NON-ASYMPTOTIC PAC-BAYES BOUNDS ON GENERALIZATION ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

Constructing non-vacuous PAC-Bayes bounds on generalization errors for unbounded risk functionals, especially in the non-asymptotic regime, is an active area of research. However, current state of the art results are applicable only in some very specialized cases. In this work, we give an integrability condition which exactly characterizes when any risk functional, for a given data set and model space, admits such bounds using the Levy-Khintchine theorem. Further, we derive a Bahadur-Rao type exact asymptotic bound, which is much sharper than a traditional Chernoff type inequality, especially in the under-sampled regime. These bounds give us the flexibility to construct data or model-dependent consistency promoting updates to a data-free prior, which provably improves the generalization performance.

1 INTRODUCTION

In this work we are interested in provable control of generalization error in model estimation, especially in the non-asymptotic or under-sampled regime. In this scenario, the number of observed samples are significantly lower than the degrees of freedom in the model, thereby leading to an ill-posed estimation problem and consequently a large expected generalization error. Such situations are extremely common in classical high dimensional statistics [Sur and Candès (2019)], for example in single cell genomics [McDavid et al. (2019)] or GWAS [Brzyski et al. (2017)] studies, the data is not only extremely high dimensional, labelled data acquisition can be error-prone, expensive and time consuming. However recently, with the plethora of wide ranging successful applications of over-parametrized deep neural networks [Zhang et al. (2016)], the question of constructing generalization bounds is of great interest in the deep learning community [Kawaguchi et al. (2017)].

From the theory of statistical learning [Vapnik (2000)] an optimal model is defined by the minimizer of a continuous non-negative risk functional $\mathrm{R} : \left(\mathcal{H}, \mathcal{P}_2\left(\mathbb{R}^d\right)\right) \to \mathbb{R}_+$ which is characterized by being differentiable in the first argument and linear in the second. Here $\mathcal{P}_2\left(\mathbb{R}^d\right)$ is the space of Borel probability measures on $\mathbb{R}^d$ with bounded variance and $\mathcal{H} \subseteq \mathrm{L}^2\left(\mathbb{R}^d\right)$ is the space of square integrable functions, represented by either parametric models like exponential families and their mixtures [Meila (2000); Nguyen (2011); Redner and Walker (1984)] or non-parametric models like infinite mixture models [Gershman and Blei (2011); Kleijn and Zhao (2013); Locatello et al. (2017); Nguyen (2011); Petrone and Veronese (2002); Ramaswamy, Wu and Ghosal (2007)], etc.

However in practice, the true measure say $\nu \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, is unknown and one only has access to a set of i.i.d. samples $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \nu$ from it. The core idea then, is to use the empirical risk functional $\mathrm{R}_{\nu_n^{\text{emp}}} := \mathrm{R}\left(\cdot, \nu_n^{\text{emp}} := \frac{1}{n}\sum_i \delta_{X_i}\right)$ as a surrogate and minimize it instead [Vapnik (2000)]. Clearly, such a minimizer, say $\hat{f}_n \in \mathcal{H}$ can be said to approximate the true optimum if and only if its generalization error i.e.

$$\underbrace{\Delta_{\nu_n^{\text{emp}}}\left(\hat{f}_n\right)}_{\text{Generalization error}} := \mathrm{R}_{\nu_n^{\text{emp}}}\left(\hat{f}_n\right) - \mathrm{R}^*$$

can be bounded almost surely, where R^* is the theoretically minimal risk, also known as the Bayes Risk [Vapnik (2000)]. Note that R^* may not be achievable inside $\mathcal{H}$. Constructing probabilistic bounds on the generalization error, is the one of the central aims of statistical learning theory.

By constructing an appropriate prior on the model space $\mathcal{H}$, one can a priori exclude models which are not expected to fit the learning problem i.e. reduce capacity. A prior needs to be constructed using only minimal assumptions [Belkin et al. (2018); Neyshabur et al. (2018); Mukherjee et al. (2006); Kleijn and Zhao (2013)], so as to ensure that it puts a positive probability in the neighborhood of the true model (2), while removing all unsuitable models. For example, in linear regression a Gaussian prior on the parameter space leads to Ridge regression which gives the optimal least squares solution [Parikh et al. (2014)] in the under-sampled regime, while a Laplace prior leads to Lasso regression which can be interpreted to be a convex relaxation to the best subset regression problem [Parikh et al. (2014)] and hence leads to relatively sparse coefficient estimates.

A principled approach to constructing priors based on problem dependent assumptions comes from the theory of Levy processes, where Gaussian process, Compound Poisson process, etc are used to induce distributions over partitions of the model space, leading to for example the Pitman-Yor process, etc. For example, a Gaussian process prior over a space of functions, assumes that the marginal distribution over any finite set of evaluations has Gaussian distribution. Such an assumption leads to smooth functions whose evaluation in a neighborhood is highly correlated, while almost independent outside its region of influence. While if we assume complete independence of all partitions of the sample space, then we end up with completely random measures like Compound Poisson processes. For more details, see [Gershman and Blei (2011)].

The Bayesian approach can be alternatively formulated from the point of view of the Minimum Description Length principle [Grünwald and Mehta (2019); Dziugaite and Roy (2017a)]. MDL is a theory of inductive and statistical inference, where statistical learning is defined as the search for the best hypothesis among a set of hypotheses, and a model in that hypothesis, that is best able to describe the regularities in the data, while compressing it the most [Foster et al. (2019); Arora et al. (2018); Shwartz-Ziv and Tishby (2017); Zhou et al. (2018)]. It can be seen as the principled formalization of the Occam's razor principle.

In the PAC-Bayesian literature, data-dependent priors have also been suggested, constructed using various techniques. One common approach is via training the prior measure on a held out portion of the data [Ambroladze et al. (2007); Parrado-Hernández et al. (2012)]. Another, following a non-parametric approach is to ensure martingale concentration by assuming stability w.r.t. small changes in the data [Dziugaite and Roy (2017b)]. In [Rivasplata et al. (2020)] approached the problem of data-dependent priors more generally by deriving PAC-Bayes inequalities for arbitrary convex error functionals, but still requiring rather strong assumptions on the exponential moments.

Usually in the literature, the risk functional is assumed to be bounded [Catoni (2007); McAllester (1999; 2013); Audibert and Bousquet (2007)], however recently results for unbounded risks have been derived under assumptions of exponentially decaying tails i.e. sub-Gaussianity of the data generating process [Germain et al. (2016)]. In this work, using Levy-Khintchine theorem [Bertoin (2006)], we characterize the condition under which the required moment generating function exists for non-negative risk functionals. Such a characterization allows us to expand the domain of applicability to general unbounded risk functionals even when the data generating process has polynomially decreasing tails. In order to that, we need to construct a stochastic process known as a subordinator using samples from our push-forward empirical measure.

Our Contribution The main focus of this work is on the construction of bounds on the estimation error. In the discussion above, we showed that such a bound can be used to construct the consistency pseudo-prior, which is a data dependent update (9) to an η-sufficient prior, specific to a given risk functional R and model space $\mathcal{H}$. These constructions are then sufficient to derive the strong PAC-Bayes bounds (10) discussed earlier.

We approach the problem of estimation error by first observing that the risk functional pushes forward the empirical measure to a measure on the non-negative real line. We show that any such push-forward measure for i.i.d. data is a Levy measure, whose moment generating function exists under very weak conditions (15). Importantly, even the boundedness of the risk functional is not necessary. Further, in this case the Levy-Khintchine representation formula [Bertoin (2006)] allows us to estimate the moment generating function by modeling the push-forward measure either by a Gamma subordinator for exponentially decreasing tail or Stable subordinator for polynomially decreasing tail, and hence derive data/model dependent PAC-Bayes bounds.

Secondly, we derive an Bahadur-Rao type exact asymptotic bound, which is much sharper than a traditional Chernoff-inequality especially in the under-sampled regime. Thereby allowing us to provide sufficient conditions to ensure δ_n-strong-consistency (3) of the empirical Risk functional. Such a bound is a result of a combination of Donsker-Varadhan duality and Berry-Essen estimates [Dembo and Zeitouni (2009)] for the convergence rates in central limit theorem.

2 PAC-BOUNDS ON GENERALIZATION ERROR

To fix notations, let $\mathrm{R} : \left(\mathcal{H}, \mathcal{P}_2\left(\mathbb{R}^d\right)\right) \to \mathbb{R}_+$ be a continuous non-negative risk functional, and $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \nu \in \mathcal{P}_2\left(\mathbb{R}^d\right)$ be i.i.d. random variables, with $\nu_n^{\text{emp}} := \frac{1}{n}\sum_i \delta_{X_i}$ as its empirical measure with the corresponding empirical risk functional $\mathrm{R}_{\nu_n^{\text{emp}}} := \mathrm{R}\left(\cdot, \nu_n^{\text{emp}} := \frac{1}{n}\sum_i \delta_{X_i}\right)$. The starting point for bounding the generalization error is the following orthogonal decomposition which isolates its deterministic and stochastic parts for any $f \in \mathcal{H}$

$$\Delta_{\nu_n^{\text{emp}}}(f) = \underbrace{\left[\mathrm{R}_\nu(f) - \mathrm{R}^*\right]}_{\text{Approximation error}} + \underbrace{\left[\mathrm{R}_{\nu_n^{\text{emp}}}(f) - \mathrm{R}_\nu(f)\right]}_{\text{Estimation error}} \tag{1}$$

Here the approximation error [Vapnik (2000)] is a deterministic functional of the model space $\mathcal{H}$ and the unknown true measure $\nu \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, while estimation error [Vapnik (2000)] is a stochastic functional dependent on the sampling process and properties of the risk functional.

2.1 PAC-BOUND APPROACH

Then we can define the model space $\mathcal{H}$ to be *η-sufficient* for the learning problem, if $\eta \geq 0$ is the smallest non-negative finite real number for which there exists some prior measure $\pi_0 \in \mathcal{P}_2(\mathcal{H})$, such that

$$\pi_0\left(\left\{f \in \mathcal{H} :\mid \mathrm{R}_\nu(f) - \mathrm{R}^* \mid < \eta\right\}\right) > 0 \tag{2}$$

If a π_0 exists for $\eta = 0$, then the model space is called *well-specified*. Clearly if a finite dimensional model space (e.g. Linear models, etc) is well-specified, then the Lebesgue measure would always satisfy such a condition. However, since for an infinite dimensional model space (e.g. RKHS, Fourier, Wavelet, etc) no Lebesgue measure exists [Kleijn and Zhao (2013)], even in the ideal case the priors need to be represented via completely random measures [Kingman (1967)] and related Levy processes like Gaussian process [Seeger (2002); Donnet et al. (2014)], Compound Poisson process [Gugushvili et al. (2019)], Pitman-Yor process [Gershman and Blei (2011)], etc.

Similarly, we can define the empirical risk to be *δ_n-strongly consistent* if for any $\epsilon > 0$ and some $f \in \mathcal{H}$, there exists a monotonically decreasing sequence $\delta_n := \delta\left(f, \nu_n^{\text{emp}}, \epsilon\right) \to 0^+$ as $n \to \infty$, such that the tail probability of the estimation error satisfies

$$\Pr\left(\mid \mathrm{R}_{\nu_n^{\text{emp}}}(f) - \mathrm{R}_\nu(f) \mid > \epsilon\right) < \delta_n,\ \delta_n \in [0, 1] \tag{3}$$

This means that if for some function $f \in \mathcal{H}$, such a δ_n exists, then the estimation error vanishes with increasing sample size. In the literature [Fu (1982); Kleijn and Zhao (2013); Mukherjee et al. (2006); Shwartz-Ziv and Tishby (2017); Wu and Ghosal (2007)], one usually considers *weak consistency* for the model space $\mathcal{H}$, by insisting on convergence (3) only in the worst case i.e. if $\sup_{f\in\mathcal{H}} \delta_n$ is monotonically decreasing with sample size. However, in that case we loose the ability to distinguish the rates of convergence for each model, which in this work is exploited as a measure of model complexity. Clearly a model can be said to have lower complexity if its estimation error vanishes faster i.e. the corresponding δ_n is smaller for the same sample size.

Therefore we can say that a model estimation problem is *well posed* if and only if the model space $\mathcal{H}$ is η-sufficient, the empirical risk estimator is δ_n-strongly consistent and an unique solution exists to the empirical risk minimization problem (4) i.e. the conditions under which an unique solution exists to

$$\inf_{f \in \text{supp}(\pi_0)} \mathrm{R}_{\nu_n^{\text{emp}}}(f) \tag{4}$$

$$\text{such that } \log \delta_n(f, \epsilon) < 0 \text{ and } \pi_0 \in \mathcal{P}_2(\mathcal{H}) \text{ is } \eta\text{-sufficient} \tag{5}$$

Note that here we define the optimization problem in terms of estimating δ_n instead of the estimation error ϵ, which is usual in the literature [Catoni (2007)].

Now the necessary conditions for a solution to the constrained optimization problem (4) is provided by the Karush-Kuhn-Tucker theorem [Parikh et al. (2014)], which states that if (f^*, λ^*) is a saddle point for the associated Lagrangian function

$$\mathrm{L}\left(f, \nu_n^{\text{emp}}, \lambda\right) \coloneqq \mathrm{R}_{\nu_n^{\text{emp}}}\left(f\right) + \lambda \log \delta_n\left(f, \epsilon\right) \tag{6}$$

where $f \in \text{supp}\left(\pi_0\right)$ and $\lambda \geq 0$, then f^* is an optimal solution. Note that here we use $\log \delta_n$, here which not only allows us to formulate the Lagrangian in a standard regularization framework, but also explicitly delineates the data-free prior π_0, from a data-dependent update to it, which we derive below (9).

Given the constraints (2) and (3), it is easy to see that, then the generalization error for f^* is bounded above by $\eta + \epsilon$ with probability at least $1 - \delta_n$ i.e.

$$\Pr\left(\mid \Delta_{\nu_n^{\text{emp}}}\left(f^*\right) \mid < \eta + \epsilon\right) > 1 - \delta_n \tag{7}$$

Such a bound on generalization error is known as a PAC-Bound i.e. a Probably-Approximately-Correct Bound [Catoni (2007)] signifying its probabilistic nature.

2.2 PAC-Bayes Approach

Now if the risk functional were assumed to be convex, the necessary conditions defined by the Karush-Kuhn-Tucker theorem are known to be sufficient and hence the solution is unique e.g. ordinary least squares regression [Parikh et al. (2014)]. Unfortunately, in many modern applications such an assumption cannot be made. However, if we can assume that the derivative of the Lagrangian functional (6) is Lipschitz continuous, then its Donsker-Varadhan dual [Donsker and Varadhan (1975)] is known to be strongly convex [Zhou (2018)] in the space of probability measures with finite second moments.

PAC-Bayesian analysis [McAllester (1999); Ambroladze et al. (2007); Audibert and Bousquet (2007); Bégin et al. (2014); Catoni (2007); Freund (1998); Germain et al. (2016); Guedj (2019)] is a framework which exploits such a duality, by defining the *pseudo-posterior* measure as the unique measure which minimizes the expected Lagrangian in the support of the prior i.e. for $\gamma > 0$

$$\hat{\pi}_n \coloneqq \arg \inf_{\pi_n \in \mathcal{P}_2(\mathcal{H})} \left\{ \mathbb{E}_{\pi_n}\left[\mathrm{L}\right] - \mathbb{E}_{\pi_0}\left[\mathrm{L}\right] + \frac{1}{\gamma}\mathrm{KL}\left(\pi_n \mid \pi_0\right) \right\} \tag{8}$$

which admits an unique analytic solution, given by the associated Gibbs measure

$$\begin{aligned} d\hat{\pi}_n &\propto \exp\left(-\gamma \mathrm{L}\left(f, \nu_n^{\text{emp}}, \lambda\right)\right) d\pi_0 \\ &\propto \exp\left(-\gamma \mathrm{R}_{\nu_n^{\text{emp}}}\left(f\right)\right) d\hat{\pi}_0^{\lambda} \end{aligned}$$

where w.l.o.g. with $\lambda > 0$,

$$d\hat{\pi}_0^{\lambda}\left(f\right) \propto \exp\left(-\lambda \log \delta_n\left(f, \epsilon\right)\right) d\pi_0\left(f\right) \tag{9}$$

We call $\hat{\pi}_0^{\lambda}$ the *consistency pseudo-prior*, and interpret it to be a data-dependent (or model-dependent) consistency promoting update to the *data-free* η-sufficient prior π_0.

Ignoring the consistency criterion i.e. for $\lambda = 0$, we end up with the standard data-independent PAC-Bayesian approach, which has had great success especially in the case of classification and regression problems with bounded risk functionals [Catoni (2007); Germain et al. (2009, 2015); Guedj (2019)]. However in the under-sampled regime, when the number of samples are smaller than the number of parameters, such optimality results do not hold. In this case, the data-free prior, which in the end is just a guess, unreasonably dominates the posterior. A data/model-dependent pseudo-prior like (9) on the other hand, provides a consistent approach to modifying data-free priors by promoting functions of lower "complexity" in the support of the prior which admit faster convergence based on the available data. This allows us to construct accurate bounds on the generalization error corresponding to both the prior as well as the posterior.

Following the same logic as above, if for any η-sufficient prior $\pi_0 \in \mathcal{P}_2\left(\mathcal{H}\right)$, we can estimate the consistency pseudo-prior (9) and its associated pseudo-posterior measure $\hat{\pi}_n$ (8), then the corresponding PAC-Bayes confidence interval on generalization error is given by

$$\Pr\left(\mid \mathbb{E}_{\hat{\pi}_0^{\lambda}}\left[\Delta_{\nu_n^{\text{emp}}}\right] \mid < \eta + \epsilon\right) > 1 - \mathbb{E}_{\hat{\pi}_0^{\lambda}}\left[\delta_n\left(f, \epsilon\right)\right]$$

and

$$\Pr\left(\mid \mathbb{E}_{\hat{\pi}_n}\left[\Delta_{\nu_n^{\text{emp}}}\right] \mid < \eta + \epsilon\right) > 1 - \mathbb{E}_{\hat{\pi}_n}\left[\delta_n\left(f, \epsilon\right)\right] \tag{10}$$

Note that since $\hat{\pi}_n \ll \hat{\pi}_0^{\lambda}$ (absolutely continuous), clearly $\mathbb{E}_{\hat{\pi}_n}\left[\delta_n\left(f, \epsilon\right)\right] < \mathbb{E}_{\hat{\pi}_0^{\lambda}}\left[\delta_n\left(f, \epsilon\right)\right] < \delta_n\left(f, \epsilon\right)$ and therefore the expected performance of sample functions under the posterior is necessarily better than under the consistency pseudo-prior, which on the other hand performs better than a fixed point-wise estimate in probability.

Therefore given a risk functional with Lipschitz continuous derivatives, if we can construct the function $\delta_n\left(f, \epsilon\right)$ which satisfies (3), then the optima of the Lagrangian functional (6) can be represented by an unique pseudo-posterior measure (8), whose samples satisfy the PAC-Bayes bound on generalization error (10). In the next section we provide such a construiction.

3 Exact Asymptotic Bound on Estimation Error

3.1 Bounding the Estimation Error

In order to calculate the data/model dependent prior update (9), we need to calculate the probability

$$\Pr\left(\mathrm{R}_{\nu_n^{\text{emp}}}\left(f\right) \geq \mathrm{R}_{\nu}\left(f\right) + \epsilon\right)$$

Let $\left\{r_i\left(f\right) \coloneqq \mathrm{R}\left(f\right)^{\sharp}\delta_{X_i}\right\}_{i=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \lambda_n$ be i.i.d. random variables, then most bounds on Estimation error in the literature are based on the Cramer-Chernoff method [Boucheron et al. (2013)]. A Chernoff-type inequality, is defined for $\Lambda_{\mathrm{R}}\left(\gamma\right) = \log \mathbb{E}\left[e^{\gamma \mathrm{R}}\right]$ we have

$$\begin{aligned}\Pr\left(\mathrm{R} \geq t\right) &\leq \exp\left(-\sup_{\gamma \in \mathbb{R}}\left\{\gamma t - \Lambda_{\mathrm{R}}\left(\lambda\right)\right\}\right) \\ &= \exp\left(-\Lambda_{\mathrm{R}}^{*}\left(t\right)\right)\end{aligned} \tag{11}$$

where $\Lambda_X^{*}\left(t\right)$ is the Legendre-Fenchel transform of Λ_{R} and $t \geq \mathbb{E}\left[\mathrm{R}\right]$. Since $\Lambda_{\mathrm{R}}\left(0\right) = 0$, Λ_{R}^{*} is a non-negative function. Further if $\mathbb{E}\left[\mathrm{R}\right]$ exists, then the convexity of the exponential function and Jensen's inequality imply $\Lambda_{\mathrm{R}}\left(\gamma\right) \geq \gamma \mathbb{E}\left[\mathrm{R}\right]$ and hence for all negative values of γ, $\gamma t - \Lambda_{\mathrm{R}}\left(\gamma\right) \leq 0$ whenever $t \geq \mathbb{E}\left[\mathrm{R}\right]$.

Now bounding the estimation error, requires constructing bounds on sums of real valued random variables. The main idea here is to set $\mathrm{R}_n = \sum_{i=1}^{n}\left(\mathrm{R}_i - \mathbb{E}\left[\mathrm{R}_i\right]\right)$, we have

$$\Lambda_{\mathrm{R}_n}\left(\gamma\right) = \log \mathbb{E}\left[e^{\gamma \sum_{i=1}^{n}\left(\mathrm{R}_i - \mathbb{E}\left[\mathrm{R}_i\right]\right)}\right] = \sum_{i=1}^{n} \log \mathbb{E}\left[e^{\gamma\left(\mathrm{R}_i - \mathbb{E}\left[\mathrm{R}_i\right]\right)}\right]$$

and hence the Cramer-Chernoff method can be applied. For example, if R_i is assumed to be bounded in a compact interval on the real line, then we end up with the well-known Hoeffding's inequality which has been extensively applied in [Audibert and Bousquet (2007); Kääriäinen and Langford (2005); McAllester (1999); Boucheron et al. (2013)] to derive PAC-Bayes bounds. Under sub-Gaussian (i.e. $\Lambda_{\mathrm{R}}\left(\lambda\right) \leq \frac{\lambda^2 \nu}{2}$ for all $\lambda \in \mathbb{R}$ with variance factor v) or sub-Gamma tail (i.e. $\Lambda_{\mathrm{R}}\left(\lambda\right) \leq \frac{\lambda^2 \nu}{2\left(1 - c\lambda\right)}$ for all $0 < \lambda < 1/c$ with variance factor v and scale factor c) assumptions over R_i, using the fact that $\log u \leq u - 1$ for $u > 0$, we can derive Bennet and Bernstein's inequalities [Boucheron et al. (2013)] respectively, which have been used extensively to derive bounds on the estimation error in various applications [Mhammedi et al. (2019); Tolstikhin and Seldin (2013); Alquier and Guedj (2018); Ambroladze et al. (2007); Arora et al. (2018); Audibert and Bousquet (2007); Holland (2019)].

In the following theorem, we show that such a method can be made much sharper than the one produced by the Cramer-Chernoff method.

Theorem 1. *Given a set of random variables $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \nu \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, consider the real valued push-forward measure $R_{\delta_{X_1}}\left(f\right)^{\sharp}\nu$. If the cumulant generating function of its centered empirical measure, $\Lambda_{\nu_n^{emp}}\left(\gamma\right) \coloneqq \log \mathbb{E}_{\nu_n^{emp}}\left[\exp \gamma\left(R_{\delta_{X_1}}\left(f\right) - R_{\nu_n^{emp}}\left(f\right)\right)\right]$ is such that $\mathcal{D}^{o}_{\Lambda_{\nu_n^{emp}}} \coloneqq$*

$\left\{\gamma \geq 0 : \Lambda_{\nu_n^{emp}}(\gamma) < \infty\right\} \supset \{0\}$, *then for any* $\epsilon \in \left\{\Lambda'_{\nu_n^{emp}}(\gamma) : \gamma \in \mathcal{D}^o_{\Lambda_{\nu_n^{emp}}}\right\}$, *we have*

$$Pr\left(R_{\nu_n^{emp}}(f) \geq R_\nu(f) + \epsilon\right) = \exp\left(-n\Lambda^*_{\nu_n^{emp}}(\epsilon)\right)\left\{\frac{\xi}{\sqrt{2\pi n}}\frac{\sqrt{\Lambda^{*''}_{\nu_n^{emp}}(\epsilon)}}{\Lambda^{*'}_{\nu_n^{emp}}(\epsilon)} + \mathcal{O}\left(n^{-3/2}\right)\right\} \tag{12}$$

$$=: \frac{1}{2}\delta_n(f,\epsilon)$$

where

$$\Lambda^*_{\nu_n^{emp}}(\epsilon) := \sup_{\gamma \in \mathcal{D}^o_{\Lambda_R}}\left\{\gamma\epsilon - \Lambda_{\nu_n^{emp}}(\gamma)\right\} \;\&\; \xi = \begin{cases} 1 & \text{if } R_{\delta_{X_1}}(f) \text{ is continuous} \\ \frac{\Lambda^{*'}_{\nu_n^{emp}}(\epsilon)}{1-\exp\left(-\Lambda^{*'}_{\nu_n^{emp}}(\epsilon)\right)} & \text{if } R_{\delta_{X_1}}(f) \text{ is discrete} \end{cases}$$

Further the corresponding asymptotic convergence rate satisfies the Chernoff's theorem i.e. this estimate cannot be dominated by another estimator.

Proof. Let $Y_i \overset{\text{i.i.d.}}{\sim} \lambda_n := \mathrm{R}(f)^\sharp \pi_n^{\text{emp}}$ for which the logarithmic moment generating function $\Lambda_n(\eta) = \log \mathbb{E}_{\mu_n}\left[e^{\eta Y_1}\right]$ exists and set $A(\epsilon) = [\epsilon, \infty)$, where $\epsilon = \Lambda'_n(\eta)$ for some positive $\eta \in \mathcal{D}^\circ_{\Lambda_n} := \{\zeta \in \mathbb{R}_+ : \Lambda_n(\zeta) < \infty\}$. Note that $\lim_{n\to\infty}\Lambda'_n(0) = \mathbb{E}_{\mathrm{R}(f)^\sharp\pi}[\psi] =: \mathrm{R}(f_\nu)$ via weak law of large numbers [Dembo and Zeitouni (2009)]. From Donsker-Varadhan duality, we know that the exponentially tilted probability measure i.e. $d\widetilde{\lambda_n}(x) = e^{\eta x - \Lambda_n(\eta)}d\lambda_n(x)$ represents the distribution of random variables Y_i normalized and centered around $\epsilon = \Lambda'_n(\eta)$ i.e.

$$Z_i := {}^{(Y_i-\epsilon)}/_{\sqrt{\Lambda''_n(\eta)}} \overset{\text{i.i.d.}}{\sim} \widetilde{\lambda_n}$$

Further, let $U_n := n^{-1/2}\sum_{i=1}^n Z_i$ and denote its cumulative distribution function by

$$H_n(u) = \Pr(U_n < u), \;\; (-\infty < u < \infty)$$

Then by construction we have $\mathrm{R}_n^{\text{emp}}(f_\nu) = \epsilon + U_n\sqrt{{\Lambda''_n(\eta)}/{n}}$, which means that

$$\begin{aligned}
\Pr\left(\mathrm{R}_n^{\text{emp}}(f_\nu) \geq \epsilon\right) &= \mathbb{E}_{\mu_n}\left[1_{\left\{\mathrm{R}_n^{\text{emp}}(f_\nu)\geq\epsilon\right\}}\right] = \mathbb{E}_{\mu_n}\left[1_{\{U_n\geq 0\}}\right] \\
&= e^{-n\Lambda^*_n(\epsilon)}\mathbb{E}_{\widetilde{\mu_n}}\left[e^{-\eta\sqrt{n\Lambda''_n(\eta)}U_n}1_{\{U_n\geq 0\}}\right] \\
&= e^{-n\Lambda^*_n(\epsilon)}\int_0^\infty e^{-\alpha(\eta)\sqrt{n}u}dH_n(u) \\
&= e^{-n\Lambda^*_n(\epsilon)}\int_0^\infty e^{-t}\left[H_n\left(\frac{t}{\alpha(\lambda)\sqrt{n}}\right) - H_n(0)\right]dt
\end{aligned}$$

using the integration by parts formula and a change of variables with $t = \alpha(\lambda)\sqrt{n}u$, where $\alpha(\lambda) = \lambda\sqrt{\Lambda''_n(\lambda)}$ and

$$\Lambda^*_n(\epsilon) := \sup_{\eta\in\mathcal{D}^\circ_\Lambda}\{\eta\epsilon - \Lambda_n(\eta)\}$$

Here $\alpha(\lambda)$ can be re-written as a function of ϵ using the equalities

$$\lambda = \Lambda^{*'}_n(\epsilon) \text{ and } \Lambda''_n(\lambda) = {}^1/_{\Lambda^{*''}_n(\epsilon)}$$

Finally, we need to estimate the integral

$$I_n := \int_0^\infty e^{-t}\left[H_n\left(\frac{t}{\alpha(\lambda)\sqrt{n}}\right) - H_n(0)\right]dt$$

To this end, consider the Berry-Esséen expansion of the cumulative distribution functions H_n which is given by [Dembo and Zeitouni (2009)]

$$\lim_{n\to\infty}\left\{\sqrt{n}\sup_u \mid H_n(u) - \Phi(u) - \frac{m_3}{6\sqrt{n}}\left(1-u^2\right)\phi(u) \mid\right\} = 0$$

where $m_3 \coloneqq \mathbb{E}_{\widetilde{\psi^\sharp f_\nu^\sharp \pi_n^{\text{emp}}}}\left[Z^3\right] < \infty$, $\phi(u) = {}^1\!/\!_{\sqrt{2\pi}} e^{-u^2/2}$ is the standard Normal density and $\Phi(u) = \int_{-\infty}^{u} \phi(t)dt$. Now since the Taylor expansion of Φ is given by

$$\Phi\left(\frac{t}{\alpha\left(\lambda\right)\sqrt{n}}\right) = \Phi\left(0\right) + \frac{1}{\sqrt{2\pi}}\left\{\frac{t}{\alpha\left(\lambda\right)\sqrt{n}} + O\left(n^{-3/2}\right)\right\}$$

therefore asymptotically we have

$$\begin{aligned} I_n &\approx \int_0^\infty e^{-t}\left[\Phi\left(\frac{t}{\alpha\left(\lambda\right)\sqrt{n}}\right) - \Phi\left(0\right) + O\left(n^{-3/2}\right)\right]dt \\ &= \int_0^\infty e^{-t}\frac{1}{\sqrt{2\pi}}\left\{\frac{t}{\alpha\left(\lambda\right)\sqrt{n}} + O\left(n^{-3/2}\right)\right\}dt \\ &= \left(\alpha\left(\lambda\right)\sqrt{2\pi n}\right)^{-1} + O\left(n^{-3/2}\right) \end{aligned}$$

Plugging in all the terms we get the final result

$$\Pr\left(\mathrm{R}_n^{\text{emp}}\left(f_\nu\right) \geq \epsilon\right) = \frac{1}{\sqrt{2\pi n}}\exp\left(-n\Lambda_n^*\left(\epsilon\right)\right)\left[\frac{\sqrt{\Lambda_n^{*\prime\prime}\left(\epsilon\right)}}{\Lambda_n^{*\prime}\left(\epsilon\right)} + O\left(n^{-3/2}\right)\right]$$

□

In the literature, usually one fixes the probability (12) say at $\alpha \in [0,1]$ and then estimates the corresponding error $\epsilon = \delta_n^{-1}\left(f,\alpha\right)$ via inversion and is necessary for further analysis. For bounds under sub-Gaussian and sub-Gamma tails derived using Cramer-Chernoff method, such an inverse can be analytically calculated. However, in our formulation we only need to estimate or bound $\Lambda^*_{\nu_n^{\text{emp}}}\left(\epsilon\right)$ accurately, in order to calculate the required Lagrangian functional (6), the prior (9) and the associated bounds on generalization error (10).

Note that even if $\Lambda^*_{\nu_n^{\text{emp}}}\left(\epsilon\right) \approx 0$, we still see δ_n converging to 0 at the rate $\mathcal{O}\left(n^{-1/2}\right)$, i.e. we get the weak law of large numbers, which is not possible to show in a Chernoff-Inequality. Therefore, by optimizing the Lagrangian functional (6) we can ensure models which admits the least amount of estimation error. However, unfortunately in the case of an interpolating model like Deep neural networks, around the optima the empirical risk usually becomes zero and then our data/model dependent correction becomes vacuous. Since when optimized using a stochastic gradient descent algorithm, the generalization error is observed (via out of sample risk) [Zhang et al. (2016)] to decrease, the optimization algorithm has an effect [London (2017)] on the approximation error and hence the η-sufficiency (2) of the prior on the model space.

3.2 Estimating the Moment Generating Function $\Lambda_{\nu_n^{\text{EMP}}}$

In order to estimate $\Lambda_n^*\left(\epsilon\right)$, we need to estimate the moment generating function of the push-forward measure $\lambda_n \coloneqq \mathrm{R}(f)^\sharp \nu_n^{\text{emp}}$. Here we want to show that push-forward measures from i.i.d. observations are necessarily Levy measures on $\mathbb{R}_+$, which allows us to use Levy-Khintchine theorem to estimate the bound. The push forward measure for each function $f \in \mathcal{H}$ i.e. $\mathrm{R}\left(f\right)^\sharp : \mathcal{P}_2\left(\mathbb{R}^d\right) \to \mathcal{R}\left(\mathbb{R}_+\right)$ is a Radon measure on the non-negative real line such that $\forall \nu \in \mathcal{P}_2\left(\mathbb{R}^d\right)$ and for all Borel subsets $A \in \mathcal{B}\left(\mathbb{R}_+\right)$

$$\int_A d\mathrm{R}\left(f\right)^\sharp \nu = \int_{\mathrm{R}(f)^{-1}(A)} f d\nu$$

If the risk functional is unbounded, the push-forward empirical measure $\lambda_n \coloneqq \mathrm{R}(f)^\sharp \nu_n^{\text{emp}}$ need not be a probability measure and can be unbounded on the non-negative real line i.e. $\lambda_n(]0,\infty[) = \infty$. It is easy to see this if the map $\mathrm{R}\left(f\right)$ is not a one to one function. For example, let ν be the probability measure on $\mathbb{R}^2$ and consider the continuous square projection map onto the first dimension $P_1 \colon \mathbb{R}^2 \to \mathbb{R}$. Set $\nu' \coloneqq P_1^\sharp \nu$, then for every $x \in \mathbb{R}$, we have $\nu'\left(x\right) = \nu\left(P_1^{-1}\left(x\right)\right) = \nu\left(\{x\} \cup \{-x\} \times \mathbb{R}\right) = 0$. But for any small open set $U \subset \mathbb{R}$ we get $\nu'\left(U\right) = \nu\left(\{U\} \cup \{-U\} \times \mathbb{R}\right) \geq \nu\left(\{U\} \times \mathbb{R}\right)$. Integrating over all such open subsets, we can see that ν' need not be finite.

Now it is not necessary for this push forward measure λ_n to even have a finite first moment, let alone a moment generating function. In order to ensure its existence, we need to put certain assumptions on λ_n. Let $\left\{r_i\left(f\right) := \mathrm{R}\left(f\right)^{\sharp}\delta_{X_i}\right\}_{i=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \lambda_n$ be a sequence of i.i.d. samples from the push-forward empirical measure. Let λ_n be a general measure on $]0,\infty[$ such that even if

$$\lambda_n(]0,\infty[) = \infty$$

assume for some $\delta > 0$ that

$$\int_{]0,\infty[} \min(\tau, r)\lambda_n(dr) < \infty$$

Observe that the integrability condition 15 implies that for every $\delta > 0$ the tail-intensity $\lambda_n(]\delta,\infty[)$ is finite i.e. there are only finitely many atoms in $]\delta,\infty[$, even when there are infinitely many atoms in $]0,\infty[$ since $\lambda_n(]0,\infty[) = \infty$. Further, the condition 15 implies that the series $\sum_{i=1}^{\infty} r_i \mathbb{I}_{\{L_i \leq \delta\}}$ converges almost surely since

$$\mathbb{E}_{\lambda_n}\left[\sum_{i=1}^{\infty} r_i \mathbb{I}_{\{r_i \leq \delta\}}\right] = \int_{]0,\delta[} r d\lambda_n(l) < \infty$$

and as there are only finitely many atoms in $]\delta,\infty[$, we have

$$\sum_{i=1}^{\infty} r_i < \infty$$

almost surely. Conversely, the series $\sum_1^{\infty} r_i$ diverges almost surely whenever the integral condition fails. This means that the above integral condition 15 provides the necessary and sufficient condition for the infinite sum to be finite for any push-forward measure λ_n.

Further as we shall see now, it also characterizes the condition under which the moment generating function exists. Consider an independent sequence $U_1, U_2...$ of i.i.d. uniform variables on $[0,1]$, and define the purely discontinuous increasing process

$$\zeta(t) := \sum_{i=1}^{\infty} r_i\left(f\right)\mathbb{I}_{\{U_i \leq t\}} = \sum_{U_i \leq t} r_i\left(f\right), \quad t \in [0,1] \tag{13}$$

Clearly the increasing process $(\zeta(t), 0 \leq t \leq 1)$ has independent and stationary increments. This means that for every $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$, the variables $\zeta(t_1) - \zeta(t_0), ..., \zeta(t_{n+1}) - \zeta(t_n)$ are independent, and $\zeta(t_{i+1}) - \zeta(t_i)$ has the same law as $\zeta(t_{i+1} - t_i)$. Such a stochastic process $\zeta(t)$ is known as a subordinator on the time interval $[0,1]$, and is an almost surely non-decreasing Levy process due to the following characterization given by the Levy-Khintchine formula.

Theorem 2. *Levy–Khintchine Formula [Bertoin (2006)]: The Laplace-Stieltjes transform of a subordinator on $\mathbb{R}_+$ has a unique representation of the form*

$$\Lambda_{\lambda_n}\left(\gamma\right) := \log \mathbb{E}\left[\exp\left(-\gamma\zeta(t)\right)\right] = -t\int_{\mathbb{R}_+}\left[1 - e^{-\gamma r}\right]\lambda_n\left(dr\right) \tag{14}$$

if and only if the push-forward measure λ_n on $\mathbb{R}_+$ satisfies, for some $\tau > 0$

$$\int_{]0,\infty[} min(\tau, r)\lambda_n(dr) < \infty \tag{15}$$

Then λ_n is known as a Levy measure.

Therefore Levy-Khintchine theorem (14) implies that $\zeta(t)$ is a subordinator if and only if λ_n, the push-forward measure defined on $\mathbb{R}_+$ is a Levy measure. Many of the well studied measures belong to the family of Levy measures, including Compound Poisson Processes, Gaussian Processes, Gamma Processes, Stable Processes, etc. Hence even if the cumulant generating function of the original measure does not exist, if the push-forward measure satisfies (15), then it's corresponding cumulant generating function exists.

3.3 APPLICATIONS

Now we show two examples with square error loss but different noise characteristics, where our method leads to consistent estimators even though the risk functional may not be bounded.

Gamma subordinators The Gamma subordinator is suitable when the tail of the push-forward loss measure decreases exponentially fast. This is the case for square error loss, with normally distributed noise. Since sum of square normal variables leads to chi-squared variates, the resulting push-forward loss can be modeled via a Gamma subordinator. Let there exist two fixed real numbers $\theta, c > 0$, such that the Levy measure given by $d\lambda_n(x) = \theta x^{-1} e^{-cx} dx$ fits the push-forward measure well. Then its moment generating function is given by $\Lambda_{\nu_n^{\text{emp}}}(\gamma) = -\theta \log(1 + \gamma/c), \gamma \geq 0$ and hence the corresponding bound on estimation error is given by

$$\Pr\left(\mid \mathrm{R}_{\nu_n^{\text{emp}}}(f) - \mathrm{R}_{\nu}(f) \mid \geq \epsilon\right) \leq \frac{e^{-n\theta\left(\log \frac{c}{\theta} - 1\right)}}{\sqrt{2\pi n}} \frac{\sqrt{\theta}}{(c\epsilon - \theta)} \exp\left(-n\left(c\epsilon - \theta \log(\epsilon)\right)\right) \quad \text{for } \epsilon > \frac{\theta}{c}$$

Stable Subordinators The Stable subordinator is suitable when the tail of the push-forward loss measure decreases polynomially fast. This is the case for square error loss, with heavy tailed noise. Let $\alpha \in]0, 1[$ and $c > 0$ be fixed parameters, such that the Levy measure given by $d\lambda_n(x) = \frac{c\alpha}{\Gamma(1-\alpha)} x^{-1-\alpha} dx$ fits the push-forward measure well. Then its moment generating function is given by $\Lambda_{\nu_n^{\text{emp}}}(\gamma) = -c\gamma^{\alpha}, \gamma \geq 0$ and hence the corresponding bound on estimation error is given by

$$\Pr\left(\mid \mathrm{R}_{\nu_n^{\text{emp}}}(f) - \mathrm{R}_{\nu}(f) \mid \geq \epsilon\right) \leq \frac{1}{\sqrt{2\pi\alpha n}} \frac{1}{\sqrt{\Lambda^*_{\nu_n^{\text{emp}}}(\epsilon)}} \exp\left(-n\Lambda^*_{\nu_n^{\text{emp}}}(\epsilon)\right)$$

where $\Lambda^*_{\nu_n^{\text{emp}}}(\epsilon) = \epsilon^{-\frac{\alpha}{1-\alpha}} \left[(\alpha c)^{\frac{1}{1-\alpha}} - c(\alpha c)^{\frac{\alpha}{1-\alpha}}\right]$. Therefore, even in the case of unbounded losses and small samples, we can estimate a bound on the estimation error for push-forward measures with both exponentially and polynomially decreasing tails.

4 CONCLUSIONS

In this work, we designed a step by step framework to construct well posed model estimation algorithms by ensuring sufficiency, consistency and uniqueness of the optimization problem. We showed that the PAC-Bayes approach admits a unique Gibbs measure, if the risk functional can be assumed to have Lipschitz continuous derivatives. We approached the problem of their unboundedness by first observing that a risk functional pushes forward the empirical measure to a Radon measure on the non-negative real line which is necessarily a Levy measure. Hence, the only necessary and sufficient condition for a resulting PAC-Bayes bound (10) to exist was shown to be an integrability condition (15) derived from the Levy-Khintchine theorem. Further, we derived a Bahadur-Rao type exact asymptotic bound, which is much sharper than a traditional Chernoff type inequality, especially in the under-sampled regime. Such a construction allowed us to derive data or model-dependent consistency promoting updates to a data-free prior, which provably improve the generalization performance. As examples we showed that if the push-forward measure can be modeled either by a Gamma subordinator for exponentially decreasing tail or Stable subordinator for polynomially decreasing tail we can derive the corresponding PAC-Bayes bounds.

REFERENCES

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine learning*, 107(5):887–902, May 2018. URL https://doi.org/10.1007/s10994-017-5690-0, 3.1

Amiran Ambroladze, Emilio Parrado-hernández, and John S Shawe-taylor. Tighter PAC-Bayes bounds. In B Schölkopf, J C Platt, and T Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 9–16. MIT Press, 2007. URL http://papers.nips.cc/paper/3058-tighter-pac-bayes-bounds.pdf 1, 2.2, 3.1

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. February 2018. URL http://arxiv.org/abs/1802.05296. 1, 3.1

Jean-Yves Audibert and Olivier Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of machine learning research: JMLR*, 8:863–889, May 2007. URL https://dl.acm.org/doi/10.5555/1248659.1248691. 1, 2.2, 3.1

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. *undefined*, 2014. URL https://www.semanticscholar.org/paper/dde1d8e12fac6ce3fcb0b4a4c94e70a5b259a391. 2.2

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. February 2018. URL http://arxiv.org/abs/1802.01396. 1

Jean Bertoin. *Random Fragmentation and Coagulation Processes*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, August 2006. URL https://play.google.com/store/books/details?id=yndbFG6medoC. 1, 1, 2

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. URL https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780199535255.001.0001/acprof-9780199535255. 3.1, 3.1

Damian Brzyski, Christine B Peterson, Piotr Sobczyk, Emmanuel J Candès, Malgorzata Bogdan, and Chiara Sabatti. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1):61–75, January 2017. URL http://dx.doi.org/10.1534/genetics.116.193987. 1

Olivier Catoni. Pac-Bayesian supervised classification: The thermodynamics of statistical learning. December 2007. URL http://arxiv.org/abs/0712.0248. 1, 2.1, 2.1, 2.2, 2.2

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, November 2009. URL https://play.google.com/store/books/details?id=d3nnjwEACAAJ. 1, 3.1

Sophie Donnet, Vincent Rivoirard, Judith Rousseau, and Catia Scricciolo. Posterior concentration rates for empirical bayes procedures, with applications to dirichlet process mixtures. June 2014. URL http://arxiv.org/abs/1406.4406. 2.1

M D Donsker and S R S Varadhan. Asymptotic evaluation of certain markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, January 1975. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160280102. 2.2

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. March 2017a. URL https://www.semanticscholar.org/paper/4f8a87535794576043803de919cbcd955cdc92f6. 1

Gintare Karolina Dziugaite and Daniel M Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. December 2017b. URL http://arxiv.org/abs/1712.09376. 1

Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In H Wallach, H Larochelle, A Beygelzimer, F dAlché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6729–6739. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8898-hypothesis-set-stability-and-generalization.pdf. 1

Yoav Freund. Self bounding learning algorithms. In *Proceedings of the eleventh annual conference on Computational learning theory - COLT' 98*, pages 247–258, New York, New York, USA, 1998. ACM Press. URL http://portal.acm.org/citation.cfm?doid=279943.279993. 2.2

James C Fu. Large sample point estimation: A large deviation theory approach. *Annals of statistics*, 10(3):762–771, 1982. URL http://www.jstor.org/stable/2240902. 2.1

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, June 2009. Association for Computing Machinery. URL https://doi.org/10.1145/1553374.1553419. 2.2

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. March 2015. URL http://arxiv.org/abs/1503.08329. 2.2

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets bayesian inference. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1884–1892. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6569-pac-bayesian-theory-meets-bayesian-inference.pdf. 1, 2.2

Samuel J Gershman and David M Blei. A tutorial on bayesian nonparametric models. June 2011. URL http://arxiv.org/abs/1106.2697. 1, 2.1

Peter D Grünwald and Nishant A Mehta. A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity. volume 98 of *Proceedings of Machine Learning Research*, pages 433–465, Chicago, Illinois, 2019. PMLR. URL http://proceedings.mlr.press/v98/grunwald19a.html. 1

Benjamin Guedj. A primer on PAC-Bayesian learning. January 2019. URL http://arxiv.org/abs/1901.05353. 2.2, 2.2

Shota Gugushvili, Ester Mariucci, and Frank van der Meulen. Decompounding discrete distributions: A non-parametric bayesian approach. March 2019. URL http://arxiv.org/abs/1903.11142. 2.1

Matthew Holland. PAC-Bayes under potentially heavy tails. In H Wallach, H Larochelle, A Beygelzimer, F dAlché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2715–2724. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8539-pac-bayes-under-potentially-heavy-tails.pdf. 3.1

Matti Kääriäinen and John Langford. A comparison of tight generalization error bounds, 2005. URL http://dx.doi.org/10.1145/1102351.1102403. 3.1

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. October 2017. URL http://arxiv.org/abs/1710.05468. 1

John Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, April 1967. URL https://msp.org/pjm/1967/21-1/p06.xhtml. 2.1

B J K Kleijn and Y Y Zhao. Criteria for posterior consistency. August 2013. URL http://arxiv.org/abs/1308.1263. 1, 2.1, 2.1

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Rätsch. Boosting variational inference: an optimization perspective. August 2017. URL http://arxiv.org/abs/1708.01733. 1

Ben London. A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2931–2940. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6886-a-pac-bayesian-analysis-of-randomized-learning-with-application-to-stochastic.pdf. 3.1

David McAllester. A PAC-Bayesian tutorial with a dropout bound. July 2013. URL http://arxiv.org/abs/1307.2118. 1

David A McAllester. Some PAC-Bayesian theorems. *Machine learning*, 37(3):355–363, December 1999. URL https://doi.org/10.1023/A:1007618624809. 1, 2.2, 3.1

Andrew McDavid, Raphael Gottardo, Noah Simon, and Mathias Drton. GRAPHICAL MODELS FOR ZERO-INFLATED SINGLE CELL GENE EXPRESSION. *The annals of applied statistics*, 13(2):848–873, June 2019. URL http://dx.doi.org/10.1214/18-AOAS1213. 1

Marina Meila. Learning with mixtures of trees. *Journal of machine learning research: JMLR*, 1: 1–48, 2000. 1

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected bernstein inequality. In H Wallach, H Larochelle, A Beygelzimer, F dAlché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 12202–12213. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9387-pac-bayes-un-expected-bernstein-inequality.pdf. 3.1

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in computational mathematics*, 25:161–193, 2006. 1, 2.1

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to Spectrally-Normalized margin bounds for neural networks. February 2018. URL https://openreview.net/pdf?id=Skz_WfbCZ. 1

Xuanlong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. (1): 370–400, September 2011. URL http://arxiv.org/abs/1109.3250. 1

Neal Parikh, Stephen P Boyd, and Others. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014. URL https://web.stanford.edu/~boyd/papers/pdf/prox_algs.pdf. 1, 2.1, 2.2

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes bounds with data dependent priors. *Journal of machine learning research: JMLR*, 13(112): 3507–3531, 2012. URL https://www.jmlr.org/papers/v13/parrado12a.html. 1

Sonia Petrone and Piero Veronese. Non parametric mixture priors based on an exponential random scheme. *Statistical methods & applications*, 11(1):1–20, February 2002. URL https://doi.org/10.1007/BF02511443. 1

Harish G Ramaswamy. Mixture proportion estimation via kernel embedding of distributions. 1

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984. URL http://www.jstor.org/stable/2030064. 1

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvari, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. June 2020. URL http://arxiv.org/abs/2006.13057. 1

Matthias Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research: JMLR*, 3(Oct):233–269, 2002. URL http://www.jmlr.org/papers/v3/seeger02a.html. 2.1

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. March 2017. URL http://arxiv.org/abs/1703.00810. 1, 2.1

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences of the United States of America*, 116(29):14516–14525, July 2019. URL http://dx.doi.org/10.1073/pnas.1810420116. 1

Ilya O Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein inequality. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 109–117. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4903-pac-bayes-empirical-bernstein-inequality.pdf. 3.1

Vladimir Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. 2000. URL https://www.springer.com/de/book/9780387987804. 1, 2

Yuefeng Wu and Subhashis Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, (0):298–331, October 2007. URL http://arxiv.org/abs/0710.2746. 1, 2.1

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. November 2016. URL http://arxiv.org/abs/1611.03530. 1, 3.1

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-Vacuous generalization bounds at the ImageNet scale: A PAC-Bayesian compression approach. April 2018. URL http://arxiv.org/abs/1804.05862. 1

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. March 2018. URL http://arxiv.org/abs/1803.06573. 2.2