construct a reference prediction rule h^*, typically an Empirical Risk Minimizer (ERM) on S_1. They then employed the decomposition

$$\mathbb{E}_\rho[L(h)] = \mathbb{E}_\rho[L(h) - L(h^*)] + L(h^*)$$

and used S_2 to give a PAC-Bayesian bound on $\mathbb{E}_\rho[L(h) - L(h^*)]$ and a single-hypothesis bound on $L(h^*)$. The quantity $\mathbb{E}_\rho[L(h) - L(h^*)]$ is known as *excess loss*. The advantage of this approach is that when $L(h^*)$ is a good approximation of $\mathbb{E}_\rho[L(h)]$, the excess loss has lower variance than the plain loss $\mathbb{E}_\rho[L(h)]$ and, therefore, is more efficient to bound, whereas the single-hypothesis bound on $L(h^*)$ does not involve the $\mathrm{KL}(\rho\|\pi)$ term. Therefore, it is generally beneficial to use excess losses in combination with data-informed priors. However, as with the previous approach, sacrificing S_1 to learn π and h^* means that the denominator in the bounds (n in Theorem 1) reduces to the size of S_2, and it does not always pay off. (We note that the excess loss is not binary and not in the $[0, 1]$ interval, and in order to exploit small variance it is actually necessary to apply a PAC-Bayes-Empirical-Bernstein-style inequality (Tolstikhin and Seldin, 2013, Mhammedi et al., 2019, Wu et al., 2021) or the PAC-Bayes-split-kl inequality (Wu and Seldin, 2022) rather than Theorem 1, but the point about reduced sample size still applies.)

Recursive PAC-Bayes (new) We introduce the following decomposition of the loss

$$\mathbb{E}_\rho[L(h)] = \mathbb{E}_\rho[L(h) - \gamma\mathbb{E}_\pi[L(h')]] + \gamma\mathbb{E}_\pi[L(h')]. \tag{1}$$

https://doi.org/10.52202/079017-0570

As before, we decompose S into two disjoint sets $S = S_1 \cup S_2$. We make the following major observations:

- The quantity $\mathbb{E}_\pi[L(h')]$ on the right is "of the same kind" as $\mathbb{E}_\rho[L(h)]$ on the left.
- We can take an uninformed prior π_0 and apply Theorem 1 (or any other suitable PAC-Bayes bound) to bound $\mathbb{E}_\pi[L(h')]$. (The KL term in the bound on $\mathbb{E}_\pi[L(h')]$ will be $\mathrm{KL}(\pi\|\pi_0)$.)
- We can restrict π to depend only on S_1, but still use all the data S in calculation of the PAC-Bayes bound on $\mathbb{E}_\pi[L(h')]$, because π is a posterior relative to π_0, and a posterior is allowed to depend on all the data, and in particular on any subset of the data. Therefore, the empirical loss $\mathbb{E}_\pi[\hat{L}(h', S)]$ can be computed on all the data S, and the denominator of the bound in Theorem 1 can be the size of S, and not the size of S_2. This is what we call *preservation of confidence information on* π, because all the data S are used to construct a confidence bound on $\mathbb{E}_\pi[L(h')]$, and not just S_2. This is in contrast to the bound on $L(h^*)$ in the approach of Mhammedi et al. (2019), which only allows to use S_2 for bounding $L(h^*)$. Note that while we use all the data S in calculation of the bound, we only use S_1 and $\mathbb{E}_\pi[\hat{L}(h', S_1)]$ in the construction of π. Nevertheless, we can still use the knowledge that we will have n samples when we reach the estimation phase, i.e., when constructing π we can leave the denominator of the bound at n, allowing more aggressive deviation from π_0.
- If we restrict π to depend only on S_1, then it is a valid prior for estimation of any posterior quantity $\mathbb{E}_\rho[\cdot]$ based on S_2. Thus, if we also restrict γ to depend only on S_1, we can use any PAC-Bayes-Empirical-Bernstein-style inequality or the PAC-Bayes-split-kl inequality to estimate the excess loss $\mathbb{E}_\rho[L(h) - \gamma\mathbb{E}_\pi[L(h')]]$ based on S_2, i.e., based on $\mathbb{E}_\rho[\hat{L}(h, S_2) - \gamma\mathbb{E}_\pi[\hat{L}(h', S_2)]]$. If $\gamma\mathbb{E}_\pi[L(h')]$ is a good approximation of $\mathbb{E}_\rho[L(h)]$ and $\mathbb{E}_\rho[L(h)]$ is not close to zero, then the excess loss $\mathbb{E}_\rho[L(h) - \gamma\mathbb{E}_\pi[L(h')]]$ is more efficient to bound than the plain loss $\mathbb{E}_\rho[L(h)]$.
- In general, since $\mathbb{E}_\rho[L(h)]$ is expected to improve on $\mathbb{E}_\pi[L(h')]$, it is natural to set $\gamma < 1$. However, γ is not allowed to depend on S_2, because otherwise $\hat{L}(h, S_2) - \gamma\mathbb{E}_\pi[\hat{L}(h', S_2)]$ becomes a biased estimate of $L(h) - \gamma\mathbb{E}_\pi[L(h')]$. We discuss the choice of γ in more detail when we present the bound and the experiments.
- Biggs and Guedj (2023) have proposed a sequential martingale-style evaluation of a martingale version of $\mathbb{E}_\rho[L(h) - L(h^*)]$ and $L(h^*)$ in the approach of Mhammedi et al., but it has not been shown to yield significant improvements yet. The same "martingalization" can be directly applied to our decomposition, but to keep things simple we stay with the basic decomposition.
- Finally, we note that we can split S_1 further and apply (1) recursively to bound $\mathbb{E}_\pi[L(h')]$.

To set notation for recursive decomposition, we use $\pi_0, \pi_1, \ldots, \pi_T$ to denote a sequence of distributions on $\mathcal{H}$, where π_0 is an uninformed prior and $\pi_T = \rho$ is the final posterior. We use $\gamma_2, \ldots, \gamma_T$ to denote a sequence of coefficients. For $t \geq 2$ we then have the recursive decomposition

$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}[L(h) - \gamma_t\mathbb{E}_{\pi_{t-1}}[L(h)]] + \gamma_t\mathbb{E}_{\pi_{t-1}}[L(h)]. \tag{2}$$

To construct $\pi_1, \ldots, \pi_T$ we split the data S into T non-overlapping subsets, $S = S_1 \cup \cdots \cup S_T$. We restrict π_t to depend on $U_t^{\mathrm{train}} = \bigcup_{s=1}^{t} S_s$ only, and we use $U_t^{\mathrm{val}} = \bigcup_{s=t}^{T} S_s$ to estimate (recursively) $\mathbb{E}_{\pi_t}[L(h)]$ (see Figures 1 and 2 in Appendix A for a graphical illustration). Note that S_t is used both for construction of π_t and for estimation of $\mathbb{E}_{\pi_t}[L(h)]$ (it is both in U_t^{train} and U_t^{val}), resulting in efficient use of the data. It is possible to use any standard PAC-Bayes bound, e.g., Theorem 1, to bound $\mathbb{E}_{\pi_1}[L(h)]$, and any PAC-Bayes-Empirical-Bernstein-style bound or the PAC-Bayes-split-kl bound to bound the excess losses $\mathbb{E}_{\pi_t}[L(h) - \gamma_t\mathbb{E}_{\pi_{t-1}}[L(h)]]$. The excess losses take more than three values, so in the next section we present a generalization of the PAC-Bayes-split-kl inequality to general discrete random variables, which may be of independent interest. The Recursive PAC-Bayes bound is presented in Section 4.

3 Split-kl and PAC-Bayes-split-kl inequalities for discrete random variables

The kl inequality is one of the tightest concentration of measure inequalities for binary random variables. Letting $\mathrm{kl}^{-1,+}(\hat{p}, \varepsilon) := \max\{p : p \in [0,1] \text{ and } \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$ denote the upper inverse of kl and $\mathrm{kl}^{-1,-}(\hat{p}, \varepsilon) := \min\{p : p \in [0,1] \text{ and } \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$ the lower inverse, it states the following.

Theorem 2 (kl Inequality (Langford, 2005, Foong et al., 2021, 2022)). *Let $Z_1, \cdots, Z_n$ be independent random variables bounded in the $[0,1]$ interval and with $\mathbb{E}[Z_i] = p$ for all i. Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ be the empirical mean. Then, for any $\delta \in (0,1)$:*

$$\mathbb{P}\left(p \geq \mathrm{kl}^{-1,+}\left(\hat{p}, \frac{1}{n}\ln\frac{1}{\delta}\right)\right) \leq \delta \qquad ; \qquad \mathbb{P}\left(p \leq \mathrm{kl}^{-1,-}\left(\hat{p}, \frac{1}{n}\ln\frac{1}{\delta}\right)\right) \leq \delta.$$

While the kl inequality is tight for binary random variables, it is loose for random variables taking more than two values due to its inability to exploit small variance. To address this shortcoming Wu and Seldin (2022) have presented the split-kl and PAC-Bayes-split-kl inequalities for ternary random variables. Ternary random variables naturally appear in a variety of applications, including analysis of excess losses, certain ways of analysing majority votes, and in learning with abstention. The bound of Wu and Seldin is based on decomposition of a ternary random variable into a pair of binary random variables and application of the kl inequality to each of them. Their decomposition yields a tight bound in the binary and ternary case, but loose otherwise. The same decomposition was used by Biggs and Guedj (2023) to derive a slight variation of the inequality, with the same limitations. We present a novel decomposition of discrete random variables into a superposition of binary random variables. Unlike the decomposition of Wu and Seldin, which only applies in the ternary case, our decomposition applies to general discrete random variables. By combining it with kl bounds for the binary elements we obtain a tight bound. The decomposition is presented formally below and illustrated graphically in Figure 3 in Appendix A.

3.1 Split-kl inequality

Let $Z \in \{b_0, \ldots, b_K\}$ be a $(K+1)$-valued random variable with $b_0 < b_1 < \cdots < b_K$. For $j \in \{1, \ldots, K\}$ define $Z_{|j} = \mathbb{1}(Z \geq b_j)$ and $\alpha_j = b_j - b_{j-1}$. Then $Z = b_0 + \sum_{j=1}^{K} \alpha_j Z_{|j}$. For a sequence $Z_1, \ldots, Z_n$ of $(K+1)$-valued random variables with the same support, let $Z_{i|j} = \mathbb{1}(Z_i \geq b_j)$ denote the elements of binary decomposition of Z_i.

Theorem 3 (Split-kl inequality for discrete random variables). *Let $Z_1, \ldots, Z_n$ be i.i.d. random variables taking values in $\{b_0, \ldots, b_K\}$ with $\mathbb{E}[Z_i] = p$ for all i. Let $\hat{p}_{|j} = \frac{1}{n}\sum_{i=1}^{n} Z_{i|j}$. Then for any $\delta \in (0,1)$:*

$$\mathbb{P}\left(p \geq b_0 + \sum_{j=1}^{K} \alpha_j \,\mathrm{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n}\ln\frac{K}{\delta}\right)\right) \leq \delta.$$

Proof. Let $p_{|j} = \mathbb{E}\left[\hat{p}_{|j}\right]$, then $p = b_0 + \sum_{j=1}^{K} \alpha_j p_{|j}$ and

$$\mathbb{P}\left(p \geq b_0 + \sum_{j=1}^{K} \alpha_j \,\mathrm{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n}\ln\frac{K}{\delta}\right)\right) \leq \mathbb{P}\left(\exists j : p_{|j} \geq \mathrm{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n}\ln\frac{K}{\delta}\right)\right) \leq \delta,$$

where the first inequality is by the decomposition of p and the second inequality is by the union bound and Theorem 2. □

3.2 PAC-Bayes-Split-kl inequality

Let $f : \mathcal{H} \times \mathcal{Z} \to \{b_0, \ldots, b_K\}$ be a $(K+1)$-valued loss function. (To connect it to the earlier examples, in the binary prediction case we would have $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with elements $Z = (X, Y)$ and $f(h, Z) = \ell(h(X), Y)$, but we will need a more general space $\mathcal{Z}$ later.) For $j \in \{1, \ldots, K\}$ let $f_{|j}(\cdot,\cdot) = \mathbb{1}(f(\cdot,\cdot) \geq b_j)$. Let $\mathcal{D}_Z$ be an unknown distribution on $\mathcal{Z}$. For $h \in \mathcal{H}$ let $F(h) = \mathbb{E}_{\mathcal{D}_Z}[f(h, Z)]$ and $F_{|j}(h) = \mathbb{E}_{\mathcal{D}_Z}[f_{|j}(h, Z)]$. Let $S = \{Z_1, \ldots, Z_n\}$ be an i.i.d. sample according to $\mathcal{D}_Z$ and $\hat{F}_{|j}(h, S) = \frac{1}{n}\sum_{i=1}^{n} f_{|j}(h, Z_i)$.

Theorem 4 (PAC-Bayes-Split-kl Inequality). *For any distribution π on $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$:*

$$\mathbb{P}\left(\exists \rho \in \mathcal{P} : \mathbb{E}_\rho[F(h)] \geq b_0 + \sum_{j=1}^{K} \alpha_j \,\mathrm{kl}^{-1,+}\left(\mathbb{E}_\rho[\hat{F}_{|j}(h, S)], \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2K\sqrt{n}}{\delta}}{n}\right)\right) \leq \delta,$$

where $\mathcal{P}$ is the set of all possible probability distributions on $\mathcal{H}$ that can depend on S.

https://doi.org/10.52202/079017-0570

Proof. We have $f(\cdot,\cdot) = b_0 + \sum_{j=1}^K \alpha_j f_{|j}(\cdot,\cdot)$ and $F(h) = b_0 + \sum_{j=1}^K \alpha_j F_{|j}(h)$. Therefore,

$$\mathbb{P}\left(\exists \rho \in \mathcal{P} : \mathbb{E}_\rho[F(h)] \geq b_0 + \sum_{j=1}^K \alpha_j \,\mathrm{kl}^{-1,+}\left(\mathbb{E}_\rho[\hat{F}_{|j}(h,S)], \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2K\sqrt{n}}{\delta}}{n}\right)\right)$$
$$\leq \mathbb{P}\left(\exists \rho \in \mathcal{P} \text{ and } \exists j : \mathbb{E}_\rho[F_{|j}(h)] \geq \mathrm{kl}^{-1,+}\left(\mathbb{E}_\rho[\hat{F}_{|j}(h,S)], \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2K\sqrt{n}}{\delta}}{n}\right)\right) \leq \delta,$$

where the first inequality is by the decomposition of F and the second inequality is by the union bound and application of Theorem 1 to $F_{|j}$ (note that $f_{|j}$ is a zero-one loss function). □

4 Recursive PAC-Bayes bound

Now we derive a Recursive PAC-Bayes bound based on the loss decomposition in equation (2). We aim to bound $\mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]]$, which we denote by

$$F_{\gamma_t,\pi_{t-1}}(h) = L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')] = \mathbb{E}_{\mathcal{D}\times\pi_{t-1}}[\ell(h(X),Y) - \gamma_t \ell(h'(X),Y)],$$

where $\mathcal{D}\times\pi_{t-1}$ is a product distribution on $\mathcal{X}\times\mathcal{Y}\times\mathcal{H}$ and $h' \in \mathcal{H}$ is sampled according to π_{t-1}. We further define

$$f_{\gamma_t}(h,(X,Y,h')) = \ell(h(X),Y) - \gamma_t \ell(h'(X),Y) \in \{-\gamma_t, 0, 1-\gamma_t, 1\},$$

then $F_{\gamma_t,\pi_{t-1}}(h) = \mathbb{E}_{\mathcal{D}\times\pi_{t-1}}[f_{\gamma_t}(h,(X,Y,h'))]$. In order to apply Theorem 4, we represent f_{γ_t} as a superposition of binary functions. For this purpose we let $\{b_{t|0}, b_{t|1}, b_{t|2}, b_{t|3}\} = \{-\gamma_t, 0, 1-\gamma_t, 1\}$ and define $f_{\gamma_t|j}(h,(X,Y,h')) = \mathbb{1}\big(f_{\gamma_t}(h,(X,Y,h')) \geq b_{t|j}\big)$. We let $F_{\gamma_t,\pi_{t-1}|j}(h) = \mathbb{E}_{\mathcal{D}\times\pi_{t-1}}[f_{\gamma_t|j}(h,(X,Y,h'))]$, then $F_{\gamma_t,\pi_{t-1}}(h) = -\gamma_t + \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) F_{\gamma_t,\pi_{t-1}|j}(h)$.

Now we construct an empirical estimate of $F_{\gamma_t,\pi_{t-1}|j}(h)$. We first let $\hat{\pi}_{t-1} = \{h_1^{\pi_{t-1}}, h_2^{\pi_{t-1}}, \dots\}$ be a sequence of prediction rules sampled independently according to π_{t-1}. We define $U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1} = \{(X_i, Y_i, h_i^{\pi_{t-1}}) : (X_i,Y_i) \in U_t^{\mathrm{val}}\}$. In words, for every sample $(X_i,Y_i) \in U_t^{\mathrm{val}}$ we sample a prediction rule $h_i^{\pi_{t-1}}$ according to π_{t-1} and put the triplet $(X_i,Y_i,h_i^{\pi_{t-1}})$ in $U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1}$. The triplets $(X_i,Y_i,h_i^{\pi_{t-1}})$ correspond to the random variables Z in Theorem 4. We note that $|U_t^{\mathrm{val}}| = |U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1}|$, and we let $n_t^{\mathrm{val}} = |U_t^{\mathrm{val}}|$. We define the empirical estimate of $F_{\gamma_t,\pi_{t-1}|j}(h)$ as $\hat{F}_{\gamma_t|j}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1}) = \frac{1}{n_t^{\mathrm{val}}} \sum_{(X,Y,h') \in U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1}} f_{\gamma_t|j}(h,(X,Y,h'))$. Note that $\mathbb{E}_{\mathcal{D}\times\pi_{t-1}}[\hat{F}_{\gamma_t|j}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1})] = F_{\gamma_t,\pi_{t-1}|j}(h)$, therefore, we can use Theorem 4 to bound $\mathbb{E}_{\pi_t}[F_{\gamma_t,\pi_{t-1}}(h)]$ using its empirical estimates. We are now ready to state the bound.

Theorem 5 (Recursive PAC-Bayes Bound)**.** *Let $S = S_1 \cup \dots \cup S_T$ be an i.i.d. sample split in an arbitrary way into T non-overlapping subsamples, and let $U_t^{\mathrm{train}} = \bigcup_{s=1}^t S_s$ and $U_t^{\mathrm{val}} = \bigcup_{s=t}^T S_s$. Let $n_t^{\mathrm{val}} = |U_t^{\mathrm{val}}|$. Let $\pi_0^*, \pi_1^*, \dots, \pi_T^*$ be a sequence of distributions on $\mathcal{H}$, where π_t^* is allowed to depend on U_t^{train}, but not the rest of the data. Let $\gamma_2, \dots, \gamma_T$ be a sequence of coefficients, where γ_t is allowed to depend on U_{t-1}^{train}, but not the rest of the data. For $t \in \{1,\dots,T\}$ let $\mathcal{P}_t$ be a set of distributions on $\mathcal{H}$, which are allowed to depend on U_t^{train}. Then for any $\delta \in (0,1)$:*

$$\mathbb{P}(\exists t \in \{1,\dots,T\} \text{ and } \pi_t \in \mathcal{P}_t : \mathbb{E}_{\pi_t}[L(h)] \geq \mathrm{B}_t(\pi_t)) \leq \delta,$$

where $\mathrm{B}_t(\pi_t)$ is a PAC-Bayes bound on $\mathbb{E}_{\pi_t}[L(h)]$ defined recursively as follows. For $t = 1$

$$\mathrm{B}_1(\pi_1) = \mathrm{kl}^{-1,+}\left(\mathbb{E}_{\pi_1}[\hat{L}(h,S)], \frac{\mathrm{KL}(\pi_1\|\pi_0^*) + \ln\frac{2T\sqrt{n}}{\delta}}{n}\right).$$

For $t \geq 2$ we let $\mathcal{E}_t(\pi_t,\gamma_t)$ denote a PAC-Bayes bound on $\mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}^*}[L(h')]]$ given by

$$\mathcal{E}_t(\pi_t,\gamma_t) = -\gamma_t + \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) \,\mathrm{kl}^{-1,+}\left(\mathbb{E}_{\pi_t}\left[\hat{F}_{\gamma_t|j}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1}^*)\right], \frac{\mathrm{KL}(\pi_t\|\pi_{t-1}^*) + \ln\frac{6T\sqrt{n_t^{\mathrm{val}}}}{\delta}}{n_t^{\mathrm{val}}}\right)$$

and then

$$\mathrm{B}_t(\pi_t) = \mathcal{E}_t(\pi_t,\gamma_t) + \gamma_t \,\mathrm{B}_{t-1}(\pi_{t-1}^*). \tag{3}$$

Proof. By Theorem 1 we have $\mathbb{P}(\exists \pi_1 \in \mathcal{P}_1 : \mathbb{E}_{\pi_1}[L(h)] \geq B_1(\pi_1)) \leq \frac{\delta}{T}$. Further, by Theorem 4 for $t \in \{2, \ldots, T\}$ we have $\mathbb{P}\Big(\exists \pi_t \in \mathcal{P}_t : \mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi^*_{t-1}}[L(h')]] \geq \mathcal{E}_t(\pi_t, \gamma_t)\Big) \leq \frac{\delta}{T}$. The theorem follows by a union bound and the recursive decomposition of the loss (2). □

Discussion

- Note that $\pi^*_1, \ldots, \pi^*_T$ can be constructed sequentially, but π^*_t can only be constructed based on the data in U^{train}_t, meaning that in the construction of π^*_t we can only rely on $\mathbb{E}_{\pi_t}\left[\hat{F}_{\gamma_t|j}(h, S_t \circ \hat{\pi}_{t-1})\right]$, but not on $\mathbb{E}_{\pi_t}\left[\hat{F}_{\gamma_t|j}(h, U^{\text{val}}_t \circ \hat{\pi}_{t-1})\right]$. Also note that S_t is part of both U^{train}_t and U^{val}_t (see Figure 1 in Appendix A for a graphical illustration). In other words, when we evaluate the bounds we can use additional data. And even though the additional data can only be used in the evaluation stage, we can still use the knowledge that we will get more data for evaluation when we construct π^*_t. For example, we can take

$$\pi^*_1 = \arg\min_{\pi} \mathrm{kl}^{-1,+}\left(\mathbb{E}_{\pi}[\hat{L}(h, S_1)], \frac{\mathrm{KL}(\pi \| \pi^*_0) + \ln \frac{2T\sqrt{n}}{\delta}}{n}\right) \tag{4}$$

and for $t \geq 2$

$$\pi^*_t = \arg\min_{\pi} \sum_{j=1}^{3} (b_{t|j} - b_{t|j-1})\, \mathrm{kl}^{-1,+}\left(\mathbb{E}_{\pi}\left[\hat{F}_{\gamma_t|j}(h, S_t \circ \hat{\pi}^*_{t-1})\right], \frac{\mathrm{KL}(\pi \| \pi^*_{t-1}) + \ln \frac{6T\sqrt{n^{\text{val}}_t}}{\delta}}{n^{\text{val}}_t}\right). \tag{5}$$

The empirical losses above are calculated on S_t corresponding to π^*_t, but the sample sizes n^{val}_t correspond to the size of the validation set U^{val}_t rather than the size of S_t. This allows to be more aggressive in deviating with π^*_t from π^*_{t-1} by sustaining larger $\mathrm{KL}(\pi^*_t \| \pi^*_{t-1})$ terms.

- Similarly, $\gamma_2, \ldots, \gamma_T$ can also be constructed sequentially, as long as γ_t only depends on U^{train}_{t-1} (otherwise $\hat{F}_{\gamma_t|j}(h, S_t \circ \hat{\pi}^*_{t-1})$ becomes a biased estimate of $F_{\gamma_t, \pi^*_{t-1}|j}(h)$).
- We naturally want to have improvement over recursion steps, meaning $\mathrm{B}_t(\pi^*_t) < \mathrm{B}_{t-1}(\pi^*_{t-1})$. Plugging this into (3), we obtain $\mathcal{E}(\pi^*_t, \gamma_t) + \gamma_t B_{t-1}(\pi^*_{t-1}) < B_{t-1}(\pi^*_{t-1})$, which implies that we want γ_t to be sufficiently small to satisfy $\gamma_t < 1 - \frac{\mathcal{E}_t(\pi^*_t, \gamma_t)}{B_{t-1}(\pi^*_{t-1})}$. At the same time, γ_t should be non-negative. Therefore, improvement over recursion steps can only be maintained as long as $\mathcal{E}_t(\pi^*_t, \gamma_t) < B_{t-1}(\pi^*_{t-1})$. We note that $\gamma_t\, \mathrm{B}_{t-1}(\pi^*_{t-1})$ term in (3) is linearly increasing in γ_t, whereas $\mathcal{E}(\pi^*_t, \gamma_t)$ is decreasing in γ_t. The value of γ_t that minimizes the trade-off depends on the data. Even though it is not allowed to use U^{val}_t for tuning γ_t, it is possible to take a grid of values of γ_t and a union bound over the grid, and then select the best value from the grid based the value of the bound evaluated on U^{val}_t.

5 Experiments

In this section, we provide an empirical comparison of our Recursive PAC-Bayes (RPB) procedure to the following prior work: i) Uninformed priors (Uninformed), (Dziugaite and Roy, 2017); ii) Data-informed priors (Informed) (Ambroladze et al., 2007, Pérez-Ortiz et al., 2021); iii) Data-informed prior + excess loss (Informed + Excess) (Mhammedi et al., 2019, Wu and Seldin, 2022). All the experiments were run on a laptop. The source code for replicating the experiments is available at Github[1].

We start with describing the details of the optimization procedure, and then present the results.

5.1 Details of the optimization and evaluation procedure

We constructed $\pi^*_1, \ldots, \pi^*_T$ sequentially using the optimization objective, (4) for π^*_1 and (5) for π^*_2 to π^*_T, and computed the bound using the recursive procedure in Theorem 5. There are a few technical details concerning convexity of the optimization procedure and infinite size of the set of prediction rules $\mathcal{H}$ that we address next.

[1] https://github.com/pyijiezhang/rpb

https://doi.org/10.52202/079017-0570

5.1.1 Convexification of the loss functions

The functions $f_{\gamma_t|j}(h,(X,Y,h'))$ defined in Section 4 are non-convex and non-differentiable: $f_{\gamma_t|j}(h,(X,Y,h')) = \mathbb{1}\big(f_{\gamma_t}(h,(X,Y,h')) \geq b_{t|j}\big) = \mathbb{1}\big(\ell(h(X),Y) - \gamma_t \ell(h'(X),Y) \geq b_{t|j}\big)$. In order to facilitate optimization, we approximate the external indicator function $\mathbb{1}(z \geq z_0)$ by a sigmoid function $\omega(z;c_1,z_0) = (1+\exp(c_1(z-z_0)))^{-1}$ with a fixed parameter $c_1 > 0$ specified in Appendix B.3.

Furthermore, since the zero-one loss $\ell(h(X),Y)$ is also non-differentiable, we adopt the cross-entropy loss, as in most modern training procedures (Pérez-Ortiz et al., 2021). Specifically, for a k-class classification problem, let $h : \mathcal{X} \to \mathbb{R}^k$ represent the function implemented by the neural network, assigning each class a real value. Let $u = h(X)$ be the assignment, with u_i being the i-th value of the vector. To convert this real-valued vector into a probability distribution over classes, we apply the softmax function $\sigma : \mathbb{R}^k \to \Delta^{k-1}$, where $\sigma(u)_i = \exp(c_2 u_i)/\sum_j \exp(c_2 u_j)$ for some $c_2 > 0$ for each entry. The cross-entropy loss $\ell^{\text{ce}} : \mathbb{R}^k \times [k] \to \mathbb{R}$ is defined by $\ell^{\text{ce}}(u,Y) = -\log(\sigma(u)_Y)$. However, since this loss is unbounded, whereas the PAC-Bayes-kl bound requires losses within $[0,1]$, we enforce a $[0,1]$-valued cross-entropy loss by mixing the output distribution with a uniform distribution $\sigma(u)$, i.e., $\tilde{\sigma}(u)_i = (1-p_{\min})\sigma(u)_i + p_{\min}/k$ for all $i \in [k]$ and for some $p_{\min} > 0$, and then rescaling it to $[0,1]$ by taking $\tilde{\ell}^{\text{ce}}(u,Y) = -\log(\tilde{\sigma}(u)_Y)/\log(k/p_{\min})$.

We emphasize that in the evaluation of the bound (using Theorem 5), we directly compute the zero-one loss and the $f_{\gamma_t|j}$ functions without employing the approximations.

5.1.2 Relaxation of the PAC-Bayes-kl bound

The PAC-Bayes-kl bound is often criticized for being unfriendly to optimization (Rodríguez-Gálvez et al., 2024). Therefore, several relaxations have been proposed, including the PAC-Bayes-classic bound (McAllester, 1999), the PAC-Bayes-λ bound (Thiemann et al., 2017), and the PAC-Bayes-quadratic bound (Rivasplata et al., 2019, Pérez-Ortiz et al., 2021), among others. In our optimization we have adopted the bound of McAllester (1999) instead of the kl-based bounds in Equation (5).

We again emphasize that in the evaluation of the bound we used the kl-based bounds in Theorem 5.

5.1.3 Estimation of $\mathbb{E}_\pi[\cdot]$

Due to the infinite size of $\mathcal{H}$ and lack of a closed-form expression for $\mathbb{E}_{\pi_1}[\hat{L}(h,S)]$ and $\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t|j}(h,U_t^{\text{val}} \circ \hat{\pi}_{t-1}^*)]$ appearing in Theorem 5, we approximate them by sampling (Pérez-Ortiz et al., 2021). For optimization, we sample one classifier for each mini-batch during stochastic gradient descent. For evaluation, we sample one classifier for each data in the corresponding evaluation dataset. Due to approximation of the empirical quantities the final bound in Theorem 5 requires an additional concentration bound. (We note that the extra bound is only required for computation of the final bound, but not for optimization of $\hat{\pi}_t^*$.) Specifically, let $\hat{\pi}_t^* = \{h_1^{\pi_t}, h_2^{\pi_t}, \ldots, h_m^{\pi_t}\}$ be m prediction rules sampled independently according to π_t. Then for any function $f(h)$ taking values in $[0,1]$ (which is the case for $\hat{L}(h,S)$ and $\hat{F}_{\gamma_t|j}(h,U_t^{\text{val}} \circ \hat{\pi}_{t-1}^*)$) and $\delta' \in (0,1)$ we have

$$\mathbb{P}\left(\mathbb{E}_{\pi_t^*}[f(h)] \geq \text{kl}^{-1,+}\left(\frac{1}{m}\sum_{i=1}^m f(h_i^{\pi_t^*}), \frac{1}{m}\log\frac{1}{\delta'}\right)\right) \leq \delta'.$$

It is worth noting that $\mathbb{E}_{\pi_t^*}[f(h)]$ is evaluated for a fixed π_t^*, meaning that there is no selection involved, and therefore no KL term appears in the bound above. We, of course, take a union bound over all the quantities being estimated.

5.2 Experimental results

We evaluated our approach and compared it to prior work using multi-class classification tasks on MNIST (LeCun and Cortes, 2010) and Fashion MNIST (Xiao et al., 2017) datasets, both with 60000 training data. The experimental setup was based on the work of Dziugaite and Roy (2017) and Pérez-Ortiz et al. (2021). Similar to them we used Gaussian distributions for all the priors and posteriors, modeled by probabilistic neural networks. Technical details are provided in B.

The empirical evaluation is presented in Table 1. For the Uninformed approach, we trained and evaluated the bound using the entire training dataset directly. For the other two baseline methods, Informed and Informed + Excess Loss, we used half of the training data to train the informed prior and an ERM h^* for the excess loss, and the other half to learn the posterior. For our Recursive PAC-Bayes, we chose $\gamma_t = 1/2$ for all t, and conducted experiments with $T = 2, 4, 6, 8$ to study the impact of recursion depth. (Each value of T corresponded to a separate run of the algorithm and a separate evaluation of the bound, i.e., they should not be seen as successive refinements.) We applied a geometric data split. Specifically, for $T = 2$ the split was (30000, 30000) points; for $T = 4$, it was (7500, 7500, 15000, 30000); for $T = 6$ it was (1875, 1875, 3750, 7500, 15000, 30000); and for $T = 8$, it was (469, 469, 937, 1875, 3750, 7500, 15000, 30000). This approach allowed the early recursion steps, which had fewer data points, to efficiently learn the prior, while preserving enough data for fine-tuning in the later steps. Note that with this approach the value of $n_t^{\mathrm{val}} = |U_t^{\mathrm{val}}| = \sum_{s=t}^{T} |S_s|$, which is in the denominator of the bounds in Theorem 5, is at least $\frac{n}{2}$.

Table 1: Comparison of the classification loss of the final posterior ρ on the entire training data, $\mathbb{E}_\rho[\hat{L}(h, S)]$ (Train 0-1), and on the testing data, $\mathbb{E}_\rho[\hat{L}(h, S^{\text{test}})]$ (Test 0-1), and the corresponding bounds for each method on MNIST and Fashion MNIST. We report the mean and one standard deviation over 5 repetitions. "Unif." abbreviates the Uniform approach, "Inf." the Informed, "Inf. + Ex." the Informed + Excess Loss, and "RPB" the Recursive PAC-Bayes.

	MNIST			Fashion MNIST		
	Train 0-1	Test 0-1	Bound	Train 0-1	Test 0-1	Bound
Uninf.	.343 (2e-3)	.335 (3e-3)	.457 (2e-3)	.382 (2e-3)	.384 (2e-3)	.464 (2e-3)
Inf.	.377 (8e-4)	.371 (6e-3)	.408 (9e-4)	.412 (1e-3)	.413 (6e-3)	.440 (1e-3)
Inf. + Ex.	.157 (2e-3)	.151 (3e-3)	.192 (2e-3)	.280 (4e-3)	.285 (5e-3)	.342 (6e-3)
RPB $T = 2$	.143 (2e-3)	.139 (3e-3)	.321 (3e-3)	.257 (3e-3)	.266 (5e-3)	.404 (3e-3)
RPB $T = 4$	.112 (1e-3)	.109 (1e-3)	.203 (8e-4)	.203 (2e-3)	.213 (3e-3)	.293 (1e-3)
RPB $T = 6$	.103 (1e-3)	.101 (1e-3)	.166 (1e-3)	.186 (4e-4)	.198 (1e-3)	.255 (1e-3)
RPB $T = 8$	**.101 (1e-3)**	**.097 (2e-3)**	**.158 (2e-3)**	**.181 (1e-3)**	**.192 (3e-3)**	**.242 (1e-3)**

Table 1 shows that even with only $T = 2$, which corresponds to the data split used in the Informed and the Informed + Excess Loss approaches, RPB achieves better test performance than prior work. As the recursion deepens, further improvements in both the test error and the bound are observed. We note that while the bound for $T = 2$ is looser compared to the Informed + Excess Loss method, deeper recursion yields bounds that are tighter. Overall, deep recursion provides substantial improvements in the bound and the test error relative to prior work.

Tables 2 and 3 provide a glimpse into the training progress of RPB with $T = 8$ by showing the evolution of the key quantities along the recursive process. Similar tables for other values of T are provided in Appendix B.4, along with training details for other methods. The tables show an impressive reduction of the KL term and significant improvement of the bound as the recursion proceeds, demonstrating effectiveness of the approach.

Table 2: Insight into the training process of the Recursive PAC-Bayes for $T = 8$ on MNIST. The table shows the evolution of $\mathcal{E}_t(\pi_t^*, \gamma_t)$, $B_t(\pi_t^*)$, and other quantities as the training progresses with t. We define $\hat{F}_{\gamma_t}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1}) = -\gamma_t + \sum_{j=1}^{3} (b_{t|j} - b_{t|j-1}) \hat{F}_{\gamma_t|j}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1})$.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\mathrm{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\mathrm{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.009 (3e-4)		.612 (9e-3)	.532 (.011)
2	59532	-0.046 (4e-3)	.031 (1e-3)	.114 (2e-3)	.421 (5e-3)	.215 (7e-3)
3	59063	.040 (3e-3)	.013 (9e-4)	.125 (3e-3)	.336 (2e-3)	.146 (3e-3)
4	58125	.049 (1e-3)	.005 (3e-4)	.099 (1e-3)	.267 (7e-4)	.120 (2e-3)
5	56250	.052 (4e-4)	.002 (1e-4)	.083 (1e-3)	.217 (1e-3)	.111 (2e-3)
6	52500	.051 (1e-3)	.001 (4e-5)	.076 (1e-3)	.185 (1e-3)	.104 (2e-3)
7	45000	.050 (1e-3)	8e-4 (6e-5)	.073 (1e-3)	.166 (1e-3)	.099 (1e-3)
8	30000	.050 (1e-3)	6e-4 (4e-5)	.074 (1e-3)	.158 (2e-3)	.097 (2e-3)

https://doi.org/10.52202/079017-0570

Table 3: Insight into the training process of the Recursive PAC-Bayes for $T = 8$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\mathrm{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\mathrm{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\mathrm{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.003 (7e-5)		.733 (7e-3)	.686 (8e-3)
2	59532	-0.043 (8e-3)	.023 (6e-4)	.104 (9e-3)	.470 (8e-3)	.309 (7e-3)
3	59063	.083 (4e-3)	.008 (3e-4)	.161 (3e-3)	.396 (4e-3)	.242 (1e-3)
4	58125	.090 (3e-3)	.004 (5e-4)	.142 (4e-3)	.341 (5e-4)	.216 (5e-3)
5	56250	.093 (3e-3)	.001 (2e-4)	.126 (3e-3)	.297 (4e-3)	.204 (4e-3)
6	52500	.090 (1e-3)	6e-4 (6e-5)	.117 (1e-3)	.265 (1e-3)	.195 (3e-3)
7	45000	.090 (1e-3)	4e-4 (2e-5)	.115 (1e-3)	.248 (1e-3)	.195 (5e-4)
8	30000	.090 (1e-3)	4e-4 (1e-5)	.117 (1e-3)	.242 (1e-3)	.192 (3e-3)

6 Discussion

We have presented the first PAC-Bayesian bound that supports sequential prior updates and preserves confidence information on the prior. The work closes a long-standing gap between Bayesian and Frequentist learning by making sequential data processing and sequential updates of prior knowledge meaningful and beneficial in the frequentist framework, as it has always been in the Bayesian framework. We have shown that apart from theoretical beauty the approach is highly beneficial in practice.

The Recursive PAC-Bayes framework is extremely rich and powerful, and leads to numerous directions for future research, some of which we briefly sketch next.

- The decomposition in (2) applies to any loss function, including unbounded losses. It would be interesting to find additional applications to it.
- While we have restricted ourselves to the zero-one loss function to illustrate the use of PAC-Bayes-split-kl, the results can be directly generalized to any bounded loss function by replacing PAC-Bayes-split-kl with PAC-Bayes-Empirical-Bernstein or PAC-Bayes-Unexpected-Bernstein, and deriving the corresponding analogue of Theorem 5 (which is straightforward).
- We have shown that the bound works well with geometric split of the data, but there are many other ways to split the data which could be studied.
- There is also a lot of space for experimentation with optimization of γ_t.
- It would be interesting to study how the bound will perform in sequential learning settings, where the data arrives sequentially, and thus the partition is dictated externally.
- There are many interesting research directions from the computational perspective. We note that for base models with linear computational complexity (e.g., neural networks) the overhead of recursion is relatively small and optimization time of Recursive PAC-Bayes is comparable to processing all data at once or in two chunks (as in data-dependent priors). For base models with superlinear computational complexity (e.g., kernel SVMs) sequential training of several small models in the recursion may actually be cheaper than training a big model based on all the data. Moreover, since the bound in Theorem 5 holds for any sequence of distributions $\pi_0^*, \pi_1^*, \ldots, \pi_T^*$, the optimization in equation (5) is allowed to be approximate. Considering that the improvement of the bounds and the test loss relative to prior work was very significant, there is space to look at the trade-off between statistical power and computational complexity. Namely, it may potentially be possible to relax the approximation of $\arg\min$ in equation (5) to gain computational speed-up at the cost of only a small compromise on the bounds and test losses.
- We note that it is possible to start the recursion at π_0. Namely, it is possible to use, for example, Theorem 2 to bound $\mathbb{E}_{\pi_0}[L(h)]$ using all the data, and apply the recursive decomposition (2) starting from π_1. Whether this would yield an advantage relative to starting the recursion at π_1, as we did, remains to be studied.

Acknowledgments and Disclosure of Funding

YW acknowledges support from the Novo Nordisk Foundation, grant number NNF21OC0070621. YZ acknowledge Ph.D. funding from Novo Nordisk A/S. BECA acknowledges funding from the ANR grant project BACKUP ANR-23-CE40-0018-01.

References

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *NeurIPS*, 2007.

Felix Biggs and Benjamin Guedj. Tighter PAC-Bayes generalisation bounds by leveraging example difficulty. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *Journal of Machine Learning Research*, 24(372), 2023.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd edition, 2006.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Andrew Foong, Wessel Bruinsma, David Burt, and Richard Turner. How tight can pac-bayes be in the small data regime? In *NeurIPS*, 2021.

Andrew Y. K. Foong, Wessel P. Bruinsma, and David R. Burt. A note on the chernoff bound for random variables in the unit interval. *arXiv preprint arXiv.2205.07880*, 2022.

Maxime Haddouche and Benjamin Guedj. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research Journal*, 2023.

Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes bounds through coin-betting. In *Proceedings of the Conference on Learning Theory (COLT)*, 2023.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6, 2005.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Andreas Maurer. A note on the PAC-Bayesian theorem. arXiv preprint cs/0411099, 2004.

David McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Conference on Learning Theory (COLT)*, 1998.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In *NeurIPS*, 2019.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 2021.

Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvari. Pac-bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019.

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110), 2024.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3, 2002.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the Conference on Learning Theory (COLT)*, 1997.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2017.

 https://doi.org/10.52202/079017-0570

Ilya Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein inequality. In *NeurIPS*, 2013.

Yi-Shan Wu and Yevgeny Seldin. Split-kl and PAC-Bayes-split-kl inequalities for ternary random variables. In *NeurIPS*, 2022.

Yi-Shan Wu, Andres Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Chebyshev-cantelli pac-bayes-bennett inequality for the weighted majority vote. In *NeurIPS*, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A Illustrations

In this appendix we provide graphical illustrations of the basic concepts presented in the paper.

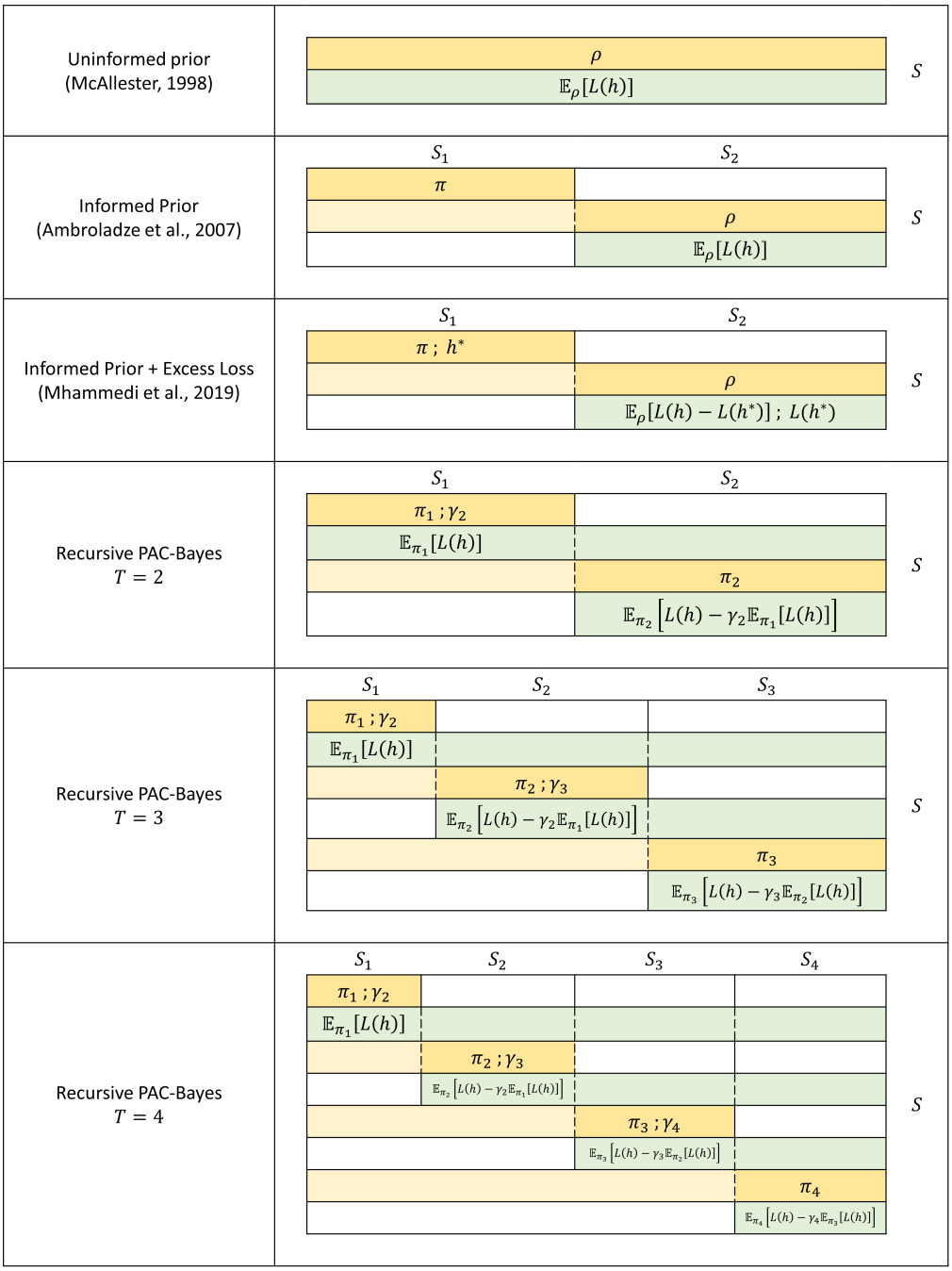

Figure 1: **Evolution of PAC-Bayes.** The figure shows how data are used by different PAC-Bayes approaches. Dark yellow shows data used directly for optimization of the indicated quantities. Light yellow shows data involved indirectly through dependence on the prior. Light green shows data used for estimation of the indicated quantities. In Recursive PAC-Bayes data are released and used sequentially chunk-by-chunk, as indicated by the dashed lines. For example, in the $T = 4$ case $\mathbb{E}_{\pi_1}[L(h)]$ is first evaluated on S_1 to construct π_1 and γ_2, then in the first recursion step on $S_1 \cup S_2$, in the second step on $S_1 \cup S_2 \cup S_3$, and in the last step on all S.

https://doi.org/10.52202/079017-0570

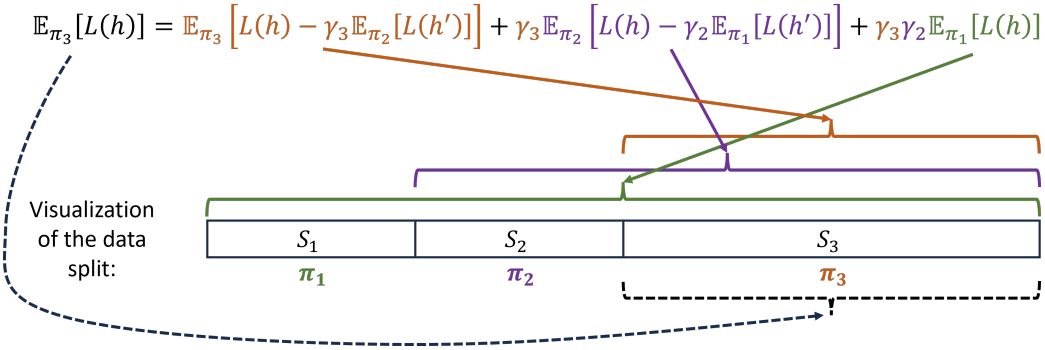

Figure 2: **Recursive Decomposition into Three Terms.** The figure illustrates recursive decomposition of $\mathbb{E}_{\pi_3}[L(h)]$ into three terms based on equation (2), and a geometric data split, as used in our experiments. The bottom line illustrates which data are used for construction of which distribution: S_1 for π_1; S_2 for π_2; and S_3 for π_3. The brackets above the data show which data are used for computing PAC-Bayes bounds for which term: $S_1 \cup S_2 \cup S_3$ for $\mathbb{E}_{\pi_1}[L(h)]$; $S_2 \cup S_3$ for $\mathbb{E}_{\pi_2}[L(h) - \gamma_2\mathbb{E}_{\pi_1}[L(h')]]$; and S_3 for $\mathbb{E}_{\pi_3}[L(h) - \gamma_3\mathbb{E}_{\pi_2}[L(h')]]$. Note that a direct computation of a PAC-Bayes bound on $\mathbb{E}_{\pi_3}[L(h)]$ would have only allowed to use the data in S_3, as shown by the black dashed line. The figure illustrates that recursive decomposition provides more efficient use of the data. We also note that initially we start with poor priors, and so the $\mathrm{KL}(\pi_t\|\pi_{t-1})$ term for small t is expected to be large, but this is compensated by a small multiplicative factor $\prod_{i=t+1}^{T}\gamma_i$ and availability of a lot of data $\bigcup_{i=t}^{T} S_i$ for computing the PAC-Bayes bound. For example, $\mathbb{E}_{\pi_1}[L(h)]$ is multiplied by $\gamma_3\gamma_2$ and we can use all the data for computing a PAC-Bayes bound on this term. By the time we reach higher t, the priors π_{t-1} get better, and the $\mathrm{KL}(\pi_t\|\pi_{t-1})$ term in the bounds gets much smaller, and additionally the bounds benefit from the small variance of the excess loss. With geometric split of the data, we use little data to quickly move π_t to a good region, and then we still have enough data for a good estimation of the later terms, like $\mathbb{E}_{\pi_3}[L(h) - \gamma_3\mathbb{E}_{\pi_2}[L(h')]]$.

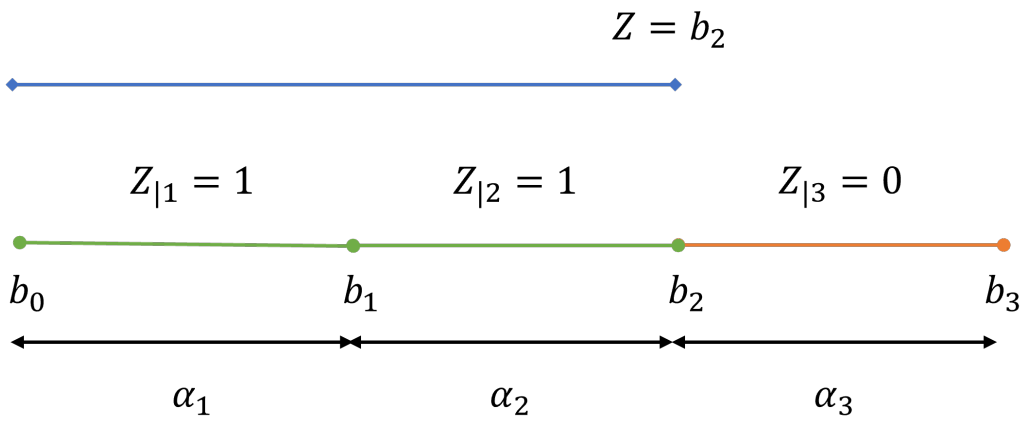

Figure 3: **Decomposition of a discrete random variable into a superposition of binary random variables.** The figure illustrates a decomposition of a discrete random variable Z with domain of four values $b_0 < b_1 < b_2 < b_3$ into a superposition of three binary random variables, $Z = b_0 + \sum_{j=1}^{3}\alpha_j Z_{|j}$. A way to think about the decomposition is to compare it to a progress bar. In the illustration Z takes value b_2, and so the random variables $Z_{|1}$ and $Z_{|2}$ corresponding to the first two segments "light up" (take value 1), whereas the random variable $Z_{|3}$ corresponding to the last segment remains "turned off" (takes value 0). The value of Z equals the sum of the lengths α_j of the "lighted up" segments.

https://doi.org/10.52202/079017-0570

B Experimental details

In this section, we provide the details of the datasets in Appendix B.1, our neural network architectures in Appendix B.2, and other details in Appendix B.3. We provide further statistics for all the methods on both datasets in Appendix B.4.

B.1 Datasets

We perform our evaluation on two datasets, MNIST (LeCun and Cortes, 2010) and Fashion MNIST (Xiao et al., 2017). We will introduce these two datasets in the following.

B.1.1 MNIST

The MNIST (Modified National Institute of Standards and Technology) dataset is one of the most renowned and widely used datasets in the field of machine learning, particularly for training and testing in the domain of image processing and computer vision. It consists of a large collection of handwritten digit images, spanning the numbers 0 through 9.

The MNIST dataset comprises a total of 70,000 grayscale images of handwritten digits, where the training set has 60,000 images and the test set has 10,000 images. Each image in the dataset is 28x28 pixels, resulting in a total of 784 pixels per image. The images are in grayscale, with pixel values ranging from 0 (black) to 255 (white). Each image is associated with a label from 0 to 9, indicating the digit that the image represents. The images are typically stored in a single flattened array of 784 elements, although they can also be represented in a 28x28 matrix format.

B.1.2 Fashion MNIST

The Fashion MNIST dataset is a contemporary alternative to the traditional MNIST dataset, created to provide a more challenging benchmark for machine learning algorithms. It consists of images of various clothing items and accessories, offering a more complex and varied dataset for image classification tasks.

The Fashion MNIST dataset contains a total of 70,000 grayscale images, where the training set has 60,000 images and the test set has 10,000 images. Each image in the dataset is 28x28 pixels, resulting in a total of 784 pixels per image. The images are in grayscale, with pixel values ranging from 0 (black) to 255 (white). Each image is associated with one of 10 categories, representing different types of fashion items. The categories are: 1. T-shirt/top 2. Trouser 3. Pullover 4. Dress 5. Coat 6. Sandal 7. Shirt 8. Sneaker 9. Bag 10. Ankle boot. Similar to MNIST, the images are stored in a single flattened array of 784 elements but can also be represented in a 28x28 matrix format.

B.2 Neural network architectures

For all methods, we adopt a family of factorized Gaussian distributions to model both priors and posteriors, characterized by the form $\pi = \mathcal{N}(w, \sigma\mathbf{I})$ where $w \in \mathbb{R}^d$ denotes the mean vector, and σ represents the scalar variance. We use feedforward neural networks for the MNIST dataset (LeCun and Cortes, 2010), while using convolutional neural networks for the Fashion MNIST dataset (Xiao et al., 2017).

Both our feedforward neural network and convolutional neural network are probabilistic, and each layer has a factorized (i.e. mean-field) Gaussian distribution.

Our feedforward neural network has the following architecture:

1. Input layer. Input size: 28×28 (flattened to 784 features).
2. Probabilistic linear layer 1. Input features: 784, output features: 600, activation: ReLU.
3. Probabilistic linear layer 2. Input features: 600, output features: 600, activation: ReLU.
4. Probabilistic linear layer 3. Input features: 600, output features: 600, activation: ReLU.
5. Probabilistic linear layer 4. Input features: 600, output features: 10, activation: Softmax.

Our convolutional neural network has the following architecture:

 https://doi.org/10.52202/079017-0570

1. Input layer. Input size: $1 \times 28 \times 28$.
2. Probabilistic convolutional layer 1. Input channels: 1, output channels: 32, kernel size: 3x3, activation: ReLU.
3. Probabilistic convolutional layer 2. Input channels: 32, output channels: 64, kernel size: 3x3, activation: ReLU.
4. Max pooling layer. Pooling size: 2x2.
5. Flattening layer. Flattens the output from the previous layers into a single vector.
6. Probabilistic linear layer 1. Input features: 9216, output features: 128, activation: ReLU.
7. Probabilistic linear layer 2 (output layer). Input features: 128, output features: 10, activation: Softmax.

B.3 Other details in the experiments

General for all methods The methods in comparisons are trained and evaluated using the procedure described in Section 2 and visually illustrated in Figure 1. We will provide some further details for each method later in the following. For all methods in comparison, we apply the optimization and evaluation method described in Section 5.1. For the approximation described in Section 5.1.1, we set the parameters $c_1 = c_2 = 5$. The lower bound for the prediction $p_{\min} = 1e - 5$. The δ in our bound and all the other methods is selected to be $\delta = 0.025$. As mentioned in Section 5.1.2, we use the PAC-Bayes-classic bound by McAllester in replacement of PAC-Bayes-kl when doing optimization. Note that for all methods, we also have to estimate the empirical loss of the posterior $\mathbb{E}_\pi[\cdot]$ described in Section 5.1.3. We also allocate the budget for the union bound for the estimation such that these estimations in the bound are controlled with probability at least $1 - \delta'$, where we chose $\delta' = 0.01$. Therefore, the ultimate bounds for all methods hold with probability at least $1 - \delta - \delta'$. Note that we do not consider such bounds during optimization but only when estimating the bounds.

For all methods, we adopt a family of factorized Gaussian distributions to model both priors and posteriors of all the learnable parameters of the classifiers, characterized by the form $\pi = \mathcal{N}(w, \sigma\mathbf{I})$ where $w \in \mathbb{R}^d$ denotes the mean vector, and σ represents the scalar variance. For all methods, we initialize an uninformed prior $\pi_0 = \mathcal{N}(w_0, \sigma_0\mathbf{I})$ that is independent of data, where the mean is randomly initialized, and the variance σ_0 is initialized to 0.03 (Pérez-Ortiz et al., 2021).

In the training process of all methods in our experiments, we set the batch size to 250, the number of training epochs to 200, and use stochastic gradient descent with a learning rate of 0.005 and a momentum of 0.95.

Uninformed priors We take π_0 defined above as the uninformed prior. We then learn the posterior ρ from the prior using the entire training dataset S, applying a PAC-Bayes bound. We evaluate the bound using, again, the entire training dataset S.

Data-informed priors We start with the same π_0 as the uninformed prior. We train the informed prior π_1 using S_1 with $|S_1| = |S|/2$ by minimizing a PAC-Bayes bound. The posterior ρ is then learned using the informed prior π_1 and the subset S_2 with $|S_2| = |S|/2$, again by minimizing a PAC-Bayes bound. The bound is evaluated using S_2.

Data-informed priors + excess loss We train the informed prior π_1 and the reference classifier h^* using S_1 that contains half of the training dataset. π_1 is obtained by minimizing a PAC-Bayes bound with the uninformed prior π_0, while the reference classifier h^* is obtained by an empirical risk minimizer (ERM). The posterior ρ is obtained by minimizing a PAC-Bayes bound on the excess loss between ρ and h^*. The prior used in the bound for both training and evaluation is the data-informed prior π_1. Therefore, the data for both training and evaluation of ρ must be the other half of data S_2.

B.4 Further results for the experiments

In this section, we report some more statistics for all methods.

For all methods, to calculate the classification loss of ρ on the testing data, $\mathbb{E}_\rho[\hat{L}(h, S^{\text{test}})]$ (Test 0-1), we sample one classifier for each data. The train 0-1 loss for all methods is computed on the entire training dataset S, while the test 0-1 loss for all methods is computed on the test dataset S_{test}.

B.4.1 Recursive PAC-Bayes

We report the additional results of Recursive PAC-Bayes on MNIST with $T = 2$ in Table 4, $T = 4$ in Table 5, and $T = 6$ in Table 6. We report Recursive PAC-Bayes on Fashion MNIST with $T = 2$ in Table 7, $T = 4$ in Table 8, and $T = 6$ in Table 9.

Table 4: Insight into the training process of the Recursive PAC-Bayes for $T = 2$ on MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.024 (3e-5)		.370 (1e-3)	.254 (2e-3)
2	30000	.013 (3e-3)	.024 (1e-4)	.136 (3e-3)	.321 (3e-3)	.139 (3e-3)

Table 5: Insight into the training process of the Recursive PAC-Bayes for $T = 4$ on MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.023 (8e-5)		.374 (1e-3)	.258 (1e-3)
2	52500	-4e-4 (1e-3)	.025 (3e-4)	.118 (1e-3)	.305 (1e-3)	.126 (2e-3)
3	45000	.053 (1e-3)	.002 (9e-5)	.087 (2e-3)	.240 (2e-3)	.114 (1e-3)
4	30000	.054 (1e-3)	.001 (2e-5)	.083 (1e-3)	.203 (8e-4)	.109 (1e-3)

Table 6: Insight into the training process of the Recursive PAC-Bayes for $T = 6$ on MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.019 (7e-5)		.425 (1e-3)	.311 (3e-3)
2	58125	-0.013 (1e-3)	.032 (6e-4)	.128 (2e-3)	.341 (3e-3)	.139 (1e-3)
3	56250	.050 (1e-3)	.003 (1e-4)	.093 (6e-4)	.264 (1e-3)	.117 (2e-3)
4	52500	.051 (1e-3)	.001 (6e-5)	.080 (9e-4)	.212 (5e-4)	.108 (2e-3)
5	45000	.051 (1e-3)	9e-4 (3e-5)	.076 (2e-3)	.182 (1e-3)	.104 (6e-4)
6	30000	.049 (1e-3)	7e-4 (3e-5)	.074 (1e-3)	.166 (1e-3)	.101 (1e-3)

Table 7: Insight into the training process of the Recursive PAC-Bayes for $T = 2$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.011 (3e-5)		.466 (1e-3)	.389 (5e-3)
2	30000	.064 (3e-3)	.013 (2e-4)	.171 (4e-3)	.404 (3e-3)	.266 (5e-3)

Table 9: Insight into the training process of the Recursive PAC-Bayes for $T = 6$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		9e-3 (8e-5)		.534 (4e-3)	.462 (5e-3)
2	58125	.013 (6e-3)	.023 (1e-3)	.151 (4e-3)	.418 (5e-3)	.254 (6e-3)
3	56250	.091 (3e-3)	3e-3 (4e-4)	.141 (2e-3)	.350 (3e-3)	.223 (1e-3)
4	52500	.090 (2e-3)	9e-4 (9e-5)	.121 (1e-3)	.296 (1e-3)	.207 (1e-3)
5	45000	.090 (1e-3)	6e-4 (3e-5)	.117 (2e-3)	.265 (2e-3)	.199 (2e-3)
6	30000	.093 (1e-3)	5e-4 (2e-5)	.122 (1e-3)	.255 (1e-3)	.198 (1e-3)

B.4.2 Uninformed priors

We report the additional results of uninformed priors (McAllester, 1998) on MNIST and Fashion MNIST in Table 10. As described earlier in Section 2, Section 5, and Section B.3, we evaluate the bound using the entire training set.

https://doi.org/10.52202/079017-0570

Table 8: Insight into the training process of the Recursive PAC-Bayes for $T = 4$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}} \circ \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \Vert \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.011 (1e-4)		.476 (4e-3)	.397 (6e-3)
2	52500	.032 (6e-4)	.017 (9e-4)	.147 (2e-3)	.386 (3e-3)	.240 (4e-3)
3	45000	.100 (3e-3)	3e-3 (1e-4)	.138 (5e-3)	.331 (4e-3)	.222 (5e-3)
4	30000	.095 (2e-3)	7e-4 (5e-5)	.128 (2e-3)	.293 (1e-3)	.213 (3e-3)

Table 10: Further details to compute the bound for the uninformed prior approach on MNIST and Fashion MNIST.

	$\mathbb{E}_\rho[\hat{L}(h, S)]$	$\frac{\text{KL}(\rho \Vert \pi_0)}{n}$	Bound	Test 0-1
MNIST	.343 (2e-3)	.023 (4e-5)	.457 (2e-3)	.335 (3e-3)
F-MNIST	.382 (2e-3)	.011 (8e-6)	.464 (2e-3)	.384 (5e-3)

B.4.3 Data-informed priors

We report the additional results of data-informed priors (Ambroladze et al., 2007) on MNIST and Fashion MNIST in Table 11. As described earlier in Section 2, Section 5, and Section B.3, we evaluate the bound using S_2 that is independent of the data-informed prior π_1.

Table 11: Further details to compute the bound for the data-informed prior on MNIST and Fashion MNIST.

	$\mathbb{E}_\rho[\hat{L}(h, S_2)]$	$\frac{\text{KL}(\rho \Vert \pi_0)}{\lvert S_2 \rvert}$	Bound	Test 0-1
MNIST	.376 (8e-4)	8e-4 (9e-6)	.408 (9e-4)	.371 (6e-3)
F-MNIST	.412 (1e-3)	4e-4 (7e-6)	.440 (1e-3)	.413 (6e-3)

B.4.4 Data-informed priors + excess loss

We report the additional results of data-informed priors + excess loss (Mhammedi et al., 2019) on MNIST and Fashion MNIST in Table 12 and 13. As described earlier in Section 2, Section 5, and Section B.3, we evaluate the bound using S_2 that is independent of the data-informed prior π_1 and the reference prediction rule h^*. The bound is composed of two parts: a bound on the excess loss of ρ with respect to h^* (Excess bound) and a single hypothesis bound on h^* (h^* bound). We report the two components of the bound in Table 12. We provide further details to compute these bounds from the losses of their corresponding quantities in Table 13.

Table 12: Details to compute the bound for the data-informed prior and excess loss on MNIST and Fashion MNIST. The table shows the bound on the excess loss of ρ with respect to h^* (Excess bound) and a single hypothesis bound on h^* (h^* bound).

	Ex. Bound	h^* Bound	Bound	Test 0-1
MNIST	.162 (1e-3)	.029 (4e-4)	.192 (2e-3)	.151 (3e-3)
F-MNIST	.196 (5e-3)	.145 (1e-3)	.342 (6e-3)	.285 (5e-3)

Table 13: Further details to compute the bound for the data-informed prior and excess loss on MNIST and Fashion MNIST. The table shows the empirical excess loss $\mathbb{E}_\rho[\hat{\Delta}(h, h^*, S_2)]$, where we define $\Delta(h, h^*, S_2) = \hat{L}(h, S_2) - \hat{L}(h^*, S_2)$, and its bound (Excess Bound). It also shows the empirical loss of the reference prediction rule $\hat{L}(h^*, S_2)$ and its bound. The computation of such bound does not involve the KL term.

	$\mathbb{E}_\rho[\hat{\Delta}(h, h^*, S_2)]$	$\frac{\text{KL}(\rho \Vert \pi_1)}{\lvert S_2 \rvert}$	Ex. Bound	$\hat{L}(h^*, S_2)$	h^* Bound	Bound
MNIST	-0.011 (3e-3)	.035 (5e-4)	.162 (1e-3)	.026 (4e-4)	.029 (4e-4)	.192 (2e-3)
F-MNIST	.104 (6e-3)	.018 (5e-4)	.196 (5e-3)	.112 (1e-3)	.145 (1e-3)	.342 (6e-3)

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: We tried to write an abstract and an introduction that reflected the rest of the paper as accurately as possible.

   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 have tried to be as honest as possible in setting the framework of this theoretical work and have explicitly mentioned the limiting assumptions of the paper, e.g. i.i.d. data.

   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?

https://doi.org/10.52202/079017-0570

Answer: [Yes]

Justification: We have clearly stated the assumptions made in the paper, and our theoretical results are accompanied by detailed proofs.

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: The datasets are available online and we have tried to provide as much detail as possible to make our experiments transparent and reproducible.

   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: The code has been uploaded, the datasets are open source.

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: We have introduced many different notations to accurately specify the different data splits and parameters used in our framework, and have tried to detail the optimization procedure to make our results as understandable as possible.

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: We provided for each experiment the corresponding standard deviations over different runs of the experiments. Please refer to the tables in the paper for more details.

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.

https://doi.org/10.52202/079017-0570

- 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: As mentioned in the paper, all the experiments were run on a personal laptop.

   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: We have not violated any of the points in 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: This is a theoretical paper with no specific societal impact.

    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: The paper poses no such risks.

    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: The open source datasets we used in the paper were properly cited, namely MNIST (LeCun and Cortes, 2010) and Fashion MNIST (Xiao et al., 2017).

    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.

https://doi.org/10.52202/079017-0570

- 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: [NA]

    Justification: We have not released any new assets.

    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: Our paper did not involve crowdsourcing nor research with human subjects.

    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: Our paper did not involve crowdsourcing nor 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.

https://doi.org/10.52202/079017-0570