Integral Probability Metrics PAC-Bayes Bounds

Ron Amit
Technion - Israel Institute of Technology
The Viterbi Faculty of Electrical Engineering
ronamit@campus.technion.ac.il

Baruch Epstein
Technion - Israel Institute of Technology
The Viterbi Faculty of Electrical Engineering
baruch.epstein@gmail.com

Shay Moran
Technion - Israel Institute of Technology
Faculty of Mathematics
The Taub Faculty of Computer Science
Google Research, Israel
smoran@technion.ac.il

Ron Meir
Technion - Israel Institute of Technology
The Viterbi Faculty of Electrical Engineering
rmeir@ee.technion.ac.il

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 improved bounds in favorable 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, 2019b; 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 data and/or the algorithm. For

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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 strategy that produces PB bounds from change-of-measure inequalities leading to bounds based on the Rényi's α-divergence, and Alquier and Guedj (2018); Ohnishi and Honorio (2021); Picard-Weibel and Guedj (2022) further extended PB bounds to other Csiszár's f-divergences.

PB theorems consider the generalization performance of stochastic predictors. These bounds are non-uniform[2] 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.[3]

There are additional notable works on data and algorithm-dependent guarantees. The classical work of Bousquet and Elisseeff (2002) and Xu and Mannor (2012) studied generalization guarantees that depend on data and algorithm-dependent stability measures. A further line of recent papers tries to incorporate noise robustness/resilience. In Miyaguchi (2019), a PAC-Bayes transportation bound is used to measure the contribution of randomization to PB. This is done via optimal transport and Lipschitzness, based on the usual KL-PB bound. The work of Wei and Ma (2019) uses data-dependent Lipschitz smoothness to improve margin bounds, and Nagarajan and Kolter (2019a) passes from standard PB to a deterministic bound by assuming noise-resilience on the training data. This property translates to the test data, implying that good training smoothness leads to good test smoothness. Finally, Yang, Sun, and Roy (2019) measure data-dependent smoothness around each hypothesis (for each sample) and merge Catoni's bound (Catoni, 2007) with Rademacher theory, to obtain fast rates.

Recent work by Aminian, Bu, Wornell, and Rodrigues (2022); Lopez and Jog (2018); 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 probability" over the sample 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 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. Recently, Chee and Loustau (2021) proposed PB bounds with the entropy regularized optimal transport distance for an online-learning setting with a finite class.

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, as opposed to PB bounds (Dziugaite & Roy, 2017; Jiang, Neyshabur, Mobahi, Krishnan, & Bengio, 2019). 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.

[2] I.e., PB bounds apply even in learning problems without uniform convergence (Definition 1).

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

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 answering these questions positively. 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. Specifically, we focus on utilizing two specific IPMs: the total variation and Wasserstein metrics 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 KL-PB bounds, and it applies to deterministic as well as stochastic prior and posterior distributions. In fact, the IPM-based PB bounds we introduce match, at worst, the rate of the UC bound used. Recently, Livni and Moran (2020) showed that the classical KL-PB theorem cannot imply meaningful distribution-free generalization bounds for 1-dimensional linear classification. In contrast, our derived IPM-PB bounds do imply such bounds, because linear classifiers satisfy uniform convergence.

We note that the work of Audibert and Bousquet (2003, 2007) showed a different approach to utilizing the UC assumption to derive PAC-Bayes bounds. Their work assumed a UC property to utilize the generic chaining technique, resulting in more refined, variance-sensitive, bounds. In contrast to our work, their bound is not fully empirical, and the assumed UC bound is not used by the resulting bound.

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 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 $O(\sqrt{W_1(Q,P)})$, where $W_1(Q,P)$ is the 1st order Wasserstein distance between posterior and prior over the unit-sphere. We conduct a numerical simulation to demonstrate the improvement of the Wasserstein PB bound over the UC bound, and, in cases of narrow prior distributions, over the KL-PB bound. The experiment also investigates the case of non-randomized predictors by setting the prior and posterior as Dirac delta measures, which f-divergence based PB bounds are unable to use.

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 $\mathcal{Z}$, [4] 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

[4] 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)).

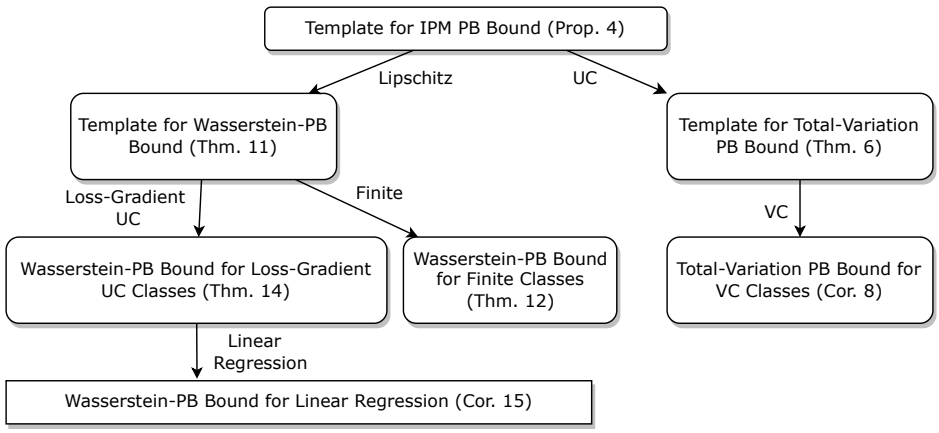


Figure 1: Claims tree diagram. $A \to B$ means that claim B is a special case of claim A, under additional assumptions on the learning problem. For full description of the assumptions, see the corresponding claims.

algorithm, the *empirical risk*, $\hat{L}_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$, can be evaluated using the training data. The generalization gap is defined by $\Delta_S(h) \stackrel{\text{def}}{=} L_{\mathcal{D}}(h) - \hat{L}_S(h)$.

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*

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

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 UC based bounds. Finally, UC bounds are usually not data-dependent, a property which is critical to explain the generalization of DNNs on real-world data (Nagarajan & Kolter, 2019b; 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

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

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 KL-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*

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

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 collection of function families $\{\mathcal{F}_S\}_{S\in\mathcal{Z}^m}$, 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 collection of bounded function sets, and a Wasserstein distance based bound, by selecting a collection of Lipschitz function sets.

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 the samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

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

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 collection of function families $\{\mathcal{F}_S\}_{S\in\mathcal{Z}^m}$, 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.

We denote by $\mathtt{kl}(p \parallel q)$ the KL divergence between two Bernoulli distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$, $p, q \in [0,1]$, that is,

$$\mathtt{kl}(p \parallel q) \stackrel{\text{def}}{=} \begin{cases} p\ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right), & \text{if } q \notin \{0,1\} \\ 0, & \text{else if } p = q \\ \text{undefined}, & \text{else.} \end{cases} \tag{6}$$

For any $h \in \mathcal{H}$, define the relative entropy of the empirical risk with respect to the expected one as $\Delta_S^{\mathtt{kl}}(h) \stackrel{\text{def}}{=} \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h))$. Similarly, and overloading notations, we define for any distribution $Q \in \mathcal{M}(\mathcal{H})$, $\Delta_S^{\mathtt{kl}}(Q) \stackrel{\text{def}}{=} \mathtt{kl}(\hat{L}_S(Q) \parallel L_D(Q))$.

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 $f_S(h) \stackrel{def}{=} m \cdot \Delta_S^{\mathtt{kl}}(h) \in \mathcal{F}_S$. Then for any prior $P \in \mathcal{M}(\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over the samples $S \sim \mathcal{D}^m$, for all $Q \in \mathcal{M}(\mathcal{H})$, we have*

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

By applying the Refined Pinsker's relaxation (McAllester (2003), Eq. 6) we can immediately derive the following, looser, but easier to interpret bound

$$\Delta_S(Q) \leq \sqrt{2\hat{L}_S(Q)\frac{\gamma_{\mathcal{F}_S}(Q,P) + \ln(2\sqrt{m}/\delta)}{m}} + 2\frac{\gamma_{\mathcal{F}_S}(Q,P) + \ln(2\sqrt{m}/\delta)}{m}. \tag{8}$$

When $\hat{L}_S(Q)$ 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

$$\mathcal{F}_M^{\infty} \stackrel{\text{def}}{=} \{f : \mathcal{H} \to [0,\infty), \|f\|_{\infty} \leq M\}, \tag{9}$$

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

$$\gamma_{\mathcal{F}_M^{\infty}}(Q,P) = \sup_{f \in \mathcal{F}_M^{\infty}} \left| \int_{\mathcal{H}} f \mathrm{d}P - \int_{\mathcal{H}} f \mathrm{d}Q \right| \stackrel{(i)}{=} M \cdot \sup_{A \in \Sigma_{\mathcal{H}}} |P(A) - Q(A)| = M D_{\mathrm{TV}}(Q,P), \tag{10}$$

where equality *(i)* holds since in the supremum it suffices to take the class of indicator functions $\{M \cdot \mathbb{1}_A(h), A \in \Sigma_{\mathcal{H}}\}$, since the functions in $\mathcal{F}_M^{\infty}$ are bounded in $[0, M]$.

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*

$$\Delta_S(Q) \leq \sqrt{u^2(m, \delta/2) D_{\mathrm{TV}}(Q,P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{11}$$

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 (11) 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 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 $O\left(\sqrt{\mathtt{VC}(\mathcal{H})/m}\right)$.

Proposition 7 (VC Bound, Boucheron, Bousquet, and Lugosi (2005))**.** *There exists some universal constant* [5] $c > 0$ *s.t. for any* $\delta \in (0,1)$ *we have*

$$\mathbb{P}\left\{\Delta_S(h) \leq c\sqrt{\frac{\mathtt{VC}(\mathcal{H}) + \ln(1/\delta)}{m}}, \forall h \in \mathcal{H}\right\} \geq 1 - \delta. \tag{12}$$

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*

$$\Delta_S(Q) \leq \sqrt{c\frac{\mathtt{VC}(\mathcal{H}) + \ln(1/\delta)}{m} D_{\mathrm{TV}}(Q,P) + \frac{\ln(m/\delta)}{2(m-1)}}. \tag{13}$$

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).[6] 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*

$$W_1(Q,P) \stackrel{def}{=} \inf_{\gamma \in \Gamma(Q,P)} \int_{\mathcal{H}\times\mathcal{H}} \rho(h,h')\mathrm{d}\gamma(h,h'), \tag{14}$$

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.

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

$$K \cdot W_1(Q,P) = \sup_{f \in \mathcal{F}_K^{Lip}} \left| \mathop{\mathbb{E}}_{h\sim P}[f(h)] - \mathop{\mathbb{E}}_{h\sim Q}[f(h)] \right|, \tag{15}$$

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

$$\sup_{(h,h')\in\mathcal{H}^2} \frac{|f(h) - f(h')|}{\rho(h,h')} \leq K. \tag{16}$$

[5] The bound of Cor. 7 originates from Talagrand (1994). As far as we know, there is no explicit value of the universal constant in the literature. Obtaining the constant involves careful computations of covering numbers and using the chaining method (e.g., based on Thm. 1.16 and 1.17 in Lugosi (2002)). Since our focus was not on the numerical evaluation of the bounds, we did not include this in our work. We note that there are other VC-type bounds with explicit constants, but with an extra $\log(m)$ factor (e.g., Vapnik (1999), Sect 3.4).

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

We can write the Kantorovich-Rubinstein duality (15) using IPM formulation (Def. 3),

$$K \cdot W_1(Q, P) = \gamma_{\mathcal{F}_K^{\text{Lip}}}(P, Q). \tag{17}$$

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, $\Delta_S^2(\cdot)$, is K-Lipschitz w.r.t. the metric ρ with some $K = K(m, \delta')$. 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*

$$\Delta_S(Q) \leq \sqrt{K(m, \delta/2) W_1(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{18}$$

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 (over samples), where the rate of $K = K(m, \delta/2)$ should be $O(1/m)$ to ensure a factor $O(1/\sqrt{m})$ multiplying the divergence between posterior and prior, as in the KL-PB bound.

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 a Wasserstein-based information-theoretic generalization bound. 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 $K = K(m, \delta)$ 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)**.** *Let $\mathcal{H}$ be a finite hypothesis class. Assume that for any fixed $z \in \mathcal{Z}$, $\ell(h, z)$ is a G-Lipschitz function in h 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*

$$\Delta_S(Q) \leq \sqrt{\frac{8G \log(4|\mathcal{H}|/\delta)}{m} W_1(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{19}$$

The proof (Appendix A.5) makes use of standard union bound arguments, Hoeffding's concentration inequality and the template Wasserstein-PB bound (Thm. 11). Notice that compared to the standard UC bound for finite classes, Thm. 12 multiplies the bound by a nonuniform factor of $\sqrt{GW_1(Q, P)}$.

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. h on*

Int$(\mathcal{H}) \times \mathcal{Z}$ *and continuous w.r.t.* h *on* $\mathcal{H} \times \mathcal{Z}$. *(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* $m \in \mathbb{N}^+$,

$$\mathbb{P}\left(\left\|\mathbb{E}_{z\sim\mathcal{D}}\nabla_h \ell(h, z) - \frac{1}{m}\sum_{i=1}^{m}\nabla_h \ell(h, z_i)\right\|_2 \leq u^{grad}(m, \delta), \forall h \in \mathit{Int}(\mathcal{H})\right) \geq 1 - \delta, \tag{20}$$

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 Int*$(\mathcal{H})$ *is the interior of* $\mathcal{H}$*. We call* u^{grad}*, the UC bound of the loss gradient.*

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* ρ *is the* L_2 *distance. Assume the learning problem satisfies the loss-gradient UC property (Def. 13), with UC bound* u*, and a UC bound 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*

$$\Delta_S(Q) \leq \sqrt{2 \cdot u(m, \delta/4) \cdot u^{grad}(m, \delta/4) \cdot W_1(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}. \tag{21}$$

The proof (Appendix A.6) is derived from the assumptions and the template Wasserstein-PB bound (Thm. 11). 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.

5.4 Linear Regression Example

Based on the Wasserstein-PB Bound for Loss-Gradient UC Classes (Thm. 14), we derive the following corollary.

Corollary 15 (Wasserstein-PB Bound for Linear Regression)**.** *Consider a data distribution of pairs* $z = (x, y)$*, where* x *is sampled from an unknown distribution supported on a* d*-dimensional ball of radius* $r > 0$*,* $\mathbb{B}_r^d \overset{def}{=} \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$*, and the target,* $y = f(x)$*, is set by an unknown, possibly random, target function* $f : \mathbb{B}_r^d \to [-1, 1]$*. The hypothesis space* $\mathcal{H}$ *is* $\mathbb{B}_{1/r}^d$ *, and the loss function is* $\ell(x, y, h) = \frac{1}{4}(h^\top x - y)^2$*. 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* $Q \in \mathcal{M}(\mathcal{H})$*,*

$$\Delta_S(Q) \leq \sqrt{2u(m, \delta/4) \cdot u^{grad}(m, \delta/4) \cdot W_1(Q, P) + \frac{\ln(2m/\delta)}{2(m-1)}}, \tag{22}$$

where $W_1(Q, P)$ *denotes the* 1st *order Wasserstein distance with the* L_2 *metric,*

$$u(m, \delta) \in O\left(\sqrt{\frac{d(1 + \ln(1/\delta))}{m}}\right), \text{ and } u^{grad}(m, \delta) \in O\left(r\sqrt{\frac{d(1 + \ln(1/\delta))}{m}}\right). \tag{23}$$

The full expression of the bound appears in the theorem's proof (Appendix A.7). Ignoring logarithmic factors we obtained a UC bound of $\tilde{O}\left(\sqrt{\frac{d}{m}}\right)$, and a Wasserstein-PB bound of $\tilde{O}\left(\sqrt{rW_1(Q, P)\frac{d}{m}}\right)$. Note that from Thm. 6 we can also deduce a TV-PB bound of order $\tilde{O}\left(\sqrt{D_{\text{TV}}(Q, P)\frac{d}{m}}\right)$. In comparison, the standard KL-PB bound is of order $\tilde{O}\left(\sqrt{\frac{\mathtt{KL}(Q \parallel P)}{m}}\right)$. The TV-PB bound is, at worst, roughly the same as the UC bound, since $D_{\text{TV}}(Q, P) \leq 1$. Note that since Q and P are distributions over a sphere of radius $1/r$, then $rW_1(Q, P) \leq 2$. Hence, the Wasserstein-PB is also, at worst, roughly the same as the UC bound. However, the KL-PB bound can be either tighter or looser, depending on P and Q. In cases where the mass of the posterior Q is concentrated in a region of the hypothesis set where the prior P is arbitrarily small, then the KL divergence can

be arbitrarily large, making the KL-PB bound extremely loose compared to the UC, TV-PB, and Wasserstein-PB bounds. The numerical experiment described in Appendix C demonstrates this by investigating different prior distributions with different widths. In particular, for posteriors and priors that are Dirac delta distributions (i.e., deterministic predictors), we show that the Wasserstein-PB considerably improves over the UC bound, while the KL-PB bound is undefined. We therefore demonstrated non-vacuous guarantees for deterministic models within the PB framework without requiring additional derandomization steps.

We can compare the bound of Thm. 14 to the bound of Corollary 8, in Neu and Lugosi (2022), which is also dependent on the Wasserstein distance between a data-dependent output (posterior) and a base measure (prior). The bound of Thm. 14 is different in the sense that *(i)* It holds with high probability instead of in expectation. *(ii)* Instead of assuming infinitely-smooth loss function with β-bounded directional derivatives, Thm. 14 assumes a loss-gradient UC class. *(iii)* The bound scales as $\tilde{O}\left(\sqrt{W_1 \cdot u \cdot u^{\text{grad}}}\right)$ instead of $\tilde{O}\left(\sqrt{W_2 \cdot \frac{d\beta}{m}}\right)$.

In our linear regression example, Cor. 15 scales as $\tilde{O}\left(\sqrt{W_1 \cdot u \cdot u^{\text{grad}}}\right) = \tilde{O}\left(\sqrt{r\sqrt{\frac{d}{m}}r\sqrt{\frac{d}{m}}}\right) = \tilde{O}\left(r\sqrt{\frac{d}{m}}\right)$. The loss is infinitely-smooth with $\beta \in O(1 + r + r^2)$, and therefore Cor. 8 of Neu and Lugosi (2022) scales as $\tilde{O}\left(\sqrt{W_2 \cdot \frac{d\beta}{m}}\right) = \tilde{O}\left(r\sqrt{\frac{d}{m}(1 + r + r^2)}\right)$, i.e., looser by a factor of $\tilde{O}\left(\sqrt{1 + r + r^2}\right)$ compared to Cor. 15.

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 distances. 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 distance 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, e.g. , Local Rademacher complexities (Bartlett, Bousquet, & Mendelson, 2005; Koltchinskii & Panchenko, 2004) are computed only on a subset of hypotheses with small empirical risk). This may allow non-vacuous bounds for large-scale models where global UC-based bounds are extremely vacuous. *(ii)* Derivation of algorithms that utilize the bounds 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 (Amit & Meir, 2018).

Acknowledgments

We thank Nadav Merlis and Daniel Soudry for helpful discussions of this work, and the anonymous reviewers for their helpful comments. SM is a Robert J. Shillman Fellow; he acknowledges support by ERC grant 802599, by ISF grant 1225/20, by BSF grant 2018385, by an Azrieli Faculty Fellowship, by Israel PBC-VATAT, and by the Technion Center for Machine Learning and Intelligent Systems (MLIS). The work of RM is partially supported by the Ollendorff Center of the Viterbi Faculty of Electrical and Computer Engineering at the Technion, and by the Skillman chair in biomedical sciences.

References

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

Alquier, P., & Guedj, B. (2018). Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, *107*(5), 887–902.

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

Amit, R., & Meir, R. (2018). Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International conference on machine learning* (pp. 205–214).

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 *International conference on machine learning* (pp. 214–223).

Audibert, J.-Y., & Bousquet, O. (2003). PAC-Bayesian generic chaining. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems (neurips)* (Vol. 16). MIT Press.

Audibert, J.-Y., & Bousquet, O. (2007). Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, *8*(4).

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).

Bartlett, P. L., Bousquet, O., & Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, *33*(4), 1497–1537.

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., & Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, *2*, 499–526.

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

Chee, A., & Loustau, S. (2021). Learning with BOT - Bregman and Optimal Transport divergences. *hal-03262687v2*.

Clement, P., & Desch, W. (2008). An elementary proof of the triangle inequality for the Wasserstein metric. *Proceedings of the American Mathematical Society*, *136*(1), 333–339.

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. In *Conference on uncertainty in artificial intelligence, (UAI)*.

Givens, C. R., & Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, *31*(2), 231–240.

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.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., & Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. In *International conference on learning representations (ICLR)*.

Koltchinskii, V., & Panchenko, D. (2004). Rademacher processes and bounding the risk of function learning. *arXiv preprint math/0405338*.

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

Livni, R., & Moran, S. (2020). A limitation of the pac-bayes framework. *Advances in Neural Information Processing Systems*, *33*, 20543–20553.

Lopez, A. T., & Jog, V. (2018). Generalization error bounds using wasserstein distances. In *2018 ieee information theory workshop (itw)* (pp. 1–5).

Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning* (pp. 1–56). Springer.

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. *Conference on Learning Theory (COLT)*.

McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. *Conference on Learning Theory (COLT)*.

Miyaguchi, K. (2019). PAC-Bayesian transportation bound. *arXiv preprint arXiv:1905.13435*.

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

Nagarajan, V., & Kolter, J. Z. (2019a). Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. *arXiv preprint arXiv:1905.13344*.

Nagarajan, V., & Kolter, J. Z. (2019b). 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).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Picard-Weibel, A., & Guedj, B. (2022). On change of measure inequalities for f-divergences. *arXiv preprint arXiv:2202.05568*.

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, ϕ-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.

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 28–76.

Thorpe, M. (2018). Introduction to optimal transport. *Notes of Course at University of Cambridge.*.

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* (Vol. 48). 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*.
Wei, C., & Ma, T. (2019). Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *arXiv preprint arXiv:1905.03684*.
Xu, H., & Mannor, S. (2012). Robustness and generalization. *Machine learning*, *86*(3), 391–423.
Yang, J., Sun, S., & Roy, D. M. (2019). Fast-rate PAC-Bayes generalization bounds via shifted rademacher processes. *arXiv preprint arXiv:1908.07585*.
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)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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]

A Appendix: Proofs

A.1 Proof of the IPM-PB Bound (Prop. 4)

Proof. The proof follows a similar structure as the classical derivation (McAllester, 2003; Shalev-Shwartz & Ben-David, 2014), except for replacing the change-of-measure inequality.

For any sample $S \in \mathcal{Z}^m$, consider the function

$$f_S(h) \stackrel{\text{def}}{=} 2(m-1)\Delta_S^2(h).$$

Since we assume $f_S \in \mathcal{F}_S$, Definition 3 implies that for any pair of probability measures $P, Q \in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \mathop{\mathbb{E}}_{h\sim P}[f_S(h)] \leq \gamma_{\mathcal{F}_S}(Q,P).$$

Therefore, by the monotonicity of $\exp(\cdot)$ we have

$$\exp\left(\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P)\right) \leq \exp\left(\mathop{\mathbb{E}}_{h\sim P}[f_S(h)]\right) \tag{24}$$

$$\leq \mathop{\mathbb{E}}_{h\sim P}\left[\exp(f_S(h))\right]. \tag{25}$$

where the last inequality is by the convexity of $\exp(\cdot)$, and by Jensen's inequality.

Taking the supremum over $Q \in \mathcal{M}(\mathcal{H})$, and an expectation over samples $S \sim \mathcal{D}^m$ we have that for any $P \in \mathcal{M}(\mathcal{H})$

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\sup_Q\left\{\exp\left(\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P)\right)\right\} &\leq \mathbb{E}_{S\sim\mathcal{D}^m}\sup_Q\left\{\mathop{\mathbb{E}}_{h\sim P}\exp(f_S(h))\right\} \\ &= \mathbb{E}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim P}\exp(f_S(h)) \\ &= \mathop{\mathbb{E}}_{h\sim P}\mathbb{E}_{S\sim\mathcal{D}^m}\exp(f_S(h)),\end{aligned} \tag{26}$$

where the last equality is obtained by the prior's independence from the sample, and from Fubini's theorem. We recall that by Hoeffding's inequality, for any $h \in \mathcal{H}$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\Delta_S(h) > u\right) \leq e^{-2mu^2},$$

which, by Lemma 5 of McAllester (2003), this imply.

$$\mathbb{E}_{S\sim\mathcal{D}^m}\exp(f_S(h)) = \mathbb{E}_{S\sim\mathcal{D}^m}\exp\big(2(m-1)\Delta_S^2(h)\big) \leq m. \tag{27}$$

Inequalities (26) and (27) imply

$$\mathbb{E}_{S\sim\mathcal{D}^m}\sup_Q\left\{\exp\left(\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P)\right)\right\} \leq m.$$

Therefore, by Markov's inequality, for any $t > 0$ we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_Q\left\{\exp\left(\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P)\right)\right\} \geq t\right) \leq \frac{m}{t}.$$

Or, equivalently,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\ln\left(\sup_Q\left\{\exp\left(\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P)\right)\right\}\right) \geq \ln(t)\right) \leq \frac{m}{t}.$$

By Lem. 18, the $\ln(\cdot)$ and $\sup(\cdot)$ operations are interchangeable, and therefore

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_Q\left\{\mathop{\mathbb{E}}_{h\sim Q}[f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P)\right\} \geq \ln(t)\right) \leq \frac{m}{t}.$$

Let $\delta \in (0,1)$, we set $t = \frac{m}{\delta}$, and plug in $f_S(h) \stackrel{\text{def}}{=} 2(m-1)\Delta_S^2(h)$ to get

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_Q\left\{\mathop{\mathbb{E}}_{h\sim Q}\left(2(m-1)\Delta_S^2(h)\right) - \gamma_{\mathcal{F}_S}(Q,P)\right\} \geq \ln(m/\delta)\right) \leq \delta.$$

Therefore, the complementary event satisfies

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_Q\left\{\mathop{\mathbb{E}}_{h\sim Q}\left(2(m-1)\Delta_S^2(h)\right) - \gamma_{\mathcal{F}_S}(Q,P)\right\} < \ln(m/\delta)\right) \geq 1-\delta.$$

I.e., for any $P \in \mathcal{M}(\mathcal{H})$, with a probability of at least $1-\delta$ over the samples $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim Q}\left(\Delta_S^2(h)\right) < \frac{\gamma_{\mathcal{F}_S}(Q,P) + \ln(m/\delta)}{2(m-1)}.$$

Jensen's inequality implies that

$$\left(\mathop{\mathbb{E}}_{h\sim Q}\Delta_S(h)\right)^2 \leq \mathop{\mathbb{E}}_{h\sim Q}\left(\Delta_S^2(h)\right) \leq \frac{\gamma_{\mathcal{F}}(Q,P) + \ln(m/\delta)}{2(m-1)}.$$

The proof is concluded by taking the square root of both sides. □

A.2 Proof of the Seeger's Type IPM-PB Bound (Prop. 5)

Proof. As in the proof of Prop. 4, we follow a similar structure as the classical derivation (Maurer, 2004; McAllester, 2003), except replacing the change-of-measure inequality.

For any sample $S \in \mathcal{Z}^m$, consider the function on $\mathcal{H}$

$$f_S(h) \stackrel{\text{def}}{=} m \cdot \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h)).$$

Note that $f_S(\cdot)$ is almost surely well-defined, since if $L_D(h) = 0$, then $\hat{L}(h) \stackrel{a.s.}{=} 0$.

Since we assume $f_S \in \mathcal{F}_S$, Definition 3 implies that for any pair of probability measures $P, Q \in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim Q}\left[f_S(h)\right] - \mathop{\mathbb{E}}_{h\sim P}\left[f_S(h)\right] \leq \gamma_{\mathcal{F}_S}(Q,P).$$

Therefore, by the monotonicity of $\exp(\cdot)$ we have

$$\begin{aligned}\exp\left(\mathop{\mathbb{E}}_{h\sim Q}\left[f_S(h)\right] - \gamma_{\mathcal{F}_S}(Q,P)\right) &\leq \exp\left(\mathop{\mathbb{E}}_{h\sim P}\left[f_S(h)\right]\right)\\ &\leq \mathop{\mathbb{E}}_{h\sim P}\left[\exp(f_S(h))\right].\end{aligned}$$

where the last inequality is by the convexity of $\exp(\cdot)$ and by Jensen's inequality.

Taking the supremum over $Q \in \mathcal{M}(\mathcal{H})$, and an expectation over samples $S \sim \mathcal{D}^m$ we have for any $P \in \mathcal{M}(\mathcal{H})$

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\sup_Q\left\{\exp\left(\mathop{\mathbb{E}}_{h\sim Q}\left[f_S(h)\right] - \gamma_{\mathcal{F}_S}(Q,P)\right)\right\} &\leq \mathbb{E}_{S\sim\mathcal{D}^m}\sup_Q\left\{\mathop{\mathbb{E}}_{h\sim P}\exp(f_S(h))\right\} \\ &= \mathbb{E}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim P}\exp(f_S(h)) \\ &= \mathop{\mathbb{E}}_{h\sim P}\mathbb{E}_{S\sim\mathcal{D}^m}\exp(f_S(h)),\end{aligned} \tag{28}$$

where the last equality is obtained by the prior's independence from the sample, and from Fubini's theorem. Using Maurer (2004), Thm. 1 we have

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\exp(f_S(h)) &= \mathbb{E}_{S\sim\mathcal{D}^m}\exp\left(m\cdot\mathtt{kl}(\hat{L}_S(h) \parallel L_D(h))\right) \\ &\leq 2\sqrt{m}.\end{aligned} \tag{29}$$

Inequalities (28) and (29) imply

$$\mathbb{E}_{S\sim\mathcal{D}^m} \sup_Q \left\{ \exp\left(\mathop{\mathbb{E}}_{h\sim Q} [f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P) \right) \right\} \leq 2\sqrt{m}.$$

Therefore, by Markov's inequality, for any $t > 0$ we have

$$\mathbb{P}_{S\sim\mathcal{D}^m} \left(\sup_Q \left\{ \exp\left(\mathop{\mathbb{E}}_{h\sim Q} [f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P) \right) \right\} \geq t \right) \leq \frac{2\sqrt{m}}{t},$$

or, equivalently,

$$\mathbb{P}_{S\sim\mathcal{D}^m} \left(\ln\left(\sup_Q \left\{ \exp\left(\mathop{\mathbb{E}}_{h\sim Q} [f_S(h)] - \gamma_{\mathcal{F}_S}(Q,P) \right) \right\} \right) \geq \ln(t) \right) \leq \frac{2\sqrt{m}}{t}.$$

By Lem. 18, the $\ln(\cdot)$ and $\sup(\cdot)$ operations are interchangeable, and therefore

Let $\delta \in (0,1)$, we set $t = \frac{2\sqrt{m}}{\delta}$, and plug in $f_S(h) \stackrel{\text{def}}{=} m \cdot \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h))$ to get

$$\mathbb{P}_{S\sim\mathcal{D}^m} \left(\sup_Q \left\{ m \mathop{\mathbb{E}}_{h\sim Q} \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h)) - \gamma_{\mathcal{F}_S}(Q,P) \right\} \geq \ln(2\sqrt{m}/\delta) \right) \leq \delta.$$

Therefore, the complementary event satisfies

$$\mathbb{P}_{S\sim\mathcal{D}^m} \left(\sup_Q \left\{ m \mathop{\mathbb{E}}_{h\sim Q} \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h)) - \gamma_{\mathcal{F}_S}(Q,P) \right\} < \ln(2\sqrt{m}/\delta) \right) \geq 1-\delta.$$

I.e., for any $P \in \mathcal{M}(\mathcal{H})$, with a probability of at least $1-\delta$ over the samples $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim Q} \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h)) < \frac{\gamma_{\mathcal{F}_S}(Q,P) + \ln(2\sqrt{m}/\delta)}{m}.$$

By the convexity of the function $\mathtt{kl}(p \parallel q)$ in the pair of parameters (p,q) and by Jensen's inequality, we have

$$\mathtt{kl}(\mathop{\mathbb{E}}_{h\sim Q} \hat{L}_S(h) \parallel \mathop{\mathbb{E}}_{h\sim Q} L_D(h)) \leq \mathop{\mathbb{E}}_{h\sim Q} \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h)).$$

Therefore we finally get that for any $P \in \mathcal{M}(\mathcal{H})$, w.p. of at least $1-\delta$ the following holds for all $Q \in \mathcal{M}(\mathcal{H})$

$$\mathtt{kl}(\hat{L}_S(Q) \parallel L_D(Q)) < \frac{\gamma_{\mathcal{F}_S}(Q,P) + \ln(2\sqrt{m}/\delta)}{m}.$$

□

A.3 Proof of the Total-Variation PAC-Bayes Bound (Thm. 6)

Proof. Let $\delta > 0$. For any fixed sample $S \in \mathcal{Z}^m$, we define $f_S(h) \stackrel{\text{def}}{=} 2(m-1)\Delta_S^2(h)$. Define $M_S \stackrel{\text{def}}{=} \sup_{h\in\mathcal{H}} \Delta_S^2(h)$. Therefore, $0 \leq f_S(h) \leq 2(m-1)M_S$, i.e., $f_S(h) \in \mathcal{F}^\infty_{2(m-1)M_S}$. This fact, together with the general IPM PB bound (Prop. 4) and Eq. (10) (equivalence of IPM to TV under the family of bounded functions) imply that for any $\delta \in (0,1)$

$$\mathbb{P}\left(\Delta_S(Q) \leq \sqrt{\frac{2(m-1)M_S D_{\mathrm{TV}}(Q,P)}{2(m-1)} + \frac{\ln(2m/\delta)}{2(m-1)}} \right) \geq 1 - \delta/2,$$

or equivalently,

$$\mathbb{P}\left(\Delta_S(Q) \leq \sqrt{M_S D_{\mathrm{TV}}(Q,P) + \frac{\ln(2m/\delta)}{2(m-1)}} \right) \geq 1 - \delta/2. \tag{30}$$

According to the UC assumption we have

$$\mathbb{P}(\Delta_S(h) \leq u(m, \delta/2), \forall h \in \mathcal{H}) \geq 1 - \delta/2,$$

and therefore, using the fact that $0 \leq \Delta_S(h) \leq 1$ we also have

$$\mathbb{P}\big(\Delta_S^2(h) \leq u^2(m, \delta/2), \forall h \in \mathcal{H}\big) \geq 1 - \delta/2,$$

which implies that the bounds also holds for the supremum M_S

$$\mathbb{P}\big(M_S \leq u^2(m, \delta/2)\big) \geq 1 - \delta/2. \tag{31}$$

To conclude the proof, we use a union bound argument and Equations (30) and (31).

□

A.4 Proof of the Template Wasserstein-PB Bound (Thm. 11)

Proof. Let $\delta > 0$. Let K_S be some Lipschitz constant of $\Delta_S^2(\cdot)$. Define $f_S(h) \stackrel{\text{def}}{=} 2(m-1)K_S$. Notice that $\Delta_S^2(h)$ is $2(m-1)K_S$-Lipschitz, i.e., $f_S(h) \in \mathcal{F}^{\text{Lip}}_{2(m-1)K_S}$. Using Proposition 4 we have

$$\mathbb{P}\left(\Delta(Q) \leq \sqrt{\frac{\gamma_{\mathcal{F}^{\text{Lip}}_{2(m-1)K_S}}(P, Q) + \ln(2m/\delta)}{2(m-1)}}\right) \geq 1 - \delta/2.$$

By equation (17) (the Kantorovich-Rubinstein duality) the inequality can rewritten as

$$\mathbb{P}\left(\Delta(Q) \leq \sqrt{\frac{2(m-1)K_S W_1(Q, P) + \ln(2m/\delta)}{2(m-1)}}\right) \geq 1 - \delta/2. \tag{32}$$

By assumption, w.p. at least $1 - \delta/2$, $\Delta_S^2(h)$ is K-Lipschitz with $K = K(m, \delta/2)$. Using a union bound argument with this event and the event of (32) concludes the proof. □

A.5 Proof of the Wasserstein-PB Bound for Finite Classes (Thm. 12)

We first prove the following lemma.

Lemma 16. *Let $\mathcal{H}$ be a finite hypothesis class. Assume that for any fixed $z \in \mathcal{Z}$, $\ell(h, z)$ is a G-Lipschitz function in $h \in \mathcal{H}$ w.r.t the metric ρ. Then for any $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(\tilde{K}_S \leq \frac{8}{m} G \log(2|\mathcal{H}|/\delta)\right) \geq 1 - \delta,$$

where for any fixed $S \in \mathcal{Z}^m$, $\tilde{K}_S$ is the sharp Lipschitz constant of $\Delta_S^2(\cdot)$, i.e.

$$\tilde{K}_S \stackrel{\text{def}}{=} \sup_{h, h' \in \mathcal{H}: h \neq h'} \frac{\left|\Delta_S^2(h) - \Delta_S^2(h')\right|}{\rho(h, h')}.$$

Proof of lemma 16. Let $\delta > 0$.

Note that for any $h, h' \in \mathcal{H}$ and $S \in \mathcal{Z}^m$, we have

$$\begin{aligned}
&\frac{|\Delta_S(h) - \Delta_S(h')|}{\rho(h, h')} \\
&= \frac{\left|\left[\mathbb{E}_{z\sim\mathcal{D}}\ell(h, z)) - \frac{1}{m}\sum_{i=1}^m \ell(h, z_i)\right] - \left[\mathbb{E}_{z\sim\mathcal{D}}\ell(h', z)) - \frac{1}{m}\sum_{i=1}^m \ell(h', z_i)\right]\right|}{\rho(h, h')} \\
&= \left|\frac{\frac{1}{m}\sum_{i=1}^m [\ell(h', z_i) - \ell(h, z_i)] - \mathbb{E}_{z\sim\mathcal{D}}[\ell(h', z) - \ell(h, z)]}{\rho(h, h')}\right|.
\end{aligned} \tag{33}$$

By assumption, for any fixed $z \in \mathcal{Z}$, $\ell(h, z)$ is a G-Lipschitz function in h w.r.t. the metric ρ. I.e., we have that $|\ell(h, z) - \ell(h', z)| \leq G\rho(h, h'), \forall h, h' \in \mathcal{H}, z \in \mathcal{Z}$. Hence, for any pair $(h, h') \in \mathcal{H}^2$, the random sequence $\{|\ell(h, z_i) - \ell(h', z_i)|\}_{i=1}^m$ is i.i.d. and bounded by $G\rho(h, h')$.

By Hoeffding's theorem, it holds with probability of at least $1 - \frac{\delta}{2|\mathcal{H}|^2}$ that

$$\left| \frac{1}{m} \sum_{i=1}^{m} [\ell(h', z_i) - \ell(h, z_i)] - \mathbb{E}_{z \sim \mathcal{D}}[\ell(h', z) - \ell(h, z)] \right| \leq \rho(h, h') G \sqrt{\frac{2 \ln(\frac{4|\mathcal{H}|^2}{\delta})}{m}}. \tag{34}$$

By using a union bound over claim (34) for all pairs of hypotheses $(h, h') \in \mathcal{H}^2$, we get that w.p. of at least $1 - \delta/2$, we have for all pairs $(h, h') \in \mathcal{H}^2$ **simultaneously** that,

$$\left| \frac{1}{m} \sum_{i=1}^{m} [\ell(h', z_i) - \ell(h, z_i)] - \mathbb{E}_{z \sim \mathcal{D}}[\ell(h', z) - \ell(h, z)] \right| \leq \rho(h, h') G \sqrt{\frac{2 \ln(\frac{4|\mathcal{H}|^2}{\delta})}{m}}. \tag{35}$$

It is well-known (e.g. Shalev-Shwartz and Ben-David (2014), Cor. 2.3) that for a finite hypothesis class and any $\delta/2 > 0$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |\Delta_S(h)| \leq \sqrt{\frac{\ln(4|\mathcal{H}|/\delta)}{2m}} \right) \geq 1 - \delta/2. \tag{36}$$

Notice that

$$\begin{aligned}
\tilde{K}_S &\stackrel{\text{def}}{=} \sup_{h, h' \in \mathcal{H}: h \neq h'} \frac{\left|\Delta_S^2(h) - \Delta_S^2(h')\right|}{\rho(h, h')} \\
&= \sup_{h, h' \in \mathcal{H}: h \neq h'} \frac{|\Delta_S(h) - \Delta_S(h')||\Delta_S(h) + \Delta_S(h')|}{\rho(h, h')} \\
&\leq \sup_{h, h' \in \mathcal{H}: h \neq h'} \frac{|\Delta_S(h) - \Delta_S(h')|}{\rho(h, h')} 2 \sup_{h'' \in \mathcal{H}} |\Delta_S(h'')| \\
&\stackrel{(33)}{=} 4 \sup_{h, h' \in \mathcal{H}: h \neq h'} \left| \frac{\frac{1}{m} \sum_{i=1}^{m} [\ell(h', z_i) - \ell(h, z_i)] - \mathbb{E}_{z \sim \mathcal{D}}[\ell(h', z) - \ell(h, z)]}{\rho(h, h')} \right| \sup_{h'' \in \mathcal{H}} |\Delta_S(h'')|.
\end{aligned}$$

The proof of the lemma is concluded by using a union bound argument with claims (35) and (36). We get that w.p. of at least $1 - \delta$ we have

$$\begin{aligned}
\tilde{K}_S &\leq 4 \sup_{h, h' \in \mathcal{H}: h \neq h'} \left\{ \frac{\rho(h, h') G \sqrt{\frac{2 \ln(4|\mathcal{H}|^2/\delta)}{m}}}{\rho(h, h')} \right\} \sqrt{\frac{\ln(4|\mathcal{H}|/\delta)}{2m}} \\
&= \frac{4}{m} G \sqrt{\ln(4|\mathcal{H}|^2/\delta) \ln(4|\mathcal{H}|/\delta)} \\
&\leq \frac{4}{m} G \ln(4|\mathcal{H}|^2/\delta) \\
&\leq \frac{8}{m} G \ln(2|\mathcal{H}|/\delta).
\end{aligned}$$

□

Proof of Theorem 12. The proof follows directly from Lemma 16 and Theorem 11. □

A.6 Proof of the Wasserstein-PB Bound for Differentiable Loss UC Classes (Thm. 14)

We first prove the following lemma.

Lemma 17. *Let $(\mathcal{H}, \rho)$ be a metric space such that $\mathcal{H} \subset \mathbb{R}^d$ is a closed convex set, and ρ is the L_2 distance. Assume that the loss function $\ell(h, z)$ is differentiable w.r.t. h on Int$(\mathcal{H}) \times \mathcal{Z}$ and continuous w.r.t. h on $\mathcal{H} \times \mathcal{Z}$. Assume the learning problem has a UC bound u (Def. 1), and a UC bound of the loss gradient, u^{grad} (Def. 13), then*

$$\mathbb{P}\Big(\tilde{K}_S \leq 2 \cdot u(m, \delta/2) \cdot u^{grad}(m, \delta/2)\Big) \geq 1 - \delta,$$

where for any fixed $S \in \mathcal{Z}^m$, $\tilde{K}_S$ is the sharp Lipschitz constant of $\Delta_S^2(\cdot)$, i.e.

$$\tilde{K}_S \overset{def}{=} \sup_{h,h' \in \mathcal{H}: h \neq h'} \frac{\left|\Delta_S^2(h) - \Delta_S^2(h')\right|}{\rho(h, h')}.$$

Proof of Lemma 17. Using the fact that loss function $\ell(h, z)$ is differentiable w.r.t. h on Int$(\mathcal{H}) \times \mathcal{Z}$ and continuous w.r.t. h on $\mathcal{H} \times \mathcal{Z}$, and by the mean value theorem and the convexity of $\mathcal{H}$, we have

$$\forall z \in \mathcal{Z}, (h, h') \in \mathcal{H}^2, \exists w_{z,h,h'} \in \mathcal{H}, \text{s.t. } \ell(h, z) - \ell(h', z) = \langle h - h', \nabla_h \ell(w_{z,h,h'}, z)\rangle, \tag{37}$$

where $\nabla_h \ell(w, z)$ denotes the gradient of $\ell(\cdot, \cdot)$ w.r.t the h variable, at the point (w, z).

Notice that

$$\begin{aligned}
\tilde{K}_S &\overset{def}{=} \sup_{h,h' \in \mathcal{H}: h \neq h'} \frac{\left|\Delta_S^2(h) - \Delta_S^2(h')\right|}{\rho(h, h')} \\
&= \sup_{h,h' \in \mathcal{H}: h \neq h'} \frac{|\Delta_S(h) - \Delta_S(h')||\Delta_S(h) + \Delta_S(h')|}{\rho(h, h')} \\
&\leq \sup_{h,h' \in \mathcal{H}: h \neq h'} \frac{|\Delta_S(h) - \Delta_S(h')|}{\rho(h, h')} 2 \sup_{h'' \in \mathcal{H}} |\Delta_S(h'')|.
\end{aligned} \tag{38}$$

We have for any $(h, h') \in \mathcal{H}, h \neq h$ that

$$\begin{aligned}
&\frac{|\Delta_S(h) - \Delta_S(h')|}{\rho(h, h')} \\
&= \left| \frac{\frac{1}{m}\sum_{i=1}^m [\ell(h', z_i) - \ell(h, z_i)] - \mathbb{E}_{z \sim \mathcal{D}}[\ell(h', z) - \ell(h, z)]}{\rho(h, h')} \right| \\
&\overset{(i)}{=} \left| \frac{\frac{1}{m}\sum_{i=1}^m \langle h - h', \nabla_h \ell(w_{z_i,h,h'}, z_i)\rangle - \mathbb{E}_{z \sim \mathcal{D}} \langle h - h', \nabla_h \ell(w_{z,h,h'}, z)\rangle}{\|h - h'\|_2} \right| \\
&\overset{(ii)}{=} \left| \frac{\langle h - h', \frac{1}{m}\sum_{i=1}^m \nabla_h \ell(w_{z_i,h,h'}, z_i) - \mathbb{E}_{z \sim \mathcal{D}} \nabla_h \ell(w_{z,h,h'}, z)\rangle}{\|h - h'\|_2} \right| \\
&\overset{(iii)}{\leq} \frac{\|h - h'\|_2 \left\|\frac{1}{m}\sum_{i=1}^m \nabla_h \ell(w_{z_i,h,h}, z_i) - \mathbb{E}_{z \sim \mathcal{D}} \nabla_h \ell(w_{z,h,h}, z)\right\|_2}{\|h - h'\|_2} \\
&= \left\| \frac{1}{m}\sum_{i=1}^m \nabla_h \ell(w_{z_i,h,h'}, z_i) - \mathbb{E}_{z \sim \mathcal{D}} \nabla_h \ell(w_{z,h,h'}, z) \right\|_2,
\end{aligned} \tag{39}$$

where *(i)* is by the mean value theorem (Eq. 37), *(ii)* is by the linearity of the sum, expectation, and the inner product, and *(iii)* is by the Cauchy–Schwarz inequality.

By the UC assumptions, we have

$$\mathbb{P}(\forall h \in \mathcal{H}, |\Delta_S(h)| \leq u(m, \delta/2)) \geq 1 - \delta/2, \tag{40}$$

and

$$\mathbb{P}\left(\forall h \in \text{Int}(\mathcal{H}), \left\| \frac{1}{m}\sum_{i=1}^m \nabla_h \ell(h, z_i) - \mathbb{E}_{z \sim \mathcal{D}} \nabla_h \ell(h, z) \right\|_2 \leq u^{\text{grad}}(m, \delta/2)\right) \geq 1 - \delta/2. \tag{41}$$

To conclude the proof, we use a union bound argument with (40) and (41), and use inequalities (38) and (39) to finally get

$$\mathbb{P}\Big(\tilde{K}_S \leq 2 \cdot u(m, \delta/2) \cdot u^{\text{grad}}(m, \delta/2)\Big) \geq 1 - \delta.$$

□

Proof of Theorem 14. The proof follows from Lemma 17 with $\delta/2$, Theorem 11 with $\delta/2$, and using the union bound. □

A.7 Proof of the Wasserstein-PB Bound for Linear Regression (Cor. 15)

Proof of Corollary 15. To meet the requirements of Theorem 14, we will prove a uniform convergence bound for the generalization gap (Def. 1), and for the loss gradient (Def. 13).

The generalization gap function for any $h \in \mathcal{H}$ can be written as

$$\begin{aligned}
\Delta_S(h) &= \mathbb{E}\ell(h,z) - \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i) \qquad (42)\\
&= \mathbb{E}\frac{1}{4}(h^\top x - y)^2 - \frac{1}{m}\sum_{i=1}^{m}\frac{1}{4}(h^\top x_i - y_i)^2\\
&= \frac{1}{4}\left(\mathbb{E}y^2 - \frac{1}{m}\sum_{i=1}^{m}y_i^2\right) - \frac{1}{2}\left(\mathbb{E}yx^\top h - \frac{1}{m}\sum_{i=1}^{m}y_i x_i^\top h\right) + \frac{1}{4}h^\top\left(\mathbb{E}xx^\top - \frac{1}{m}\sum_{i=1}^{m}x_i x_i^\top\right)h.
\end{aligned}$$

Let $\delta > 0$. We will now bound in high probability each of the three term above.

First term. The variables $y_1^2, \ldots, y_m^2$ are independent random variables in the range $[0,1]$, therefore by Hoeffding's inequality

$$\mathbb{P}\left(\left|\mathbb{E}y^2 - \frac{1}{m}\sum_{i=1}^{m}y_i^2\right| \le \sqrt{\frac{\ln(6/\delta)}{2m}}\right) \ge 1 - \delta/3. \qquad (43)$$

I.e.,

$$\mathbb{P}\left(\frac{1}{4}\left|\mathbb{E}y^2 - \frac{1}{m}\sum_{i=1}^{m}y_i^2\right| \le \sqrt{\frac{\ln(6/\delta)}{32m}}\right) \ge 1 - \delta/3. \qquad (44)$$

Second term. Note that yx is r-sub-Gaussian random vector in $\mathbb{R}^d$, since, for any s in the unit-sphere $\mathbb{S}_1^{d-1}$, $s^\top yx$ is r-sub-Gaussian (since it is a.s. bounded in $[-r,r]$). Therefore $\{y_i x_i\}_{i=1}^m$ are independent r-sub-Gaussian random vectors. Using Thm. 1 of Hsu, Kakade, and Zhang (2012), we have that for any $t > 0$,

$$\mathbb{P}\left(\left\|\mathbb{E}yx - \frac{1}{m}\sum_{i=1}^{m}y_i x_i\right\|_2^2 > \frac{r^2}{m}\left(d + 2d\sqrt{t} + 2t\right)\right) \le \exp(-t).$$

Therefore, we have

$$\mathbb{P}\left(\left\|\mathbb{E}yx - \frac{1}{m}\sum_{i=1}^{m}y_i x_i\right\|_2^2 < \frac{r^2}{m}\left(d + 2d\sqrt{\ln(3/\delta)} + 2\ln(3/\delta)\right)\right) \ge 1 - \delta/3. \qquad (45)$$

Then, by the Cauchy–Schwarz inequality we have that w.p. of at least $1 - \delta/3$, for all $h \in \mathbb{B}_{1/r}^d$

$$\begin{aligned}
\frac{1}{2}\left|h^\top\left(\mathbb{E}yx^\top - \frac{1}{m}\sum_{i=1}^{m}y_i x_i^\top\right)\right| &\le \frac{1}{2}\|h\|_2\left\|\mathbb{E}yx - \frac{1}{m}\sum_{i=1}^{m}y_i x_i\right\|_2 \qquad (46)\\
&< \frac{1}{2\sqrt{m}}\sqrt{d + 2d\sqrt{\ln(3/\delta)} + 2\ln(3/\delta)}.
\end{aligned}$$

Third term. By Theorem 6.5 of Wainwright (2019) (constants from Thm. of Bastani, Simchi-Levi, and Zhu (2021) Lem. 22), we have that w.p. of at least $1 - \delta/3$

$$\frac{1}{4}\left\|\frac{1}{m}\sum_{i=1}^{m}x_i x_i^\top - \mathbb{E}\left(xx^\top\right)\right\|_{\mathrm{op}} \le 8r^2\max\left\{\sqrt{\frac{5d + 2\ln\left(\frac{6}{\delta}\right)}{m}}, \frac{5d + 2\ln\left(\frac{6}{\delta}\right)}{m}\right\}, \qquad (47)$$

where $\|\cdot\|_{\text{op}}$ is the ℓ_2 operator-norm, that can be defined by $\|A\|_{\text{op}} \stackrel{\text{def}}{=} \sup_{u,v\in\mathbb{S}_1^{d-1}} \left|u^\top A v\right|, \forall A \in \mathbb{R}^{d\times d}$, where $\mathbb{S}_1^{d-1}$ is the the unit sphere in $\mathbb{R}^d$. Therefore, we conclude that w.p. of at least $1-\delta/3$

$$\forall h \in \mathbb{B}^d_{1/r}, \frac{1}{4}\left|h^\top\left(\frac{1}{m}\sum_{i=1}^m x_i x_i^\top - \mathbb{E}\big(xx^\top\big)\right)h\right| \leq 8\max\left\{\sqrt{\frac{5d+2\ln\left(\frac{6}{\delta}\right)}{m}}, \frac{5d+2\ln\left(\frac{6}{\delta}\right)}{m}\right\}. \tag{48}$$

Taking the absolute value of (42) and using the triangle inequality, the union bound, and inequalities (44), (46), and (48), we get that

$$\mathbb{P}(\forall h \in \mathcal{H}, \Delta_S(h) \leq u(m,\delta)) \geq 1-\delta,$$

where we defined

$$\begin{aligned} u(m,\delta) \stackrel{\text{def}}{=} & \sqrt{\frac{\ln(6/\delta)}{32m}} + \sqrt{\frac{d+2d\sqrt{\ln(3/\delta)}+2\ln(3/\delta)}{4m}} \\ & + 8\max\left\{\sqrt{\frac{5d+2\ln\left(\frac{6}{\delta}\right)}{m}}, \frac{5d+2\ln\left(\frac{6}{\delta}\right)}{m}\right\}. \end{aligned} \tag{49}$$

Therefore, $u(m,\delta) \in O\left(\sqrt{\frac{d(1+\ln(1/\delta))}{m}}\right)$.

Next, we wish to prove a uniform convergence bound for the loss gradient, in Euclidean norm. Note that for any $h \in \mathcal{H}$

$$\begin{aligned} & \mathbb{E}_{z\sim\mathcal{D}}\nabla_h \ell(h,z) - \frac{1}{m}\sum_{i=1}^m \nabla_h \ell(h,z_i) \\ &= \mathbb{E}_{z\sim\mathcal{D}}\nabla_h \frac{1}{4}(h^\top x - y)^2 - \frac{1}{m}\sum_{i=1}^m \nabla_h \frac{1}{4}(h^\top x_i - y_i)^2 \\ &= \frac{1}{2}\mathbb{E}(h^\top x - y)x^\top - \frac{1}{m}\sum_{i=1}^m \frac{1}{2}(h^\top x_i - y_i)x_i^\top \\ &= \frac{1}{2}h^\top\left(\mathbb{E}xx^\top - \frac{1}{m}\sum_{i=1}^m x_i x_i^\top\right) - \frac{1}{2}\left(\mathbb{E}yx^\top - \frac{1}{m}\sum_{i=1}^m y_i x_i^\top\right). \end{aligned} \tag{50}$$

To bound the L_2 norm of the first term of the equation above, we use similar argument as in (47), and the fact that the operator-norm can be defined equivalently by $\forall A \in \mathbb{R}^{d\times d}, \|A\|_{\text{op}} = \sup_{v\in\mathbb{S}_1^{d-1}} \|Av\|_2$, to get that w.p. of at least $1-\delta/2$

$$\forall h \in \mathbb{B}^d_{1/r}, \left\|\frac{1}{2}h^\top\left(\frac{1}{m}\sum_{i=1}^m x_i x_i^\top - \mathbb{E}\big(xx^\top\big)\right)\right\|_2 \leq 16r\max\left\{\sqrt{\frac{5d+2\ln\left(\frac{4}{\delta}\right)}{m}}, \frac{5d+2\ln\left(\frac{4}{\delta}\right)}{m}\right\}. \tag{51}$$

To bound the L_2 norm of the second term, we use the same argument as in (45), and get

$$\mathbb{P}\left(\frac{1}{2}\left\|\mathbb{E}yx - \frac{1}{m}\sum_{i=1}^m y_i x_i\right\|_2 < \frac{r}{2\sqrt{m}}\sqrt{d+2d\sqrt{\ln(2/\delta)}+2\ln(2/\delta)}\right) \geq 1-\delta/2. \tag{52}$$

Now, taking the norm of equality (50) and using the triangle inequality, inequalities (51) and (52), and the union bound we get that

$$\mathbb{P}\left(\forall h \in \mathcal{H}, \left\|\mathbb{E}_{z\sim\mathcal{D}}\nabla_h \ell(h,z) - \frac{1}{m}\sum_{i=1}^m \nabla_h \ell(h,z_i)\right\|_2 \leq u^{\text{grad}}(m,\delta)\right) \geq 1-\delta, \tag{53}$$

where we defined

$$u^{\text{grad}}(m,\delta) \stackrel{\text{def}}{=} 16r \max\left\{\sqrt{\frac{5d+2\ln\left(\frac{4}{\delta}\right)}{m}}, \frac{5d+2\ln\left(\frac{4}{\delta}\right)}{m}\right\} + r\sqrt{\frac{d+2d\sqrt{\ln(2/\delta)}+2\ln(2/\delta)}{4m}}. \tag{54}$$

Therefore, $u^{\text{grad}}(m,\delta) \in O\left(r\sqrt{\frac{d(1+\ln(1/\delta))}{m}}\right)$.

Notice that the loss is bounded in $[0,1]$, since

$$\ell(x,y,h) = \frac{1}{4}(h^\top x - y)^2 \leq \frac{1}{4}2\big((h^\top x)^2 + y^2\big) \leq \frac{1}{2}\Big(\|h\|_2^2\,\|x\|_2^2 + 1\Big) \leq 1.$$

Therefore we can use Theorem 14 to conclude the proof. □

B Appendix: An Example of a Seeger Type Bound

To derive an analogous Seeger's type theorem to Thm. 6, we need to prove uniform convergence of the kl-gap, $\Delta_S^{\mathtt{kl}}(h) \stackrel{\text{def}}{=} \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h))$, rather than the usual gap $\Delta_S(h) = L_D(h) - \hat{L}_S(h)$.

For example, consider the binary classification and finite $\mathcal{H}$ case.

For each $h \in \mathcal{H}$, we bound $\Delta_S^{\mathtt{kl}}(h) = \mathtt{kl}(\hat{L}_S(h) \parallel L_D(h))$ using the concentration inequality from Dembo and Zeitouni (2009) Thm. 2.2.3. (see also Mardia, Jiao, Tánczos, Nowak, and Weissman (2019) Lem. 8), which holds since $\hat{L}_S(h)$ is an empirical average of m Bernoulli i.i.d variables with mean $L_D(h)$. For any $\varepsilon > 0$, we have

$$\mathbb{P}\big(\Delta_S^{\mathtt{kl}}(h) < \varepsilon\big) \geq 1 - 2\exp(-m\varepsilon).$$

Using a union bound argument we get

$$\mathbb{P}\big(\forall h \in \mathcal{H}, \Delta_S^{\mathtt{kl}}(h) < \varepsilon\big) \geq 1 - 2|\mathcal{H}|\exp(-m\varepsilon).$$

Therefore, for any $\delta \in (0,1)$ we can get

$$\mathbb{P}\left(\forall h \in \mathcal{H}, \Delta_S^{\mathtt{kl}}(h) < \frac{\ln(2|\mathcal{H}|/\delta)}{m}\right) \geq 1 - \delta. \tag{55}$$

Let $\mathcal{F}^{\infty}_{\ln(4|\mathcal{H}|/\delta)}$ be the family of functions as defined in Eq. 9, i.e., functions that are bounded in the ∞-norm by $\ln(4|\mathcal{H}|/\delta)$, for which the IPM is $\ln(4|\mathcal{H}|/\delta)D_{\text{TV}}(Q,P)$.

By (55), w.p. at least $1 - \delta/2$ we have that $m\Delta_S^{\mathtt{kl}}(h) \in \mathcal{F}^{\infty}_{\ln(4|\mathcal{H}|/\delta)}$

Now we can use the Seeger's type IPM-PB bound (Prop. 5) and a union bound argument, to get that with probability at least $1-\delta$ over the samples $S \sim \mathcal{D}^m$, the following inequality holds for all $Q \in \mathcal{M}(\mathcal{H})$

$$\Delta_S(Q) \leq \sqrt{2\hat{L}_S(Q)\frac{\ln(4|\mathcal{H}|/\delta)D_{\text{TV}}(Q,P) + \ln(4\sqrt{m}/\delta)}{m}} + 2\frac{\ln(4|\mathcal{H}|/\delta)D_{\text{TV}}(Q,P) + \ln(4\sqrt{m}/\delta)}{m}.$$

C Appendix: Numerical Demonstration Details

This section describes the experiment that implements the setting of Corollary 15 (Wasserstein-PB Bound for Linear Regression). The code is available at: https://github.com/ron-amit/pac_bayes_reg.

The sample distribution. The unknown data distribution $\mathcal{D}$ is determined by a latent vector $g \in \mathbb{R}^d$, drawn once per experiment instance from a uniform distribution over $\mathbb{B}_{0.1}$. The dimension

is $d = 10$. For each sample $(x, y) \sim \mathcal{D}$, x is drawn uniformly from $\mathbb{B}_{0.1}$ and $y = f(x)$ is set by $f(x) = \text{clip}_{[-1,1]}\{g^\top x + \xi\}$ where,

$$\underset{[a,b]}{\text{clip}}(t) \stackrel{\text{def}}{=} \begin{cases} a, & t < a \\ t, & a \leq t \leq b \\ b & t > b, \end{cases}$$

for any $a, b, t \in \mathbb{R}$, and ξ is drawn uniformly from $[-0.5, 0.5]$. The motivation for this choice of $\mathcal{D}$ is to have an underlying linear structure in the data, corrupted by noise. The clipping ensures that the loss values are in the range $[0, 1]$.

The prior and posterior distributions. The hypothesis space is an r-radius ball $\mathcal{H} = \mathbb{B}_r$, with $r = 1$. The prior and posterior distributions over $\mathcal{H}$ are set as projected Gaussian distributions. Let $P_{\mathbb{B}} : \mathbb{R}^d \to \mathbb{B}_r$ be a projection operator onto $\mathbb{B}_r$ Let $\tilde{P}$ be a Gaussian measure over $\mathbb{R}^d$, $\mathcal{N}(\mu_P, \sigma_P^2 I)$, where $\mu_P = \bar{\mathbf{0}}$, and σ_P is a fixed constant that will be specified later. The prior is defined as $P = P_{\mathbb{B}} \sharp \tilde{P}$, i.e. , as the push-forward measure of $\tilde{P}$ under the projection $P_{\mathbb{B}}$. The family of posteriors we are considering are projected Gaussian distributions, $\mathcal{Q} \stackrel{\text{def}}{=} \left\{ P_{\mathbb{B}} \sharp \tilde{Q} : \tilde{Q} = \mathcal{N}(\mu_Q, \sigma_Q^2 I), \mu_Q, \in \mathbb{B}_{r_Q} \right\}$, where σ_Q is a fixed constant that will be specified later and the maximal norm of μ_Q is $r_Q = 0.05$.

The Wasserstein distance. Since there is no closed-form formula for the 1st order Wasserstein distance between Gaussian distributions projected onto a ball $W_1(Q, P)$, we will instead use an upper bound. We use Lemma 19 (Sect. D) to bound this distance with the distance of the corresponding pre-projection measures, $W_1(\tilde{Q}, \tilde{P})$, where $\tilde{Q}$ and $\tilde{P}$ are the corresponding pre-projection measures. Note that our choice of parameters ensures that the lemma condition holds:

$$r^2 \geq \max\left\{ \|\mu_Q\|_2^2 + \|\Sigma_Q\|_F^2, \|\mu_P\|_2^2 + \|\Sigma_P\|_F^2 \right\} = \max\left\{ \|\mu_Q\|_2^2 + d\sigma_Q^2, \|\mu_P\|_2^2 + d\sigma_P^2 \right\}.$$

We also use the fact that $W_1(\tilde{Q}, \tilde{P}) \leq W_2(\tilde{Q}, \tilde{P})$ (Givens and Shortt (1984), Prop. 3) and the analytic formula for the 2nd order Wasserstein distance between two Gaussian distributions (Givens and Shortt (1984), Prop. 7) to finally get a closed-form upper bound,

$$\begin{aligned} W_1(Q, P) &\stackrel{\text{Lem. 19}}{\leq} \sqrt{\|\mu_Q - \mu_P\|_2^2 + \mathbf{Tr}\left(\Sigma_Q + \Sigma_P - 2\left(\Sigma_Q^{1/2}\Sigma_P\Sigma_Q^{1/2}\right)^{1/2}\right)} \\ &+ \sqrt{\frac{\pi}{2}\|\Sigma_Q\|_{2,2}}\,\text{erfc}\left(\frac{r - \sqrt{\|\mu_Q\|_2^2 + \|\Sigma_Q\|_F^2}}{\sqrt{2\|\Sigma_Q\|_{2,2}}}\right) \\ &+ \sqrt{\frac{\pi}{2}\|\Sigma_P\|_{2,2}}\,\text{erfc}\left(\frac{r - \sqrt{\|\mu_P\|_2^2 + \|\Sigma_P\|_F^2}}{\sqrt{2\|\Sigma_P\|_{2,2}}}\right) \\ &= \sqrt{\|\mu_Q - \mu_P\|_2^2 + d(\sigma_Q - \sigma_P)^2} \\ &+ \sqrt{\frac{\pi}{2}}\sigma_Q\,\text{erfc}\left(\frac{r - \sqrt{\|\mu_Q\|_2^2 + d\sigma_Q^2}}{\sqrt{2}\sigma_Q}\right) + \sqrt{\frac{\pi}{2}}\sigma_P\,\text{erfc}\left(\frac{r - \sqrt{\|\mu_P\|_2^2 + d\sigma_P^2}}{\sqrt{2}\sigma_P}\right) \\ &\stackrel{\text{def}}{=} W_{\text{bound}}(\mu_Q). \end{aligned} \tag{56}$$

Notice that in the limit of $\sigma_Q, \sigma_P \to 0$, the bound becomes $\|\mu_Q - \mu_P\|_2$, which is equivalent to the Wasserstein distance between two Dirac measures at μ_Q and μ_P.

The empirical risk term. To compute the expectation of the empirical risk w.r.t. the posterior, $\mathbb{E}_{h\sim Q}\hat{L}(h)$, we derive a closed-form formula using the structure and of the loss and the posterior distribution [7] . Given a dataset $S = \{(x_i, y_i)\}_{i=1}^m$, denote $X \in \mathbb{R}^{m\times d}$ as a matrix whose rows are the

[7] In cases where the loss is a more complicated function (but still differentiable), one can approximate the expectation over the posterior with the reparametrization trick (D. P. Kingma & Welling, 2013), similarly to Amit and Meir (2018); Dziugaite and Roy (2017).

vectors x_i, and denote $Y \in \mathbb{R}^{m\times 1}$ as a vector whose entries are y_i. Denote $\hat{J}_{(X,Y)}(\mu_Q) \stackrel{\text{def}}{=} \mathbb{E}_{h\sim Q}\hat{L}(h)$. Then we have

$$\begin{aligned}
\hat{J}_{(X,Y)}(\mu_Q) &= \mathbb{E}_{h\sim\mathcal{N}(\mu_Q,\sigma_Q^2 I)}\frac{1}{m}\sum_{i=1}^{m}\frac{1}{4}(h^\top x_i - y_i)^2 \\
&= \frac{1}{4m}\mathbb{E}_{h\sim\mathcal{N}(\mu_Q,\sigma_Q^2 I)}\left\|Xh^\top - Y\right\|_2^2 \\
&= \frac{1}{4m}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left\|X(\mu_Q+\sigma_Q\epsilon)^\top - Y\right\|_2^2 \\
&= \frac{1}{4m}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left\|\sigma_Q X\epsilon^\top + X\mu_Q^\top - Y\right\|_2^2 \\
&= \frac{1}{4m}\Big(\left\|X\mu_Q^\top - Y\right\|_2^2 + \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\sigma_Q^2\,\mathbf{Tr}(\epsilon X^\top X\epsilon^\top)\Big) \\
&= \frac{1}{4m}\Big(\left\|X\mu_Q^\top - Y\right\|_2^2 + \sigma_Q^2\,\mathbf{Tr}(X^\top X)\Big) \\
&= \frac{1}{4m}\Big(\left\|X\mu_Q^\top - Y\right\|_2^2 + \sigma_Q^2\left\|X\right\|_F^2\Big).
\end{aligned} \tag{57}$$

The explicit Wasserstein-PB bound. According to Cor. 15, given a training set $S = (X, Y)$, the upper bound on the expected risk $L(Q) \stackrel{\text{def}}{=} \mathbb{E}_{h\sim Q}L(h)$ is

$$J^{\text{WPB}}_{(X,Y)}(\mu_Q) \stackrel{\text{def}}{=} \hat{J}_{(X,Y)}(\mu_Q) + \sqrt{2u(m,\delta/4)\cdot u^{\text{grad}}(m,\delta/4)\cdot W_{\text{bound}}(\mu_Q) + \frac{\ln(2m/\delta)}{2(m-1)}}, \tag{58}$$

where $u(m,\delta)$ is defined in (49), $u^{\text{grad}}(m,\delta)$ is defined in (54), and $W_{\text{bound}}(\mu_Q)$ is the upper bound over $W_1(Q,P)$ defined in (56).

The explicit KL-PB bound. For the KL-PB bound, we use the classic PB bound (Prop. 2) with the KL-divergence replaced by an upper bound that has a closed-form expression. By the data-processing inequality we have that $\mathtt{KL}(Q \parallel P) = \mathtt{KL}(P_{\mathbb{B}}\sharp\tilde{Q} \parallel P_{\mathbb{B}}\sharp\tilde{P}) \le \mathtt{KL}(\tilde{Q} \parallel \tilde{P})$, and $\mathtt{KL}(\tilde{Q} \parallel \tilde{P})$ can be computed using the analytic formula for the KL-divergence between two Gaussian distributions. Therefore, the upper bound we use is

$$J^{\text{KL-PB}}_{(X,Y)}(\mu_Q) \stackrel{\text{def}}{=} \hat{J}_{(X,Y)}(\mu_Q) + \sqrt{\frac{\frac{\|\mu_Q-\mu_P\|_2^2}{2\sigma_P^2} + d\Big(\ln\Big(\frac{\sigma_P}{\sigma_Q}\Big) + \frac{\sigma_Q^2}{2\sigma_P^2} - \frac{1}{2}\Big) + \ln(m/\delta)}{2(m-1)}}. \tag{59}$$

Experiment Procedure: We repeat the experiment for 10 repetitions, to account for the randomness of the data and optimization in each run. In each run, *(i)* the task data distribution $\mathcal{D}$ is generated as described above, *(ii)* A training set of m samples is generated. *(iii)* The posterior mean vector μ_Q is learned using the Adam Optimizer (D. Kingma & Ba, 2015) that minimizes either $J^{\text{KL-PB}}_{(X,Y)}(\mu_Q)$ or $J^{\text{WPB}}_{(X,Y)}(\mu_Q)$ (as will be specified later), where the learning rate is set as 10^{-3}, and the maximal batch size is 256. The gradients are computed using automatic differentiation by the PyTorch framework (Paszke et al., 2019). After each gradient step, the parameter μ_Q is projected to $\mathbb{B}_{r_Q}$.

Results. Table 1 show the results when we set the prior parameter σ_P as 10^{-2}, and Table 2 shows the results for $\sigma_P = 10^{-4}$, both use $\sigma_Q = 10^{-3}$. The optimization objective for those two setups is the KLPB bound, $J^{\text{KL-PB}}_{(X,Y)}(\mu_Q)$.

The third setup (Table 3) investigates Dirac posteriors ("a deterministic model"). In this setup we set $\sigma_Q = \sigma_P = 0$, and the optimization objective is set to be $J^{\text{WPB}}_{(X,Y)}(\mu_Q)$. Note that since $\sigma_P = 0$ then the KL-divergence is undefined, while the W_1 distance equals exactly $\|\mu_Q - \mu_P\|_2$.

Figures 2a, 2b and 2c show the corresponding plots. The 'Training loss' column shows the empirical risk (57), i.e., the averaged loss of the learned posterior on the training data. The 'Test loss' column shows the average loss of the learned posterior on a separate 'test' set of 10000 samples drawn from $\mathcal{D}$. In all the evaluated bounds, we use the confidence parameter $\delta = 0.05$. The 'UC bound' shows

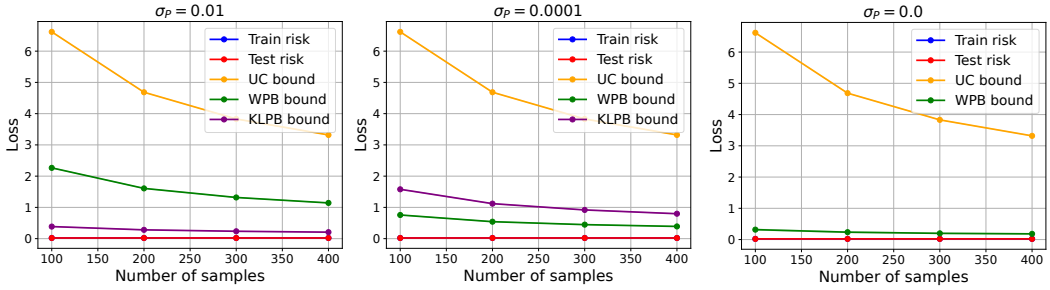
Table 1: Linear regression experiment with $\sigma_{\mathbf{P}} = \mathbf{10^{-2}}, \sigma_{\mathbf{Q}} = \mathbf{10^{-3}}$. Each cell shows the mean over 10 independent runs and the 95% confidence interval in parenthesis.

# samples	Train risk	Test risk	UC bound	WPB bound	KLPB bound
100	0.0211 (0.0010)	0.0208 (0.0001)	6.6176 (0.0010)	2.2652 (0.0010)	0.3861 (0.0010)
200	0.0206 (0.0009)	0.0208 (0.0001)	4.6850 (0.0009)	1.6080 (0.0009)	0.2814 (0.0009)
300	0.0214 (0.0006)	0.0209 (0.0001)	3.8298 (0.0006)	1.3177 (0.0006)	0.2357 (0.0006)
400	0.0205 (0.0005)	0.0208 (0.0001)	3.3187 (0.0005)	1.1433 (0.0005)	0.2070 (0.0005)

Table 2: Linear regression experiment with $\sigma_{\mathbf{P}} = \mathbf{10^{-4}}, \sigma_{\mathbf{Q}} = \mathbf{10^{-3}}$. Each cell shows the mean over 10 independent runs and the 95% confidence interval in parenthesis.

# samples	Train risk	Test risk	UC bound	WPB bound	KLPB bound
100	0.0211 (0.0010)	0.0208 (0.0001)	6.6176 (0.0010)	0.7569 (0.0010)	1.5787 (0.0010)
200	0.0206 (0.0009)	0.0208 (0.0001)	4.6850 (0.0009)	0.5424 (0.0009)	1.1199 (0.0009)
300	0.0214 (0.0006)	0.0209 (0.0001)	3.8298 (0.0006)	0.4482 (0.0006)	0.9186 (0.0006)
400	0.0205 (0.0005)	0.0208 (0.0001)	3.3187 (0.0005)	0.3906 (0.0005)	0.7974 (0.0005)

the sum of the empirical risk and the UC generalization gap bound (49). The 'WPB bound' is the Wasserstein-PB bound evaluated by equation (58), and the 'KLPB bound' is evaluated by equation (59). The results clearly show the improved tightness of the WPB bound over the UC bound, for the two choices of a prior distribution. The KLPB bound, also shows relatively tight values, as expected from an algorithm- and data-dependent bound. However, for the narrower prior distribution ($\sigma_P = 10^{-4}$), the KLPB bound is significantly looser than the WPB bound. That is expected from the properties of the KL-divergence, which can tend to ∞ if $\sigma_P \to 0$, as opposed to the Wasserstein distance. In the extreme case of $\sigma_P = 0$ the KLPB bound is undefined, while the WPB exhibits a considerable improvement over the UC bound. The results confirm that the WPB generally improves over UC bounds, and may be tighter than the KLPB bound, depending on the prior and posterior distributions.



(a) $\sigma_{\mathbf{P}} = \mathbf{10^{-2}}, \sigma_{\mathbf{Q}} = \mathbf{10^{-3}}$ (b) $\sigma_{\mathbf{P}} = \mathbf{10^{-4}}, \sigma_{\mathbf{Q}} = \mathbf{10^{-3}}$ (c) $\sigma_{\mathbf{P}} = \sigma_{\mathbf{Q}} = \mathbf{0}$

Figure 2: Linear regression experiment. Note that the the 95% confidence interval is too small to be discernible in the plots, and that blue train risk plot is not visible since it is very close to the test risk.

D Appendix: Technical Lemmas

Lemma 18. *Let $A \subset \mathbb{R}$ be bounded and non-empty, and let $f : (0, \infty) \to \mathbb{R}$ be continuous and monotone non-decreasing. Then $f(\sup A) = \sup f(A)$, where we defined $\sup A \stackrel{def}{=} \sup_{a \in A} a$, and $\sup f(A) \stackrel{def}{=} \sup_{a \in A} f(a)$; that is, $f(A)$ is the image of the set A under f.*

Proof. First notice that $f(\sup A) \geq \sup f(A)$ by monotonicity, because $a \leq \sup A$ for all $a \in A$.

For the other inequality, $f(\sup A) \leq \sup f(A)$ we need to also use continuity: let $\varepsilon > 0$; by continuity there exists $\delta > 0$ such that for every $a \in A$ such that $a \geq \sup A - \delta$ it holds that

Table 3: Linear regression experiment with $\sigma_{\mathbf{P}} = \mathbf{0}, \sigma_{\mathbf{Q}} = \mathbf{0}$. Each cell shows the mean over 10 independent runs and the 95% confidence interval in parenthesis.

# samples	Train risk	Test risk	UC bound	WPB bound	KLPB bound
100	0.0211 (0.0010)	0.0208 (0.0001)	6.6176 (0.0010)	0.3175 (0.0177)	undefined
200	0.0206 (0.0009)	0.0208 (0.0001)	4.6850 (0.0009)	0.2363 (0.0136)	undefined
300	0.0214 (0.0006)	0.0209 (0.0001)	3.8298 (0.0006)	0.1989 (0.0087)	undefined
400	0.0205 (0.0005)	0.0208 (0.0001)	3.3187 (0.0005)	0.1824 (0.0127)	undefined

$f(a) \geq f(\sup A) - \varepsilon$ (there exists such a by the definition of the supremum). By monotonicity this implies that $\sup f(A) \geq f(\sup A) - \varepsilon$. Since the latter inequality holds for every $\varepsilon > 0$, we conclude that $\sup f(A) \geq f(\sup A)$ as required. □

Lemma 19 (Wasserstein distance between truncated Gaussian distributions)**.** *Let $X^{(1)}$ and $X^{(2)}$ be the Gaussian random vectors in $\mathbb{R}^d$, with distributions $\mathcal{N}(\mu_1, \Sigma_1)$, and $\mathcal{N}(\mu_1, \Sigma_2)$ respectively. Let $P_{\mathbb{B}_r} : \mathbb{R}^d \to \mathbb{B}_r$ be a projection operator onto the an r-radius ball around the origin, $P_{\mathbb{B}_r}(x) \stackrel{def}{=} \operatorname{argmin}_{x' \in \mathbb{B}_r} \|x - x'\|_2$, where $\mathbb{B}_r = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq r\right\}$. Assume that $r \geq \sqrt{\|\mu_j\|_2^2 + \|\Sigma_j\|_F^2}$ for $j = 1, 2$. Denote the distribution measures of $X^{(1)}$ and $X^{(2)}$ as ν_1 and ν_2 respectively. Let $P_{\mathbb{B}_r}\sharp\nu_1$ and $P_{\mathbb{B}_r}\sharp\nu_2$ be the push-forward measures of ν_1 and ν_2, respectively, under the operator $P_{\mathbb{B}_r}$. Then*

$$W_1(P_{\mathbb{B}_r}\sharp\nu_1, P_{\mathbb{B}_r}\sharp\nu_2) \leq W_1(\nu_1, \nu_2) + \sum_{j=1}^{2} \sqrt{\frac{\pi}{2} \|\Sigma_j\|_{2,2}} \operatorname{erfc}\left(\frac{r - \sqrt{\|\mu_j\|_2^2 + \|\Sigma_j\|_F^2}}{\sqrt{2\|\Sigma_j\|_{2,2}}}\right),$$

where $W_1(Q, P)$ denotes the 1st order Wasserstein distance with the L_2 metric.

Proof. Using the triangle inequality of Wasserstein distances (Clement & Desch, 2008; Thorpe, 2018) twice we get

$$W_1(P_{\mathbb{B}_r}\sharp\nu_1, P_{\mathbb{B}_r}\sharp\nu_2) \leq W_1(P_{\mathbb{B}_r}\sharp\nu_1, \nu_1) + W_1(\nu_1, \nu_2) + W_1(\nu_2, P_{\mathbb{B}_r}\sharp\nu_2). \tag{60}$$

Notice that

$$\begin{aligned} W_1(\nu_2, P_{\mathbb{B}_r}\sharp\nu_2) &= \inf_{\gamma \in \Gamma(\nu_2, P_{\mathbb{B}_r}\sharp\nu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|_2 \, \mathrm{d}\gamma(x, x') \\ &\leq \int_{\mathbb{R}^d} \|x - P_{\mathbb{B}_r}(x)\|_2 \, \mathrm{d}\nu_2(x) \\ &\stackrel{(i)}{=} \int_{\mathbb{R}^d} [\|x\|_2 - r]_+ \mathrm{d}\nu_2(x) \\ &= \mathbb{E}\Big\{[\|X_2\|_2 - r]_+\Big\} \\ &\stackrel{(ii)}{=} \mathbb{E}\big\{[V - r]_+\big\} \\ &\stackrel{(iii)}{=} \int_{t=0}^{\infty} \mathbb{P}\big\{[V - r]_+ > t\big\} \mathrm{d}t \\ &\stackrel{(iv)}{=} \int_{t=0}^{\infty} \mathbb{P}\{V > r + t\} \mathrm{d}t, \end{aligned} \tag{61}$$

where in *(i)* we used the notation $[t]_+ \stackrel{\text{def}}{=} \begin{cases} t & t > 0 \\ 0 & t \leq 0 \end{cases}$ and the equality holds since $P_{\mathbb{B}_r}(x) = x$ for $\|x\| \leq r$, and $\|P_{\mathbb{B}_r}(x) - x\| = \|x\|_2 - r$ for $\|x\| > r$ (by the properties of the projection onto the a L_2 ball). In *(ii)* we use the definition of the random variable $V \stackrel{\text{def}}{=} \left\|X^{(2)}\right\|_2$. In *(iii)* we used the tail sum formula for the expectation. Equality *(iv)* holds since the corresponding events are equivalent.

Let Z be a standard Gaussian random vector in $\mathbb{R}^d$, i.e., $Z \sim \mathcal{N}(\bar{\mathbf{0}}, I)$.

By Wainwright (2019), Thm. 2.26 (one-sided variant) we have that

$$\mathbb{P}\{f(Z) - \mathbb{E}[f(Z)] \geq s\} \leq \exp\left(-\frac{s^2}{2L^2}\right),$$

for any $s \geq 0$ and for any function $f : \mathbb{R}^d \to \mathbb{R}$ that is L-Lipschitz w.r.t. the L_2 metric.

In particular, for the function $f(z) \stackrel{\text{def}}{=} \left\|\mu_2 + \Sigma_2^{1/2} z\right\|_2$ we have for any $s \geq 0$

$$\begin{aligned}\mathbb{P}\left\{\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2 - \mathbb{E}\left[\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2\right] \geq s\right\} &\leq \exp\left(-\frac{s^2}{2\left\|\Sigma_2^{1/2}\right\|_{2,2}^2}\right) \\ &\stackrel{(i)}{=} \exp\left(-\frac{s^2}{2\left\|\Sigma_2\right\|_{2,2}}\right),\end{aligned} \tag{62}$$

since f is Lipschitz with constant $\left\|\Sigma_2^{1/2}\right\|_{2,2}$ where $\|\cdot\|_{2,2}$ is the operator norm defined by $\|A\|_{2,2} = \sup_{\|x\|_2=1} \|Ax\|_2$ for any $A \in \mathbb{R}^{d\times d}$. Equality *(i)* holds since

$$\left\|\Sigma_2^{1/2}\right\|_{2,2}^2 = \left(\sup_{\|x\|_2=1} \left\|\Sigma_2^{1/2} x\right\|_2\right)^2 = \sup_{\|x\|_2=1} \left\|\Sigma_2^{1/2} x\right\|_2^2 = \sup_{\|x\|_2=1} x^\top \Sigma_2 x = \|\Sigma_2\|_{2,2}.$$

Notice that

$$\begin{aligned}\mathbb{E}\left[\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2\right] &= \mathbb{E}\left[\sqrt{\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2^2}\right] \\ &\stackrel{(i)}{\leq} \sqrt{\mathbb{E}\left\{\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2^2\right\}} \\ &= \sqrt{\mathbb{E}\left\{\left(\mu_2 + \Sigma_2^{1/2} Z\right)^\top \left(\mu_2 + \Sigma_2^{1/2} Z\right)\right\}} \\ &\stackrel{(ii)}{=} \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2},\end{aligned}$$

where *(i)* is by Jensen's inequality, and in *(ii)* we used the fact that $Z \sim \mathcal{N}(\bar{\mathbf{0}}, I)$ and $\|\cdot\|_F$ denotes the Frobenius Norm, defined by $\|A\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i,j} A_{i,j}^2}$.

Therefore, using (62), we get

$$\mathbb{P}\left\{\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2 - \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2} \geq s\right\} \leq \exp\left(-\frac{s^2}{2\|\Sigma_2\|_{2,2}}\right).$$

Note that the random variable $V \stackrel{\text{def}}{=} \left\|X^{(2)}\right\|_2$ is equal, in distribution, to the random variable $\left\|\mu_2 + \Sigma_2^{1/2} Z\right\|_2$, and therefore we also have for $s \geq 0$ that

$$\mathbb{P}\left\{V - \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2} \geq s\right\} \leq \exp\left(-\frac{s^2}{2\|\Sigma_2\|_{2,2}}\right).$$

For any $t \geq 0$, set $s := t + r - \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2}$. Since we assume that $r \geq \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2}$, we have that $s \geq 0$. Therefore we have

$$\mathbb{P}\{V > r + t\} \leq \exp\left(-\frac{\left(t + r - \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2}\right)^2}{2\|\Sigma_2\|_{2,2}}\right).$$

Hence, by (61) we have

$$
\begin{aligned}
W_1(\nu_2, P_{\mathbb{B}_r}\sharp\nu_2) &\leq \int_{t=0}^{\infty} \mathbb{P}\{V > r + t\}\mathrm{d}t \\
&\leq \int_{t=0}^{\infty} \exp\left(-\frac{\left(t + r - \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2}\right)^2}{2\,\|\Sigma_2\|_{2,2}} \right) \mathrm{d}t \\
&= \sqrt{\frac{\pi}{2}\,\|\Sigma_2\|_{2,2}}\,\mathrm{erfc}\left(\frac{r - \sqrt{\|\mu_2\|_2^2 + \|\Sigma_2\|_F^2}}{\sqrt{2\,\|\Sigma_2\|_{2,2}}} \right).
\end{aligned}
$$

By symmetry we have a similar bound for $W_1(P_{\mathbb{B}_r}\sharp\nu_1, \nu_1)$. To conclude, using (60) we get

$$
\begin{aligned}
W_1(P_{\mathbb{B}_r}\sharp\nu_1, P_{\mathbb{B}_r}\sharp\nu_2) &\leq W_1(P_{\mathbb{B}_r}\sharp\nu_1, \nu_1) + W_1(\nu_1, \nu_2) + W_1(\nu_2, P_{\mathbb{B}_r}\sharp\nu_2) \\
&\leq W_1(\nu_1, \nu_2) + \sum_{j=1}^{2} \sqrt{\frac{\pi}{2}\,\|\Sigma_j\|_{2,2}}\,\mathrm{erfc}\left(\frac{r - \sqrt{\|\mu_j\|_2^2 + \|\Sigma_j\|_F^2}}{\sqrt{2\,\|\Sigma_j\|_{2,2}}} \right).
\end{aligned}
$$

□