Integral Probability Metrics PAC-Bayes Bounds

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a PAC-Bayes-style generalization bound which enables the replacement of the KL-divergence with a variety of *Integral Probability Metrics* (IPM). We provide instances of this bound with the IPM being the *total variation metric* and the *Wasserstein distance*. A notable feature of the obtained bounds is that they naturally interpolate between classical uniform convergence bounds in the worst case (when the prior and posterior are far away from each other), and preferable bounds in better cases (when the posterior and prior are close). This illustrates the possibility of reinforcing classical generalization bounds with algorithm- and data-dependent components, thus making them more suitable to analyze algorithms that use a large hypothesis space.

1 Introduction and Related Work

Classical statistical learning theory is based on a worst-case perspective which can be too pessimistic to model practical machine learning. In reality, data is rarely worst-case, and experiments demonstrate learning tasks that are solved with much less data than predicted by traditional theory. A primary manifestation of the traditional worst-case perspective is demonstrated by *uniform convergence* (UC); a genre of generalization bounds which form the backbone of the classical theory (Vapnik, 1999). These bounds guarantee that the generalization gap of *all* hypotheses in the output-space of the algorithm *simultaneously* vanish as the training-set size grows. The key algorithmic insight these bounds provide is summarized by the *Empirical Risk Minimization* principle (ERM), which asserts that it suffices to output *any* hypothesis in the class which minimizes the empirical risk.

Consequently, UC arguments provide non-trivial guarantees only if the hypothesis class used by the algorithm is *restricted* (e.g. has low Rademacher complexity or bounded VC dimension). In contrast, practical learning approaches such as deep learning algorithms use huge hypothesis classes whose VC dimensions rapidly increase with the size and depth of the underlying network. Hence, the rate guaranteed by UC arguments is often much slower than the rate observed in practice (Bachmann, Moosavi-Dezfooli, & Hofmann, 2021; Nagarajan & Kolter, 2019; Neyshabur, Bhojanapalli, McAllester, & Srebro, 2017; Zhang, Bengio, Hardt, Recht, & Vinyals, 2017). A further shortcoming of UC bounds, and the associated ERM principle, is that they are algorithm- and data-independent;[1] that is, they do not utilize beneficial properties of the the data and/or the algorithm. For example, in practice, regularized algorithms often perform better than Empirical Risk Minimization, but this cannot be expressed by UC bounds and the ERM principle.

The PAC-Bayes (PB) framework is a prominent example of a theoretical framework that does not require the UC property. This framework was pioneered by Shawe-Taylor and Williamson (1997) and McAllester (1998) and developed in later papers, e.g. (Catoni, 2007; Lever, Laviolette, & Shawe-Taylor, 2013; Maurer, 2004; McAllester, 2003; Seeger, 2002); see Guedj (2019) and Alquier (2021) for extensive surveys. Begin, Germain, Laviolette, and Roy (2016) introduced a general

[1] More precisely, UC bounds only depend on the hypothesis space.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

strategy that produces PB bounds from change-of-measure inequalities leading to bounds based on the Rényi's α-divergence, and Ohnishi and Honorio (2021) further extended PB bounds to other Csiszár's f-divergences.

PB theorems consider the generalization performance of stochastic (Gibbs) predictors. These bounds are non-uniform by nature, and are algorithm and data-dependent. They are usually based on a complexity term that depends on the Kullback-Leibler (KL) divergence between a data-dependent posterior distribution and a data-independent prior distribution.[2] Recent work by Aminian, Bu, Wornell, and Rodrigues (2022); Rodríguez Gálvez, Bassi, Thobaben, and Skoglund (2021); Wang, Diaz, Santos Filho, and Calmon (2019) and Neu and Lugosi (2022) proved information-theoretic bounds on the *expected* generalization gap using the Wasserstein and the total-variation (TV) distances. Our work is within the PB framework, and therefore enjoys the following advantages: *(i)* The bounds are "in high probabily" rather than in expectation. *(ii)* PB bounds are sample dependent, i.e., bound the generalization gap for a specific sample-dependent posterior, while information-theoretic bounds are formulated as expectation over all sample sets, thereby providing a basis for empirical algorithms, e.g., Alquier (2021); Dziugaite and Roy (2017). *(iii)* The reference measure in PB it can be any sample-independent distribution, while information-theoretic bounds consider a specific reference. Our work introduces uniform convergence assumptions, while the above-mentioned papers each used different assumptions.

In the modern deep learning regime, measures of the hypothesis class complexity used in UC bounds, such as the VC dimension or Rademacher complexity, are enormous, making the bounds extremely loose for any reasonable number of samples. However, these complexity measures often have closed-form formulas for models such as neural networks, which show explicitly the effect of the model architecture (number of layer, activation functions etc.). This in contrast to PB bounds, in which the dependence on the hypothesis class is less explicit (but see Anthony and Bartlett (1999); Neyshabur, Tomioka, and Srebro (2015) for exceptions for neural networks). Therefore, we believe that extension of UC bounds to incorporate data- and algorithm-dependence can facilitate the design of better performing architectures. A further advantage of PB bounds is their non-uniformity (the generalization gap bound depends on the learning output), hence we can use the bound as a minimization objective for a structural minimization algorithm, where the complexity term acts as a regularizer. In cases where the hypothesis class is very large, but we have some prior knowledge on which hypotheses are more likely to have low population loss (e.g. prefer simpler hypothesis as suggested by Occam's Razor), then in PB one can inject this knowledge as the prior distribution, effectively lowering the generalization bound.

Can the rich theory of UC bounds be extended to help explain generalization with modern large scale models? Can this theory be used to prove data and algorithm dependent guarantees? In this paper, we take a step in the direction of positive answers to these questions. To achieve this goal, we show a new technique to incorporate UC bounds within the PAC-Bayes framework. We prove new PB bounds with Integral Probability Metric (IPM) (Müller, 1997) to measure distances between distributions, rather than the standard KL or f-divergences used so far. This IPM framework allows greater flexibility, as it does not require the support of the posterior to be a subset of the prior's support (absolute continuity) as in standard PB bounds, and it applies to deterministic as well as stochastic prior and posterior distributions. Specifically, we focus on utilizing two specific IPMs: the total variation and Wasserstein metrics

The Total Variation PAC-Bayes (TVPB) bound (Thm. 6) applies in any setting where uniform convergence (UC) (Def. 1) holds, an assumption that is satisfied by many natural learning problems. For example, in binary problems, UC holds with a rate of $O(\sqrt{\mathtt{VC}(\mathcal{H})/m})$, where $\mathtt{VC}(\mathcal{H})$ is the VC dimension (Vapnik, 1999). As observed in practice, for large models of deep neural networks with very large VC dimensions the learning rate on natural datasets is often much faster than predicted UC bounds. To explain this gap, we must turn to data and algorithm-dependent bounds. The TVPB improves the gap bound to be $O\left(\sqrt{\mathtt{VC}(\mathcal{H})D_{\mathrm{TV}}(Q,P)/m}\right)$, effectively multiplying the VC dimension by the total variation distance of the posterior from the prior. Intuitively, simpler posteriors (closer to the prior assumed before observing the data) lead to a better generalization gap. Compared to the vanilla KL-PB, the TV distance can be small even in cases where the KL distance can be very large, and in any case, the bound only improves over the original UC bound. In addition, the TV-PB bounds incorporate properties of $\mathcal{H}$ via the VC dim. We also explore settings where the generalization gap

[2] But see Rivasplata, Kuzborskij, Szepesvári, and Shawe-Taylor (2020) for data-dependent priors.

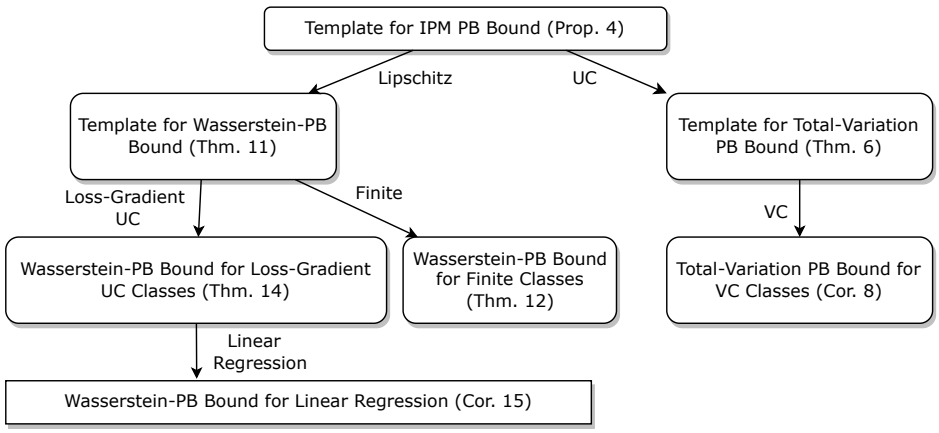

Figure 1: Claims tree diagram. $A \to B$ means that claim B is a special case of claim A, under additional assumptions on the learning problem. For full description of the assumptions, see the corresponding claims.

function exhibits a certain smoothness property, and show a PB bound with the Wasserstein metric (Thm. 11), We analyze this smoothness property and show an explicit Wasserstein-based bound in the finite hypothesis class setting and in a linear regression setting. In the latter setting, we show that a standard UC bound can be improved by a factor of $O(\sqrt{W(Q,P)})$, where $W(Q,P)$ is the Wasserstein distance between posterior and prior over the unit-sphere. Figure 1 illustrates graphically the organization of the claims in the paper.

2 Preliminaries

2.1 The Learning Problem

We begin with a short description a standard supervised learning task. Consider a domain $\mathcal{Z}$, [3] a distribution $\mathcal{D}$ over $\mathcal{Z}$, a hypothesis set $\mathcal{H}$, and finally, a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$. The tuple $(\mathcal{D}, \mathcal{H}, \ell)$ defines a learning problem: The learning algorithm receives as input a training set $S = \{z_i\}_{i=1}^m \in \mathcal{Z}^m$ sampled i.i.d from $\mathcal{D}$ and selects an hypothesis $h \in \mathcal{H}$. The performance of h is measured by the *expected risk*, $L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathbb{E}_{z\sim\mathcal{D}}\ell(h,z)$. While the expected risk is unavailable to algorithm, the *empirical risk*, $L_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m \ell(h,z_i)$, can be evaluated using the training data. The generalization gap is defined by $\Delta_S(h) \stackrel{\text{def}}{=} L_{\mathcal{D}}(h) - L_S(h)$.

Several of our results will assume that the learning problem $(\mathcal{D}, \mathcal{H}, \ell)$ satisfies the *uniform convergence property*; i.e. the existence of a uniform upper bound on the generalization gap which applies simultaneously to all hypotheses in $\mathcal{H}$.

Definition 1 (Uniform convergence, (Vapnik & Chervonenkis, 2015))**.** *The learning problem $(\mathcal{D}, \mathcal{H}, \ell)$ satisfies the uniform convergence property, if there exists a bound function $u(m,\delta') > 0$ s.t. for any $m \in \mathbb{N}^+$, $\delta \in (0,1)$ we have*

$$\mathbb{P}\{|\Delta_S(h)| \le u(m,\delta), \forall h \in \mathcal{H}\} \ge 1-\delta \quad \textit{and} \quad u(m,\delta) \underset{m\to\infty}{\to} 0. \tag{1}$$

UC type bounds are a major part of the foundations of theoretical machine learning. Unfortunately, they suffers from a few drawbacks. First, currently known bounds tend to be extremely loose in many cases, most notably for deep networks. Second, the setup does not provide a natural way to encode prior knowledge into the bounds, particularly when dealing with deep networks - the hypothesis set is usually rich enough to express all relevant functions, and the training algorithms that might utilize some prior knowledge are not themselves a part of the PAC framework. Finally, PAC bounds are

[3] This formulation allows for greater generality than the standard $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and mis-classification loss setting. In particular, it can describe a number of unsupervised learning problems (Seldin and Tishby (2010)).

usually not data-dependent, a property which is critical to explain the generalization of DNNs on real-world data (Nagarajan & Kolter, 2019; Zhang et al., 2017).

2.2 PAC-Bayes Bounds

Let $\mathcal{M}(\mathcal{H})$ denote the set of all probability measures over $\mathcal{H}$. For any probability measure $Q \in \mathcal{M}(\mathcal{H})$, we define the expected loss, empirical loss and generalization gap by

$$L_{\mathcal{D}}(Q) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{h\sim Q} L_{\mathcal{D}}(h) \quad ; \quad L_S(Q) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{h\sim Q} L_S(h) \quad ; \quad \Delta_S(Q) \stackrel{\text{def}}{=} L_{\mathcal{D}}(Q) - L_S(Q). \tag{2}$$

PAC-Bayes theory bounds the expected loss, simultaneously for all "posterior" (sample-dependent) probability measures $Q \in \mathcal{M}(\mathcal{H})$, with high probability over the samples $S \sim \mathcal{D}^m$, given any "prior" (sample-independent) probability measure $P \in \mathcal{M}(\mathcal{H})$. A key feature of most PAC-Bayes bounds is their dependence on the KL divergence between the two distributions P, Q, $\mathtt{KL}(Q \parallel P) \stackrel{\text{def}}{=} \int_{\mathcal{H}} \ln(\frac{\mathrm{d}Q}{\mathrm{d}P})\mathrm{d}Q$, where $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is the Radon–Nikodym derivative of Q w.r.t. P. While KL is a natural measure of divergence between probability distributions, it restricts the applicability of the resulting bounds to cases where the support of Q is contained in the support of P. The following bound was introduced by McAllester (1998).

Proposition 2 (Classical PB Bound)**.** *For any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over the samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\Delta_S(Q) \leq \sqrt{\frac{\mathtt{KL}(Q \parallel P) + \ln(m/\delta)}{2(m-1)}}. \tag{3}$$

3 A Template for IPM PAC-Bayes Bounds

3.1 General IPM-PB Bound

Definition 3 (Integral Probability Metric)**.** *(Müller, 1997; Sriperumbudur, Fukumizu, Gretton, Schölkopf, & Lanckriet, 2009, 2012) The Integral Probability Metric (IPM) between two probability measures P and Q over $\mathcal{H}$ is defined as*

$$\gamma_{\mathcal{F}}(Q,P) \stackrel{\text{def}}{=} \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{H}} f \mathrm{d}P - \int_{\mathcal{H}} f \mathrm{d}Q \right| = \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{h\sim P}[f_S(h)] - \mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] \right| \tag{4}$$

where $\mathcal{F}$ is a set of real-valued bounded functions $\mathcal{H} \to \mathbb{R}$.

By definition, IPM distance measures are symmetric and non-negative. Note that the KL-divergence is not a special case of IPM, rather it belongs to the family of f-divergences, that intersect with IPM only at the Total-Variation (Sriperumbudur et al., 2009, 2012).

The following proposition assumes that for any fixed sample $S' \in \mathcal{Z}^m$, the function $f_{S'}(h) \stackrel{\text{def}}{=} 2(m-1)\Delta^2_{S'}(h)$ is a member of a family of functions that depend on the sample, denoted $\mathcal{F}_{S'}$. Thus, the IPM-PB bound allows us to 'convert' some knowledge we have about the properties of the generalization gap function $\Delta_S(h)$ to a generalization bound.

Since we do not specify yet the family of functions $\mathcal{F}_S$, the bound does not convey an explicit rate, and it should rather be seen as a **template**. In the next sections, we will derive explicit bounds with specific IPMs divergences. Namely, we will derive a total-variation distance based bound by selecting a family of bounded functions, and a Wasserstein distance based bound, by selecting a family of Lipschitz functions.

Proposition 4 (Template for IPM PB Bound)**.** *For any fixed dataset $S' \in \mathcal{Z}^m, m \in \mathbb{N}^+$, let $\mathcal{F}_{S'}$ be a family of bounded and measurable functions $\mathcal{H} \to \mathbb{R}$. Assume that for any number of samples, m, and sample $S' \in \mathcal{Z}^m$, the function $2(m-1)\Delta^2_{S'}(\cdot)$ is in $\mathcal{F}_{S'}$. Then, for any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$,*

$$\Delta_S(Q) \leq \sqrt{\frac{\gamma_{\mathcal{F}_S}(Q,P) + \ln(m/\delta)}{2(m-1)}}. \tag{5}$$

The proof is in Appendix A.1. The main idea is to use the IPM definition and the assumption as a change-of-measure inequality, instead of the variational formula by Donsker and Varadhan (1975), which is used in the classical KL-PB bound proof. The rest of the proof is similar to the classical derivation (McAllester, 2003; Shalev-Shwartz & Ben-David, 2014).

Note that, similarly to the classical KL-PB bound, Proposition 4 does not require the UC property to hold. However, in the next sections we will see that assuming an existence of a UC bound $u(m, \delta)$, and selecting a particular function family $\mathcal{F}_S$, will result in explicit bounds that can improve upon the worst-case nature of the original $u(m, \delta)$ bound.

3.2 Template for Seeger Type IPM PAC-Bayes Bound

The work of Seeger (2002) and Maurer (2004) presented a different form of the PAC-Bayes theorem with fast $O(1/m)$ rate as the dominant term, if the empirical risk is low. Let $\mathtt{kl}(p \parallel q)$ be the KL divergence between two Bernoulli distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$, $0 < p, q < 1$, that is, $\mathtt{kl}(p \parallel q) \stackrel{\text{def}}{=} p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right)$. For any $h \in \mathcal{H}$, define the relative entropy of the empirical risk with respect to the expected one as $\Delta_S^{\mathtt{kl}}(h) \stackrel{\text{def}}{=} \mathtt{kl}(L_S(h) \parallel L_D(h))$, Similarly, and overloading notations, we define for any distribution $Q \in \mathcal{M}(\mathcal{H})$, $\Delta_S^{\mathtt{kl}}(Q) \stackrel{\text{def}}{=} \mathtt{kl}(L_S(Q) \parallel L_D(Q))$.

By replacing the KL-based change-of-measure inequality step in the proof of (Maurer, 2004, Thm. 5) with the IPM property (Def. 3) we get a similar bound for IPM measures (see a detailed proof in Appendix A.2).

Proposition 5 (Template for Seeger Type IPM PAC-Bayes Bound)**.** *Assume $f_S(h) \stackrel{\text{def}}{=} m \cdot \Delta_S^{\mathtt{kl}}(h) \in \mathcal{F}_S$. We have that with probability at least $1-\delta$ over the samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$*

$$\Delta_S^{\mathtt{kl}}(Q) \leq \frac{\gamma_{\mathcal{F}_S}(Q, P) + \ln(2\sqrt{m}/\delta)}{m}. \tag{6}$$

By applying the technique in Eq. 6 of McAllester (2003) we can immediately derive the following, looser, but easier to interpret bound

$$\Delta_S(Q) \leq \sqrt{2L_S(Q)\frac{\gamma_{\mathcal{F}_S}(Q, P) + \ln(2\sqrt{m}/\delta)}{m}} + 2\frac{\gamma_{\mathcal{F}_S}(Q, P) + \ln(2\sqrt{m}/\delta)}{m}. \tag{7}$$

When $L_S(Q)$ is small (as is typical with modern deep networks), the final term determines the convergence rate. We defer the investigation of PB bounds derived from Prop. 5 to Appendix B. In the following sections we focus on investigating the implication of the Template IPM PB Bound of Prop. 4.

4 Total-Variation PAC-Bayes Bounds

In this section, we investigate a PB bound with the total-variation (TV) distance, $D_{\mathrm{TV}}(Q, P) \stackrel{\text{def}}{=} \sup_{A \in \sigma_{\mathcal{H}}} |P(A) - Q(A)|$, where $\sigma_{\mathcal{H}}$ is the standard Borel sigma-algebra associated with $\mathcal{H}$. The TV distance can be described as an IPM with the family of functions

$$\mathcal{F}_M^{\infty} \stackrel{\text{def}}{=} \{f : \mathcal{H} \to \mathbb{R}, \|f\|_{\infty} \leq M\} \tag{8}$$

for any $M \geq 0$. To see this, note that

$$\gamma_{\mathcal{F}_M^{\infty}}(Q, P) = \sup_{f \in \mathcal{F}_M^{\infty}} \left| \int_{\mathcal{H}} f \mathrm{d}P - \int_{\mathcal{H}} f \mathrm{d}Q \right| \stackrel{(i)}{=} M \cdot \sup_{A \in \sigma_{\mathcal{H}}} |P(A) - Q(A)| = M D_{\mathrm{TV}}(Q, P) \tag{9}$$

where equality *(i)* holds since in the supremum it suffices to take the class of indicator functions $\{M \mathbb{1}_A(h), A \in \sigma_{\mathcal{H}}\}$.

Theorem 6 (Template for Total-Variation PB Bound)**.** *Assume that there exists some uniform convergence bound $u(m, \delta')$ (Definition 1), then, for any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0, 1)$, with probability of at least $1-\delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\Delta_S(Q) \leq \sqrt{u^2(m, \delta/2) D_{\mathrm{TV}}(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{10}$$

The proof (Appendix A.3) follows directly from the general IPM PB bound (Prop. 4) and the uniform convergence assumption, using a union bound argument. This bound can be seen as a template to be used to derive explicit PB bounds, by plugging in existing UC bounds. Note that while we require the existence of UC bound, the resulting bound is nonuniform (since it depends on the data-dependent posterior).

Compared to the original UC bound, $u(m, \delta)$, the bound in (10) is roughly multiplied by a factor of $\sqrt{D_{\mathrm{TV}}(Q, P)} \in [0, 1]$, ensuring tighter guarantees, especially if the posterior is close to the prior.

For example, consider a binary classification case, with the the zero-one loss function and $\mathtt{VC}(\mathcal{H})$ class $\mathcal{H}$. The well-known UC theorem states that the generalization gap converges uniformly at a rate $O(\sqrt{\mathtt{VC}(\mathcal{H})/m})$.

Proposition 7 (VC Bound, Boucheron, Bousquet, and Lugosi (2005))**.** *There exists some universal constant $c > 0$ s.t. for any $\delta \in (0, 1)$ we have*

$$\mathbb{P}\left\{\Delta_S(h) \leq c\sqrt{\frac{\mathtt{VC}(\mathcal{H}) + \ln(1/\delta)}{m}}, \forall h \in \mathcal{H}\right\} \geq 1 - \delta \tag{11}$$

Using Thm. 6, we derive the following algorithm and data-dependent bound.

Corollary 8 (Total-Variation PB Bound for VC Classes)**.** *Consider a binary classification problem, with the zero-one loss, and hypothesis class $\mathcal{H}$, with finite VC dimension, $\mathtt{VC}(\mathcal{H})$. There exists some universal constant $c > 0$ s.t. for any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\Delta_S(Q) \leq \sqrt{c\frac{\mathtt{VC}(\mathcal{H}) + \ln(1/\delta)}{m} D_{\mathrm{TV}}(Q, P) + \frac{\ln(m/\delta)}{2(m-1)}}. \tag{12}$$

Compared to the UC bound of Prop. 7, Cor. 8 multiplies the dominant term of the bound by a nonuniform (data and algorithm-dependent) factor of $\sqrt{D_{\mathrm{TV}}(Q, P)}$, which is guaranteed to tighten the bound.

Note that the total-variation based bound of Aminian et al. (2022) and Rodríguez Gálvez et al. (2021) assume Lipschitz loss function, while our TV bound allows non continuous loss functions such as the zero-one loss. The TV based bound of Wang et al. (2019) (Thm. 1) is not directly comparable, since the empirical risk term is multiplied by a factor that goes to infinity for TV distance that goes to 1.

5 Wasserstein PAC-Bayes Bounds

5.1 Template for Wasserstein-PB Bound

In this section, we provide a PAC-Bayes generalization bound with the Wasserstein metric between posterior and prior and a certain smoothness parameter of the generalization gap function. We explore learning settings for which $\mathcal{H}$ can be paired with a distance metric $\rho : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_{\geq 0}$ s.t. $(\mathcal{H}, \rho)$ is a Polish metric space (complete, separable metric space).[4] Given the distance metric ρ, we can define the Wasserstein distance between any two probability measures on $\mathcal{H}$.

Definition 9 (Wasserstein Distance)**.** *For any two probability measures P, Q on $\mathcal{H}$ with finite first moment, the 1st order Wasserstein distance is*

$$W(Q, P) \stackrel{def}{=} \inf_{\gamma \in \Gamma(Q,P)} \int_{\mathcal{H}\times\mathcal{H}} \rho(h, h')\mathrm{d}\gamma(h, h'), \tag{13}$$

where $\Gamma(Q, P)$ denotes the set of all couplings of Q and P, that is, the set of all joint measure on $\mathcal{H} \times \mathcal{H}$ whose marginals are Q and P.

The following proposition gives a dual representation for the first-order Wasserstein distance.

[4] See Villani (2006) Ch. 1, for a discussion of this assumption.

Proposition 10 (Kantorovich-Rubinstein Duality (Villani, 2006))**.** *For any $0 < K$, and any two probability measures $P, Q \in \mathcal{M}(\mathcal{H})$,*

$$K \cdot W(Q, P) = \sup_{f \in \mathcal{F}_K^{Lip}} \left| \mathop{\mathbb{E}}_{h \sim P} [f(h)] - \mathop{\mathbb{E}}_{h \sim Q} [f(h)] \right| \tag{14}$$

where $\mathcal{F}_K^{Lip}$ is the set of K-Lipschitz functions w.r.t. $\rho(h, h')$, i.e. functions that satisfy

$$\sup_{(h,h') \in \mathcal{H}^2} \frac{|f(h) - f(h')|}{\rho(h, h')} \leq K \tag{15}$$

We can write the Kantorovich-Rubinstein duality (14) using IPM formulation (Def. 3) and get

$$K \cdot W(Q, P) = \gamma_{\mathcal{F}_K^{\text{Lip}}}(P, Q) \tag{16}$$

Using this duality we will prove the following bound.

Theorem 11 (Template for Wasserstein-PB Bound)**.** *Assume that for any $\delta' \in (0, 1]$, w.p. at least $1 - \delta'$ over the sampling $S \sim \mathcal{D}^m$, the squared generalization gap function, $\Delta_S^2(h)$, is $K(m, \delta')$-Lipschitz w.r.t. the metric ρ. Then, for any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\Delta_S(Q) \leq \sqrt{K(m, \delta/2) W(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{17}$$

The proof (Appendix A.4) follows directly from Proposition 4 and the assumption, via the union bound. Theorem 11 can be seen as a template to be used for deriving Wasserstein-PB bounds in various learning settings where the $\Delta_S^2(\cdot)$ is K-Lipschitz with high probability, where the rate of K is preferably $O(1/m)$. Such a result can be challenging to prove since it requires uniform convergence of the slope between any two hypotheses in $\mathcal{H}$. Next, we will show specific learning settings where this property holds and the resulting generalization bounds.

We note that recent work Neu and Lugosi (2022) also establishes Wasserstein-BP bounds. This work assumed infinitely smooth loss functions and established bounds on the expected generalization gap, rather than high-probability bounds. In fact, Neu and Lugosi (2022) noted that obtaining such bounds is an open problem. Observe, though, that our bound depends on the Lipschitz constant $K(m, \delta)$ which needs to assessed; see sections 5.2 and 5.3 for specific examples. The general problem remains open.

At a more pragmatic level, we note that learning algorithms derived from minimizing Wasserstein based PB bounds have an added benefit of more stable optimization compared to KL based approaches, due to lower gradient variance, as noted in Arjovsky, Chintala, and Bottou (2017), and this distance measure can be approximated efficiently from finite samples (Cuturi, 2013; Weed & Bach, 2017)

5.2 Wasserstein-PB Bound for Finite Classes

We first investigate the simple case of a finite hypothesis class with a loss function ℓ which is G-Lipschitz w.r.t. the metric ρ. Note that for finite classes, UC always holds. We derive the following bound from Thm. 11.

Theorem 12 (Wasserstein-PB Bound for Finite Classes)**.** *For any finite hypothesis set $\mathcal{H}$, G-Lipschitz loss function, prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\Delta_S(Q) \leq \sqrt{\frac{8G \log(2|\mathcal{H}|/\delta)}{m} W(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{18}$$

The proof (Appendix A.5) makes use of standard union bound arguments and Hoeffding's concentration inequality. Notice that compared to the standard UC bound for finite classes, Thm. 12 multiplies the bound by a nonuniform factor of $\sqrt{GW(Q, P)}$.

5.3 Wasserstein-PB Bound for Loss-Gradient UC Classes

In this section we show a Wasserstein-PB bound for learning problems $(\mathcal{D}, \mathcal{H}, \ell)$, with $\mathcal{H} \subset \mathbb{R}^d$, for some dimension $d \in \mathbb{N}^+$, that satisfy the standard UC property, and, additionally, satisfy UC property for the loss gradient, as defined below.

Definition 13 (Loss-Gradient UC Property)**.** *A learning problem $(\mathcal{D}, \mathcal{H}, \ell)$, with $\mathcal{H} \subset \mathbb{R}^d$, is said to satisfy the **loss-gradient UC property**, if: (i) The loss function $\ell(h, z)$ is differentiable w.r.t. the hypothesis h. (ii) The problem satisfies the uniform convergence property (Def. 1). (iii) The empirical average of the loss gradient converges uniformly in L_2 norm sense to its mean. I.e., there exists a bound function $u^{grad}(m, \delta) > 0$, s.t. for any $\delta \in (0, 1)$ we have $u^{grad}(m, \delta) \underset{m\to\infty}{\to} 0$ and for any $m \in \mathbb{N}^+$,*

$$\mathbb{P}\left(\left\|\mathbb{E}_{z\sim\mathcal{D}}\nabla_h \ell(h, z) - \frac{1}{m}\sum_{i=1}^{m}\nabla_h \ell(h, z_i)\right\|_2 \leq u^{grad}(m, \delta), \forall h \in \mathit{Int}(\mathcal{H})\right) \geq 1 - \delta, \tag{19}$$

where $\nabla_h \ell(h, z)$ denotes the gradient of $\nabla_h \ell(h, z)$ w.r.t. h, for a fixed $z \in \mathcal{Z}$, and $\mathit{Int}(\mathcal{H})$ is the interior of $\mathcal{H}$.

Theorem 14 (Wasserstein-PB Bound for Loss-Gradient UC Classes)**.** *Let $(\mathcal{H}, \rho)$ be a metric space such that $\mathcal{H} \subset \mathbb{R}^d$ is a closed and convex set, and $\rho(h, h') \overset{def}{=} \|h - h'\|_2$. Assume that for any fixed $z \in \mathcal{Z}$, $\ell(h, z)$ is continuously differentiable for any $h \in \mathit{Int}(\mathcal{H})$. Assume the learning problem satisfies the loss-gradient UC property (Def. 13), with a uniform convergence bound u, and a uniform convergence of the loss gradient, u^{grad}. Then for any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

$$\Delta_S(Q) \leq \sqrt{2 \cdot u(m, \delta/4) \cdot u^{grad}(m, \delta/4) \cdot W(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{20}$$

The proof is in Appendix A.6. In learning problems that satisfy the loss-gradient UC property, we often have $u(m, \delta/4), u^{\text{grad}}(m, \delta/4) \in O(1/\sqrt{m})$, and then the resulting bound is $O(1/\sqrt{m})$. We provide full analysis that shows such a rate for the following linear regression example.

Corollary 15 (Wasserstein-PB Bound for Linear Regression)**.** *Let $z = (x, y)$, where x is sampled from a zero-mean σ sub-Gaussian distribution in $\mathbb{R}^d$, and the target, $y = f(x)$, is set by an unknown, possibly random, target function $\mathbb{R}^d \to [-\sigma, \sigma]$. The hypothesis space $\mathcal{H}$ is the unit $(d-1)$-sphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, and the loss function is $\ell(x, y, h) = (h^\top x - y)^2$. For any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0, 1)$, with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$,*

$$\Delta_S(Q) \leq \sqrt{2u(m, \delta/4) \cdot u^{grad}(m, \delta/4) \cdot W(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}, \tag{21}$$

where

$$u(m, \delta) \text{ and } u^{grad}(m, \delta) \text{ are in } O\left(\frac{\sigma^2}{\sqrt{m}}\sqrt{d\left(1 + \sqrt{1/\delta}\right) + \ln(1/\delta)}\right) \tag{22}$$

The full expression of the bound appears in the theorem's proof (Appendix A.7). The Wasserstein-PB bound we get is of order $O\left(u(m, \delta)\sqrt{W(Q, P)}\right)$.

6 Discussion

We have presented high-probability PB bounds based on integral probability metrics, that extend standard PB bounds based on KL divergence and more recent f-divergence and α-divergence based bounds, to a new class of divergences. Our bounds interpolate between classic UC bounds and PB bounds, by allowing data- and algorithm-dependent complexity terms. As in all PB results, our bounds suggest improved rates when the PB posterior is close to the prior. While we have extended high-probability PB bounds for IPMs to novel divergence measures, it is still an open question to do so without the UC assumption.

Possible directions for future research include: *(i)* Deriving high probability IPM-PB bounds (e.g. Wasserstein or TV based), without global UC assumptions (possibly using localization based approaches). *(ii)* Numerical investigation of the bounds and the derivation of algorithms that utilize them as minimization objectives. Based on the optimization advantages of Wasserstein based costs mentioned in Sec. 5.1, these could lead to enhanced practical utility. Such an advantage could play an important role in meta-learning schemes where PB methods have been widely used in recent years.

References

Alquier, P. (2021). User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*.

Aminian, G., Bu, Y., Wornell, G., & Rodrigues, M. (2022). Tighter expected generalization error bounds via convexity of information measures. *arXiv preprint arXiv:2202.12150*.

Anthony, M., & Bartlett, P. L. (1999). *Neural network learning: Theoretical foundations* (Vol. 9). cambridge university press Cambridge.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *PMLR*.

Bachmann, G., Moosavi-Dezfooli, S.-M., & Hofmann, T. (2021). Uniform convergence, adversarial spheres and a simple remedy. In *International conference on machine learning* (pp. 490–499).

Bastani, H., Simchi-Levi, D., & Zhu, R. (2021). Meta dynamic pricing: Transfer learning across experiments. *Management Science*.

Begin, L., Germain, P., Laviolette, F., & Roy, J.-F. (2016). PAC-Bayesian bounds based on the rényi divergence. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, *9*, 323–375.

Bousquet, O., Boucheron, S., & Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer school on machine learning* (pp. 169–207).

Catoni, O. (2007). PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *NIPS 2013*.

Dembo, A., & Zeitouni, O. (2009). Ldp for finite dimensional spaces. In *Large deviations techniques and applications* (pp. 11–70). Springer.

Donsker, M. D., & Varadhan, S. R. S. (1975). Asymptotic evaluation of certain markov process expectations for large time. *Comm. Pure Appl. Math.*.

Dziugaite, G. K., & Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR abs/1703.11008*.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*.

Hsu, D., Kakade, S., & Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, *17*, 1–6.

Lever, G., Laviolette, F., & Shawe-Taylor, J. (2013). Tighter PAC-Bayes bounds through distribution-dependent priors. *Theor. Comput. Sci.*, *473*(Feb), 4–28.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., & Weissman, T. (2019, 11). Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, *9*(4), 813-850.

Maurer, A. (2004). A note on the PAC Bayesian theorem. *CoRR abs/cs/0411099*.

McAllester, D. (1998). Some PAC-Bayesian theorems. *COLT*.

McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. *Proceedings of COLT 2003, 203-215*.

Müller, A. (1997). Stochastic orders generated by integrals: a unified study. *Advances in Applied probability*, *29*(2), 414–428.

Nagarajan, V., & Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, *32*.

Neu, G., & Lugosi, G. (2022). Generalization bounds via convex analysis. *arXiv preprint arXiv:2202.04985*.

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *Advances in neural information processing systems*, *30*.

Neyshabur, B., Tomioka, R., & Srebro, N. (2015). Norm-based capacity control in neural networks. In *Conference on learning theory* (pp. 1376–1401).

Ohnishi, Y., & Honorio, J. (2021). Novel change of measure inequalities with applications to PAC-Bayesian bounds and monte carlo estimation. In *International conference on artificial intelligence and statistics* (pp. 1711–1719).

Rivasplata, O., Kuzborskij, I., Szepesvári, C., & Shawe-Taylor, J. (2020). Pac-bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems*, *33*, 16833–16845.

Rodríguez Gálvez, B., Bassi, G., Thobaben, R., & Skoglund, M. (2021). Tighter expected generalization error bounds via wasserstein distance. *Advances in Neural Information Processing Systems*, *34*.

Seeger, M. (2002). PAC-Bayesian generalisation bounds for gaussian processes. *Journal of Machine Learning Research, 3:233-269*.

Seldin, Y., & Tishby, N. (2010). PAC-Bayesian analysis of co-clustering and beyond. *JMLR 2010*.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shawe-Taylor, J., & Williamson, R. C. (1997). A PAC analysis of a Bayesian estimator. *Proceedings of the International Conference on Computational Learning Theory (COLT)*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., & Lanckriet, G. R. (2009). On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., & Lanckriet, G. R. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, *6*, 1550–1599.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V., & Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity* (pp. 11–30). Springer.

Villani, C. (2006). *Optimal transport: old and new*. Springer Science and Business Media.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Wang, H., Diaz, M., Santos Filho, J. C. S., & Calmon, F. P. (2019). An information-theoretic view of generalization via wasserstein distance. In *2019 ieee international symposium on information theory (isit)* (pp. 577–581).

Weed, J., & Bach, F. (2017). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli 25 (4A), 2620-2648*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations, ICLR*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]