Information Complexity and Generalization Bounds

Pradeep Kr. Banerjee
MPI MiS
Email: pradeep@mis.mpg.de

Guido Montúfar
UCLA and MPI MiS
Email: montufar@math.ucla.edu

***Abstract*—We present a unifying picture of PAC-Bayesian and mutual information-based upper bounds on the generalization error of randomized learning algorithms. As we show, Tong Zhang's information exponential inequality (IEI) gives a general recipe for constructing bounds of both flavors. We show that several important results in the literature can be obtained as simple corollaries of the IEI under different assumptions on the loss function. Moreover, we obtain new bounds for data-dependent priors and unbounded loss functions. Optimizing the bounds gives rise to the Gibbs algorithm, for which we discuss two practical examples for learning with neural networks, namely, Entropy- and PAC-Bayes- SGD. Further, we use an Occam's factor argument to show a PAC-Bayes bound that incorporates second-order curvature information of the training loss.**

The submitted version had a transcription error in Propositions 12 and 20. Corrected places are highlighted in blue.

I. Introduction

The generalization capability of a learning algorithm is intrinsically related to the information that the output hypothesis reveals about the input training dataset: The lesser the information revealed, the better the generalization. This argument has been formalized in recent years by appealing to different notions of information stability [1]–[8]. Information stability quantifies the sensitivity of a learning algorithm to local perturbations of its input, and draws on a rich tradition of earlier work on algorithmic [9]–[11], and distributional [12]–[14] stability in adaptive data analysis. Closely related to the information stability approach is the so-called PAC-Bayesian approach to data-dependent generalization bounds, originally due to McAllester [15]–[17]. While these two approaches have evolved independently of each other, a principle objective of this paper is to present them under a unified framework.

We consider the standard apparatus of statistical learning theory [18]. We have an example domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ of the instances and labels, a hypothesis space $\mathcal{W}$, a fixed loss function $\ell : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$, and a training sample S, which is an n-tuple $(Z_1, \dots, Z_n)$ of i.i.d. random elements of $\mathcal{Z}$ drawn according to some unknown distribution μ. A learning algorithm is a Markov kernel $P_{W|S}$ that maps input training samples S to conditional distributions of hypotheses W in $\mathcal{W}$. This defines a joint distribution $P_{SW} = P_S P_{W|S}$, $P_S = \mu^{\otimes n}$, and a corresponding marginal distribution P_W. The *true risk* of a hypothesis $w \in \mathcal{W}$ on μ is $L_\mu(w) := \mathbb{E}_\mu[\ell(w, Z)]$, and its *empirical risk* on the training sample S is $L_S(w) := \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$. Our goal is to control the *generalization error*, $\mathrm{g}(W, S) := L_\mu(W) - L_S(W)$, either in expectation, or with high probability. One difficulty in achieving this goal is the nontrivial statistical dependency between the sample S and the learned hypothesis W.

For controlling the generalization error in expectation, we can rewrite the true risk of a given hypothesis w as $L_\mu(w) = \mathbb{E}_{S' \sim \mu^{\otimes n}}[L_{S'}(w)]$, where $S' = (Z'_1, \dots, Z'_n)$ is an i.i.d. sample. Then the expected generalization error can be written as a difference of two expectations of the same loss function,

$$\mathbb{E}_{SW}[\mathrm{g}(W, S)] = \mathbb{E}_{P_S \otimes P_W}[L_S(W)] - \mathbb{E}_{P_{SW}}[L_S(W)],$$

where the second expectation is taken w.r.t. the joint distribution of the training sample and the output hypothesis, while the first expectation is taken w.r.t. the product of the two marginal distributions. Hence the expected generalization error reflects the dependence of the output W on the input S. This dependence can also be measured by their mutual information as has been shown in recent works [2]–[8]. We refer to such bounds as mutual information-based generalization bounds.

Alternatively, we may wish to control the generalization error of the learning algorithm $P_{W|S}$ with high probability over the training sample S. The expected generalization error over hypotheses chosen from the distribution P (*posterior*) output by the learning algorithm, i.e., $\mathbb{E}_P[\mathrm{g}(W, S)]$, can be upper-bounded with high probability under P_S by the KL divergence between P and an arbitrary reference distribution Q (*prior*), that is selected *before* the draw of the training sample S. For any Q, these bounds hold uniformly for all P, and are called PAC-Bayesian bounds [15]–[17], [19]–[28], where PAC stands for Probably Approximately Correct. Bounds of this type are useful when we have a fixed dataset $s \in \mathcal{Z}^n$ and a new hypothesis is sampled from P every time the algorithm is used. Choosing the posterior to minimize a PAC-Bayesian bound leads to the well-known Gibbs algorithm [2], [19], [24], [29]. On the other hand, for a fixed posterior P, $\mathbb{E}_S[D(P\|Q)]$ is minimized by the *oracle prior*, $Q^\star = \mathbb{E}_S[P_{W|S}(\cdot|S)]$. Note $\mathbb{E}_S[D(P\|Q^\star)]$ is just the mutual information $I(S; W)$, which is the key quantity controlling the expected generalization error in [2]–[4].

Summary of contributions. We present a unified framework for deriving PAC-Bayesian and mutual information-based generalization bounds, starting from a fundamental information-theoretic inequality, Lemma 4, due to Tong Zhang [19]. Besides recovering several well-known bounds of both flavors, such as the Xu-Raginsky mutual information-bound in Corollary 7, we also obtain new bounds for data-dependent priors and unbounded loss functions. In particular, the bounds in Theorem 6

and Propositions 10, 12, 19, 20 and 22 are new. Optimizing these bounds w.r.t. the posterior gives rise to variants of the Gibbs algorithm, for which we discuss two practical examples and show how a classical bound due to Catoni [20] can be adapted for use in PAC-Bayes-SGD [30]. We also show a PAC-Bayes bound motivated by an Occam's factor argument, in relation to flat minima in neural networks.

II. Preliminaries

We write $\mathcal{M}(\mathcal{W})$ to denote the family of probability measures over a set $\mathcal{W}$, and $\mathcal{K}(\mathcal{S},\mathcal{W})$ to denote the set of Markov kernels from $\mathcal{S}$ to $\mathcal{W}$. Proposition 1 collects some well-known facts about the cumulant generating function $\Lambda_X(\beta)=\ln\mathbb{E}[e^{\beta X}]$ of a random variable X for $\beta>0$ (see, e.g., [31, §2], and [19]):

Proposition 1 (Facts about cumulant generating function)**.**

1) $\Lambda_X(\beta)$ *is convex in* β.
2) $\frac{1}{\beta}\Lambda_X(\beta)$ *is an increasing function of* β.
3) $\Lambda_{X-\mathbb{E}[X]}(0)=\Lambda'_{X-\mathbb{E}[X]}(0)=0$.
4) *For real constants* a,b, $\Lambda_{aX+b}(\beta)=\Lambda_X(a\beta)+b$.
5) $\Lambda_X(\beta)\le\beta\mathbb{E}[X]+\frac{\beta^2}{2}\mathrm{Var}(X)+O(\beta^3)$, *where* $\mathrm{Var}(X)=\mathbb{E}\big[(X-\mathbb{E}[X])^2\big]$.
6) *If* $X\in[0,1]$, *then* $\frac{1}{\beta}\Lambda_X(\beta)\le\frac{1}{\beta}\ln\big(1-(1-e^{\beta})\mathbb{E}[X]\big)$, *with equality when* $X\in\{0,1\}$ *is Bernoulli.*
7) X *is* σ-sub-Gaussian *if* $\Lambda_{X-\mathbb{E}[X]}(\beta)\le\frac{\beta^2\sigma^2}{2}$.
8) X *is* (σ,c)-sub-gamma *if* $\Lambda_{X-\mathbb{E}[X]}(\beta)\le\frac{\beta^2\sigma^2}{2(1-c)}$ *for all* $\beta\in(0,\frac{1}{c})$.

We note the following characterization of the inverse of the Fenchel-Legendre dual of a smooth convex function:

Lemma 2 ([31, Lemma 2.4])**.** *Let* ψ *be a convex and continuously differentiable function defined on the interval* $[0,b)$, *where* $0<b\le\infty$. *Assume that* $\psi(0)=\psi'(0)=0$. *Then, the Legendre dual of* ψ, *defined as* $\psi^*(t):=\sup_{\beta\in[0,b)}\{\beta t-\psi(\beta)\}$, *is a nonnegative convex and nondecreasing function on* $[0,\infty)$ *with* $\psi^*(0)=0$. *Moreover, for every* $y\ge 0$, *the set* $\{t\ge 0:\psi^*(t)>y\}$ *is non-empty and the generalized inverse of* ψ^* *defined by* $\psi^{*-1}(y)=\inf\{t\ge 0:\psi^*(t)>y\}$ *can also be written as* $\psi^{*-1}(y)=\inf_{\beta\in(0,b)}\frac{y+\psi(\beta)}{\beta}$.

We will need the following property of a Gibbs distribution:

Lemma 3 ([19, Proposition 3.1])**.** *For any real-valued measurable function* f *on* $\mathcal{W}$, *any real* $\beta>0$, *and any* $P,Q\in\mathcal{M}(\mathcal{W})$ *such that* $D(P\|Q)<\infty$, *we have* $\beta^{-1}D(P\|P^*)=\mathbb{E}_P[f(W)]+\beta^{-1}D(P\|Q)+\beta^{-1}\ln\mathbb{E}_Q[e^{-\beta f(W)}]$, *where* P^* *is the Gibbs distribution* $P^*(\mathrm{d}w):=\frac{e^{-\beta f(w)}}{\mathbb{E}_Q[e^{-\beta f(W')}]}Q(\mathrm{d}w)$. *Consequently,*

$$\inf_{P\in\mathcal{M}(\mathcal{W})}\big\{\mathbb{E}_P f(W)+\beta^{-1}D(P\|Q)\big\}=-\beta^{-1}\ln\mathbb{E}_Q[e^{-\beta f(W)}].$$

Finally, we recall the *golden formula*: For all $Q\in\mathcal{M}(\mathcal{W})$ such that $D(P_W\|Q)<\infty$, we have

$$I(S;W)=D(P_{W|S}\|Q|P_S)-D(P_W\|Q),\tag{1}$$

where $D(P_{W|S}\|Q|P_S)=\int_{\mathcal{Z}^n}D(P_{W|S=s}\|Q)\mu^{\otimes n}(\mathrm{d}s)$.

All information-theoretic quantities are expressed in *nats*, and we will consistently use the natural logarithm, unless specified otherwise. All proofs are relegated to the full version of the article available on the link below the abstract.

III. One bound to rule them all! Well, almost

A. The Information Exponential Inequality (IEI)

For any real $\beta>0$, define

$$M_\beta(w)=-\beta^{-1}\Lambda_{-\ell(w,Z)}(\beta)=-\beta^{-1}\ln\mathbb{E}_\mu[e^{-\beta\ell(w,Z)}],\tag{2}$$

which acts as a surrogate for $L_\mu(w)$. Following [23], we call this quantity the *annealed expectation*.

Lemma 4 (Information exponential inequality (IEI) [19, Lemma 2.1])**.** *For any prior* $Q\in\mathcal{M}(\mathcal{W})$, *any real-valued loss function* ℓ *on* $\mathcal{W}\times\mathcal{Z}$, *and any posterior distribution* $P\ll Q$ *over* $\mathcal{W}$ *that depends on an i.i.d. training sample* S, *we have* $\mathbb{E}_S\exp\big\{n\beta\mathbb{E}_P[M_\beta(W)-L_S(W)]-D(P\|Q)\big\}\le 1$.

The IEI implies bounds both in probability and in expectation for the quantity $n\beta\mathbb{E}_P[M_\beta(W)-L_S(W)]-D(P\|Q)$, and is the key tool for showing the following theorem due to Tong Zhang that holds for unbounded loss functions:

Theorem 5 ([19, Theorem 2.1])**.** *Let* μ *be a distribution over* $\mathcal{Z}$, *and let* S *be an i.i.d. training sample from* μ. *Let* $Q\in\mathcal{M}(\mathcal{W})$ *be a prior distribution that does not depend on* S, *and let* ℓ *be a real-valued loss function on* $\mathcal{W}\times\mathcal{Z}$. *Let* $\beta>0$, *and let* $\delta\in(0,1]$. *Then, with probability of at least* $1-\delta$ *over the choice of* $S\sim\mu^{\otimes n}$, *for all distributions* $P\ll Q$ *over* $\mathcal{W}$ *(even such that depend on* S*), we have:*

$$\mathbb{E}_P[M_\beta(W)]\le\mathbb{E}_P[L_S(W)]+\frac{1}{n\beta}\left(D(P\|Q)+\ln\frac{1}{\delta}\right).\tag{3}$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{SW}[M_\beta(W)]\le\mathbb{E}_{SW}[L_S(W)]+\frac{1}{n\beta}D(P\|Q|S).\tag{4}$$

It is useful to replace the annealed expectation $M_\beta(w)$ in (3) and (4) with the true risk $L_\mu(w)$. By Proposition 1 item *1)* and Jensen's inequality, we have $M_\beta(w)\le L_\mu(w)$. For general loss functions, Proposition 1 item *5)* is useful for getting bounds in the opposite direction. By items *4)*, *7)* and *8)* of Proposition 1, if for all $w\in\mathcal{W}$, $\ell(w,Z)$ is σ-sub-Gaussian, resp., (σ,c)-sub-gamma under μ, then we have for all $w\in\mathcal{W}$ and $\beta>0$, $L_\mu(w)\le M_\beta(w)+\frac{\beta}{2}\sigma^2$, resp., $L_\mu(w)\le M_\beta(w)+\frac{\beta}{2(1-c)}\sigma^2$. More generally, we note the following result, which follows as a corollary to Theorem 5 and Lemma 2:

Theorem 6. *Suppose that there exist a convex function* $\psi\colon\mathbb{R}_{\ge 0}\to\mathbb{R}$ *satisfying* $\psi(0)=\psi'(0)=0$, *such that*

$$\sup_{w\in\mathcal{W}}[L_\mu(w)-M_\beta(w)]\le\frac{\psi(\beta)}{\beta},\ \forall\beta>0.\tag{5}$$

Then, under the setting of Theorem 5, with probability of at least $1-\delta$ *over the choice of* $S\sim\mu^{\otimes n}$, *for all distributions*

$P \ll Q$ over $\mathcal{W}$ (even such that depend on S), we have

$$\mathbb{E}_P[\mathrm{g}(W,S)] \le \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\psi(\beta)}{\beta}. \quad (6)$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{SW}[\mathrm{g}(W,S)] \le \psi^{*-1}\left(\frac{D(P\|Q|P_S)}{n}\right). \quad (7)$$

As discussed in the introduction, by the golden formula (1), under the oracle prior $Q^\star = \mathbb{E}_S[P_{W|S}]$, $\mathbb{E}_S[D(P\|Q^\star)] = I(S;W)$. If $\ell(w,Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then we can take $\psi(\beta) = \frac{\beta^2\sigma^2}{2}$ and $\psi^{*-1}(y) = \sqrt{2\sigma^2 y}$ [31, §2.3], in which case we recover the following bound in expectation due to Xu and Raginsky [2]:

Corollary 7. *If $\ell(w,Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then $\mathbb{E}_{SW}[\mathrm{g}(W,S)] \le \sqrt{2\sigma^2 I(S;W)/n}$.*

Likewise, for a (σ,c)-sub-gamma under μ, we can take $\psi(\beta) = \frac{\beta^2\sigma^2}{2(1-c)}$ and $\psi^{*-1}(y) = \sqrt{2\sigma^2 y} + cy$ [31, §2.4], which gives the following result:

Corollary 8. *If $\ell(w,Z)$ is (σ,c)-sub-gamma under μ for all $w \in \mathcal{W}$, then $\mathbb{E}_{SW}[\mathrm{g}(W,S)] \le \sqrt{2\sigma^2 I(S;W)/n} + cI(S;W)/n$.*

Fixing $\beta = 1$ in (6), we recover [26, Corollary 5]:

Corollary 9. *Consider the setting in Theorem 5. If the loss ℓ is (σ,c)-sub-gamma with $c < 1$, then with probability of at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions $P \ll Q$ over $\mathcal{W}$, $\mathbb{E}_P[\mathrm{g}(W,S)] \le \frac{1}{n}(D(P\|Q) + \ln(1/\delta)) + \frac{\sigma^2}{2(1-c)}$.*

The condition $c < 1$ guarantees that $\beta = 1 \in (0, \frac{1}{c})$ when the sub-gamma condition in Proposition 1 *8)* is satisfied. In the limit $c \to 0_+$, a sub-gamma loss reduces to the sub-Gaussian loss [31, §2.4], and we recover [26, Corollary 4]. We can optimize β in (6) at a small cost using the union bound:

Proposition 10. *Consider the setting in Theorem 5. If $\ell(w,Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0,1]$, for all $\beta \in (0,v]$, with probability of at least $1-\delta$, we have*

$$\mathbb{E}_P[\mathrm{g}(W,S)] \le \frac{\alpha}{n\beta}\left(D(P\|Q) + \ln\frac{\log_\alpha\sqrt{n} + K}{\delta}\right) + \frac{\beta\sigma^2}{2},$$

where $K = \max\{\log_\alpha\left(\frac{v\sigma}{\sqrt{2\alpha}}\right), 0\} + e$.

B. The Conditional Mutual Information (CMI) bound

One drawback of the mutual information-based bounds in Corollaries 7 and 8 is that $I(S;W)$ can be unbounded in many practical situations of interest [5], [7]. CMI-based bounds [7], [8] address this issue by conditioning on a superset of the training sample called the *supersample*, in effect, normalizing the information content of each datum to one bit. As nicely articulated by Steinke and Zakynthinou [7], intuitively, the difference between the CMI- and MI-based approaches is that between "recognizing" vs. "reconstructing" the input, given the output of the algorithm. Recognizing the input is formalized by considering a i.i.d. supersample $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ consisting of $n \times 2$ data points, which comprises of n "true" input data points mixed with n "ghost" data points. A selector variable $U \in \{0,1\}^n$ chooses the input samples from the supersample, uniformly at random. Given the output of the algorithm, CMI then measures how well it is possible to distinguish the true inputs from their ghosts. We note the following definition:

Definition 11 (CMI of an algorithm $P_{W|S}$ [7])**.** *Let μ be a probability distribution on $\mathcal{Z}$ and let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ consist of $2n$ i.i.d. samples drawn from μ. Let $U = (U_1,\dots,U_n) \in \{0,1\}^n$ be uniformly random and independent from $\tilde{Z}$ and the randomness of the algorithm. Define $S := \tilde{Z}_U \in \mathcal{Z}^n$ by $(\tilde{Z}_U)_i = \tilde{Z}_{i,U_i+1}$ for all $i \in [n]$, i.e., S is the subset of $\tilde{Z}$ indexed by U. Then the conditional mutual information (CMI) of an algorithm $P_{W|S}$ w.r.t. μ is $\mathsf{CMI}_\mu(P_{W|S}) := I(W;U|\tilde{Z})$.*

In [7, Theorem 2(1)], it is shown that for a $[0,1]$-valued loss, $\mathbb{E}_{SW}[\mathrm{g}(W,S)] \le \sqrt{2\cdot\mathsf{CMI}_\mu(P_{W|S})/n}$. Unlike the mutual information $I(S;W)$ that can be potentially unbounded, $\mathsf{CMI}_\mu(P_{W|S})$ is bounded above by $n\ln 2$.

We give a high probability version of the CMI bound in Proposition 12. Let $\bar{U} = (\bar{U}_1,\dots,\bar{U}_n)$ be a vector obtained by inverting all the bits of U, and define $\bar{S} = \tilde{Z}_{\bar{U}}$. S and $\bar{S}$ have a common marginal distribution, $\mu^{\otimes n}$. The algorithm maps the input $S = \tilde{Z}_U$ to a random element W of $\mathcal{W}$. Since $\bar{S} \perp\!\!\!\perp W$, we can define the generalization error as $\mathrm{g}(W,\tilde{Z},U) := L_{\bar{S}}(W) - L_S(W)$, where $L_{\bar{S}}(w) := \frac{1}{n}\sum_{i=1}^n \ell(w,(\tilde{Z}_{\bar{U}})_i)$, and $L_S(w) := \frac{1}{n}\sum_{i=1}^n \ell(w,(\tilde{Z}_U)_i)$. Given a realization of the supersample $\tilde{Z} = \tilde{z}$ and selector variable $U = u$, we write $Q \equiv Q_{W|\tilde{z}}$ and $P \equiv P_{W|\tilde{z},u}$ for, resp., the prior and the posterior distribution. Then the following bounds hold for all such prior and posterior distributions:

Proposition 12. *For any $[0,1]$-valued loss function ℓ, for any $\beta > 0$ and $\delta \in (0,1]$, with probability of at least $1-\delta$ over a draw of $\tilde{Z},U$ as defined above, we have:*

$$\mathbb{E}_P[\mathrm{g}(W,\tilde{Z},U)] \le \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\beta}{2}. \quad (8)$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{W,\tilde{Z},U}[\mathrm{g}(W,\tilde{Z},U)] \le \sqrt{\frac{2\cdot D(P\|Q|P_{\tilde{Z},U})}{n}}. \quad (9)$$

Using the same reasoning as earlier, supplanting the associated oracle prior recovers the bound in expectation $\sqrt{2\cdot\mathsf{CMI}_\mu(P_{W|S})/n}$ in [7, Theorem 2(1)]. In the full version, we also show the following "single-draw" bound: $\Pr_{W,\tilde{Z},U}\left(|\mathrm{g}(W,\tilde{Z},U)| > \varepsilon\right) \le O(\mathsf{CMI}_\mu(P_{W|S})/n\varepsilon^2)$.

C. Recovering classical PAC-Bayesian bounds

By Proposition 1 item *6)*, for a $\{0,1\}$-valued loss, we have $M_\beta(w) = -\beta^{-1}\ln\big(1-(1-e^{-\beta})L_\mu(w)\big) =: \Phi_\beta(L_\mu(w))$. Φ_β is an increasing one-to-one mapping of the unit interval onto itself, and is convex for $\beta > 0$ [20]. The inverse of Φ_β is given by $\Phi_\beta^{-1}(x) = \frac{1-e^{-\beta x}}{1-e^{-\beta}}$, $x \in (0,1)$, and we recover Catoni's PAC-Bayesian bound:

Theorem 13 (Catoni's bound [20, Theorem 1.2.6])**.** *For any $\{0,1\}$-valued loss ℓ, any distribution μ, prior $Q \in \mathcal{M}(\mathcal{W})$, any real $\beta > 0$, and any $\delta \in (0,1]$, with probability of at least $1-\delta$ over $S \sim \mu^{\otimes n}$, we have for all $P \ll Q$ over $\mathcal{W}$:*

$$\mathbb{E}_P[L_\mu(W)] \le \Phi_\beta^{-1}\Big\{\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta}\Big]\Big\}.$$

Starting from Catoni's bound and using a series of approximations, one can recover [20, Theorem 1.2.1] and McAllester's "Linear PAC-Bayes bound" [17, Theorem 2]. For loss functions bounded in $[0,1]$, we elaborate in the full version on other approximations that lead to several well-known PAC-Bayesian inequalities such as the "PAC-Bayes-KL inequality" [21], [32].

Remark 14 (Related work)**.** *A variation of the IEI for the special case of the 0-1 loss appears in the monograph by Catoni [20, Eq.1.2], and has been rediscovered more recently for the sub-Gaussian loss in [33], [34]. The statements of [33, Corollary 3, Eq.20] and [34, Corollary 6, Eq.95] which are analogues of our Propositions 10 and 12, resp., are incorrect as they assume that β can be optimized "for free," when in fact we have to pay a union bound price for optimizing β, which is selected* before *the draw of the training sample. We also note two related works that focus exclusively on unifying either PAC-Bayesian bounds for the 0-1 loss [35], or information-theoretic bounds for the sub-Gaussian loss [36].*

IV. Differentially private data-dependent priors

A PAC-Bayesian bound such as (3) stipulates that the prior Q be chosen before the draw of the training sample S. Q may, however, depend on the data generating distribution μ [37]. To have a good control over the KL term in (3), it is desirable that Q be "aligned" with the data-dependent posterior P. One way to achieve this goal is to choose Q based on S in a differentially private fashion so that Q is stable to local perturbations in S [38]. We can then treat Q "as if" it is independent of S. Here, the key quantity of interest is the approximate max-information between the input S and the data-dependent prior. We shall make these notions precise.

For $\alpha \ge 0$, the *α-approximate max-divergence* is defined as $D_\infty^\alpha(P\|Q) = \ln \sup_{\mathcal{O}\subseteq\mathcal{X}:\, P(\mathcal{O})>\alpha} \frac{P(\mathcal{O})-\alpha}{Q(\mathcal{O})}$. The *max-divergence* $D_\infty(P\|Q)$ is defined as $D_\infty^\alpha(P\|Q)$ for $\alpha = 0$. For a pair of variables (X,Y) with joint law P_{XY} and marginals P_X and P_Y, the *α-approximate max-information* between X and Y is defined as $I_\infty^\alpha(X;Y) = D_\infty^\alpha(P_{XY}\|P_X \otimes P_Y)$. The *max-information* $I_\infty(X;Y)$ is defined to be $I_\infty^\alpha(X;Y)$ for $\alpha = 0$. $I_\infty(X;Y)$ is an upper bound on the ordinary mutual information $I(X;Y)$ [12].

Definition 15 (Differential Privacy [39])**.** *For any $\epsilon > 0$ and $\delta \in [0,1]$, an algorithm $P_{W|S}$ is said to be* (ϵ,δ)-differentially private *if for all pairs of datasets $s, s' \in \mathcal{Z}^n$ that differ in a single element, $D_\infty^\delta(P_{W|S=s}\|P_{W|S=s'}) \le \epsilon$. The case $\delta = 0$ is called pure differential privacy.*

Definition 16 (Max-Information of an algorithm [12])**.** *We say that an algorithm $P_{W|S}$ has α-approximate max-information of k, denoted as $I_{\infty,\mu}^\alpha(P_{W|S}, n) \le k$, if for every distribution μ over $\mathcal{Z}$, we have $I_\infty^\alpha(S;W) \le k$ when $S \sim \mu^{\otimes n}$.*

It follows from the definition of α-approximate max-information that if an algorithm $P_{W|S}$ has bounded approximate max-information, then we can control the probability of "bad events" that may arise as a result of the dependence of the output W on the input S [12]. Let $S' \perp\!\!\!\perp W$ be an independent sample with the same distribution as S. If for some $\alpha \ge 0$, $I_\infty^\alpha(S;W) \le k$, then for any event $\mathcal{O} \subseteq \mathcal{Z}^n \times \mathcal{W}$, we have

$$\Pr((S,W) \in \mathcal{O}) \le e^k \cdot \Pr((S',W) \in \mathcal{O}) + \alpha. \tag{10}$$

Pure differential privacy implies a bound on the approximate max-information:

Theorem 17 (Pure differential privacy and α-approximate max-information [12, Theorem 20])**.** *If $P_{W|S}$ is an $(\epsilon,0)$-differentially private algorithm, then $I_{\infty,\mu}(P_{W|S},n) \le \epsilon n$, and for any $\alpha > 0$, $I_{\infty,\mu}^\alpha(P_{W|S},n) \le n\epsilon^2/2 + \epsilon\sqrt{n\ln(2/\alpha)/2}$.*

Remark 18. *The result above is extended to (ϵ,δ)-differential privacy in [13, Theorem 3.1]: If $P_{W|S}$ is an (ϵ,δ)-differentially private algorithm for $\epsilon \in (0,1/2]$ and $\delta \in (0,\epsilon)$, then for $\alpha = O(n\sqrt{\delta/\epsilon})$, $I_{\infty,\mu}^\alpha(\mathcal{A},n) = O(n\epsilon^2 + n\sqrt{\delta/\epsilon})$.*

Proposition 19. *Consider the setting in Theorem 5. Let $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{W})$ be an $(\epsilon,0)$-differentially private algorithm. Then with probability of at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, for all $P \in \mathcal{M}(\mathcal{W})$,*

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{D(P\|Q^0(S)) + f_n(\delta,\epsilon)}{n\beta}.$$

where $f_n(\delta,\epsilon) := \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}}$.

By Remark 18, the result above can be extended to (ϵ,δ)-differentially private priors. Proposition 19 is valid for any loss function and is similar in spirit to the traditional PAC-Bayesian bounds in [38, Theorem 4.2], and [28, Equation 7], which however only apply when the loss function is bounded in $[0,1]$. We can also bound the expected generalization error. The next result follows from Theorem 17 and Proposition 12:

Proposition 20. *Consider the setting in Proposition 12 with $\mathrm{g}(W,\tilde{Z},U)$, and P as defined there. Let $Q^0 \in \mathcal{K}(\mathcal{Z}^{n\times 2} \times \{0,1\}^n, \mathcal{W})$ be an $(\epsilon,0)$-differentially private algorithm. Then with probability of at least $1-\delta$ over a draw of $\tilde{Z},U$, for all P, we have*

$$\mathbb{E}_P[\mathrm{g}(W,\tilde{Z},U)] \le \frac{D(P\|Q^0(\tilde{Z},U)) + f_n(\delta,\epsilon)}{n\beta} + \frac{\beta}{2}, \tag{11}$$

where $f_n(\delta,\epsilon) := \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}}$.

The main advantage of the max-information formulation is that we can get high probability guarantees at the cost of a $O(n\epsilon^2 + \epsilon\sqrt{n\ln 1/\delta})$ correction term. This cost is compensated for by a lower KL complexity since the prior is more "aligned" with the data-dependent posterior than when chosen independently of the data. As is well-known [12], [14],

a small mutual information between the data and the prior will not ensure that bad events will happen with low probability.

V. Information complexity minimization

Given any prior Q, minimizing the right hand side of (3) gives rise to the *Information Complexity Minimization (ICM)* framework [19], [40]. Concretely, for a given prior Q and hypothesis set $\mathcal{G} \subseteq \mathcal{M}(\mathcal{W})$, define the *Optimal Information Complexity (*OIC*)* at a given β as

$$\mathrm{OIC}^{\beta}_{\mathcal{G}} := \inf_{P \in \mathcal{G}} \Big\{ \mathbb{E}_P[L_S(W)] + (n\beta)^{-1} D(P\|Q) \Big\}. \quad (12)$$

When $\mathcal{G} = \mathcal{M}(\mathcal{W})$, applying Lemma 3 to $f(w) = nL_S(w)$, and writing β for $n\beta$, we recover the Gibbs algorithm, $P^\star$, in which case the OIC evaluates to the *(extended) stochastic complexity* $-\frac{1}{\beta}\ln \mathbb{E}_Q[e^{-\beta L_S(W)}]$ [41], [42]. The latter in turn coincides with the log Bayesian marginal likelihood for $\beta = 1$ and the log loss [19], [40], [43]. We briefly discuss two practical examples of ICM for learning with neural networks (NNs).

A. PAC-Bayes-SGD

PAC-Bayes-SGD is an approach to computing nonvacuous generalization bounds for overparameterized NN classifiers trained with stochastic gradient descent (SGD) [30], [44], [45]. These bounds are obtained by *retraining* the network using an objective derived from a PAC-Bayes bound, starting from the solution found by SGD (or in fact any other procedure) for the training loss $L_S(w)$ w.r.t. w. We show how the bound in Theorem 13 can be adapted for use in PAC-Bayes-SGD.

Consider a binary classification setting with examples domain $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and loss $\ell \colon \mathbb{R}^k \times \mathcal{Z} \to \{0,1\}$. Each $w \in \mathcal{W}$ corresponds to a classifier $f_w \colon \mathcal{X} \to \{0,1\}$ that can be interpreted as a deterministic NN with parameters in $\mathbb{R}^k$. For trainable parameters $w_P \in \mathbb{R}^k$, $\gamma \in \mathbb{R}^k_+$, $\lambda \in \mathbb{R}_+$, let $\mathcal{G}$ be the set of all Gaussian posteriors of the form $P = \mathcal{N}(w_P, \mathrm{diag}(\gamma))$ and let $Q = \mathcal{N}(w_0, \lambda I_k)$ be a prior centered at a non-trainable random initialization, $w_0 \in \mathbb{R}^k$. We use a convex surrogate of the 0-1 loss, and the reparameterization trick $w = w_P + \nu \odot \sqrt{\gamma}$, $\nu \sim \mathcal{N}(0, I_k)$ [46] to compute an unbiased estimate of the gradient of the PAC-Bayes bound in Theorem 13 w.r.t. the parameters w_P, γ, λ and β. Computing the expectation $\mathbb{E}_P[L_S(f_W)]$ is difficult in practice. Instead, we use a Monte Carlo estimate $\hat{L}_S(f_W) = \frac{1}{m}\sum_{i=1}^m L_S(f_{W_i})$, where $W_i \overset{\text{i.i.d.}}{\sim} P$. Then Theorem 13 takes the form: For any $\delta, \delta' \in (0,1)$, fixed $\alpha > 1$, $c \in (0,1)$, $b \in \mathbb{N}$, and $m, n \in \mathbb{N}$, with probability of at least $1 - \delta - \delta'$ over a draw of $S \sim \mu^{\otimes n}$ and $W \sim (P)^{\otimes m}$,

$$\mathbb{E}_P[L_\mu(f_W)] \le \inf_{P \in \mathcal{G}, \beta > 1, \lambda \in (0,c)} \Phi_\beta^{-1}\Big\{ \mathbb{E}_P[\hat{L}_S(f_W)] + \frac{\alpha}{n\beta} D(P\|Q) + R(\lambda, \beta; \delta, \delta') \Big\},$$

where $R = \frac{2\alpha}{n\beta}\ln\Big(\frac{\ln \alpha^2 \beta n}{\ln \alpha}\Big) + \frac{\alpha}{n\beta}\ln\Big[\frac{\pi^2 b^2}{6\delta}\big(\ln\frac{c}{\lambda}\big)^2\Big] + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta'}}$ accounts for the cost of optimizing the parameters β, λ, and using the Monte Carlo estimate of the empirical risk. For large n, m, R is negligible, and the optimization is dominated by the IC term, $\mathbb{E}_P[\hat{L}_S(f_W)] + \alpha(n\beta)^{-1} D(P\|Q)$.

B. Entropy-SGD

A related approach is *Entropy-SGD* [47], which instead directly minimizes the stochastic complexity $-\frac{1}{\beta}\ln \mathbb{E}_Q e^{-\beta L_S(W)}$. This entails optimizing the prior Q, when a PAC-Bayesian bound such as (3) stipulates that the prior be fixed before the draw of the training sample S. A way out is to sample Q in a differentially private fashion, and this forms the basis of the Entropy-SGLD algorithm [48]. For $Q = \mathcal{N}\big(w, (\beta\gamma)^{-1} I_k\big)$, the stochastic complexity can be equivalently written as (up to constant terms) $-\frac{1}{\beta}\ln \int_{w' \in \mathbb{R}^k} e^{-\beta\left[L_S(w') + \frac{\gamma}{2}\|w - w'\|^2\right]} \mathrm{d}w'$, which can be interpreted as a measure of *flatness* of the loss landscape that measures the log-volume of low-loss parameter configurations around w. More generally, from the perspective of ICM, both Entropy- and PAC-Bayes-SGD can be viewed as optimization schemes that search for flat minima solutions.

C. PAC-Bayes and Occam's factor

Lemma 21 gives the form of the optimal posterior under a quadratic approximation of the loss around a local minimizer:

Lemma 21. *Consider a quadratic approximation of the training loss around a local minimizer w_P, $\tilde{L}_S(w) = \frac{1}{2}(w - w_P)^\top H (w - w_P)$, a fixed prior $Q = \mathcal{N}(w_Q, \lambda^{-1} I_k)$, and a posterior distribution of the form $P = \mathcal{N}(w_P, \Sigma_P)$. Then the solution to the convex optimization problem $\min_{\Sigma_P} \mathbb{E}_P[\tilde{L}_S(W)] + (n\beta)^{-1} D(P\|Q)$, is given by $\Sigma_P^\star = H_\lambda^{-1}$, where $H_\lambda := (n\beta H + \lambda I_k)$. Here we assume $\lambda > 0$ is sufficiently large so that H_λ is positive definite.*

We can use a posterior of the form $P = \mathcal{N}(w_P, H_\lambda^{-1})$ to get the following PAC-Bayesian bound that incorporates second-order curvature information of the training loss:

Proposition 22. *Let $\{\lambda_i\}_{i=1}^k$ be the eigenvalues of H_λ and suppose that $\lambda_i \ge \lambda > 0$ for all i. Let $Q = \mathcal{N}(w_Q, \lambda^{-1} I_k)$ be a prior, and let $P = \mathcal{N}(w_P, H_\lambda^{-1})$. Then with probability of at least $1 - \delta$ over a draw of the sample S, we have*

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\ln\frac{1}{\delta} + \frac{1}{n\beta}\left(\frac{\lambda}{2}\|w_Q - w_P\|^2 + \frac{1}{2}\sum_{i=1}^k \ln\frac{\lambda_i}{\lambda}\right). \quad (13)$$

The log-ratio term $\frac{1}{2}\sum_{i=1}^k \ln\frac{\lambda_i}{\lambda} = -\ln\sqrt{\det\frac{\lambda}{H_\lambda}}$ in (13) is the negative logarithm of the *Occam factor* [49], [50]. The log-Occam factor is the differential entropy of the Gaussian posterior with a scaled covariance $\lambda(H_\lambda)^{-1}$, and can be interpreted as the amount of information we gain about the model's parameters after seeing the training data. From the perspective of ICM, minimizing the right hand side of (13) w.r.t. the posterior leads to solutions with higher entropy and hence wider minima.

Acknowledgment

This project has received funding from the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (grant agreement n° 757983).

REFERENCES

[1] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016, pp. 1232–1240.

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, pp. 2524–2533.

[3] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1475–1479.

[4] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," in *International Conference on Algorithmic Learning Theory (ALT)*, 2018, pp. 25–55.

[5] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 587–591.

[6] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 582–586.

[7] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference On Learning Theory*, 2020, pp. 3437–3452.

[8] ——, "Open problem: Information complexity of VC learning," in *Conference on Learning Theory*, 2020, pp. 3857–3863.

[9] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. Mar, pp. 499–526, 2002.

[10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *The Journal of Machine Learning Research*, vol. 11, pp. 2635–2670, 2010.

[11] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, 2016, pp. 1046–1059.

[12] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Advances in Neural Information Processing Systems*, 2015, pp. 2350–2358.

[13] R. Rogers, A. Roth, A. Smith, and O. Thakkar, "Max-information, differential privacy, and post-selection hypothesis testing," in *57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2016, pp. 487–494.

[14] V. Feldman and T. Steinke, "Calibrating noise to variance in adaptive data analysis," in *Conference On Learning Theory*, 2018, pp. 535–544.

[15] D. A. McAllester, "PAC-Bayesian model averaging," in *Proceedings of the 12th Annual Conference on Computational learning theory*. ACM, 1999, pp. 164–170.

[16] ——, "Some PAC-Bayesian theorems," *Machine Learning*, vol. 37, no. 3, pp. 355–363, 1999.

[17] ——, "A PAC-Bayesian tutorial with a dropout bound," *arXiv preprint arXiv:1307.2118*, 2013.

[18] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[19] T. Zhang, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[20] O. Catoni, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007, vol. 56.

[21] A. Maurer, "A note on the PAC Bayesian theorem," *arXiv preprint cs/0411099*, 2004.

[22] T. van Erven, "PAC-Bayes mini-tutorial: A continuous union bound," *arXiv preprint arXiv:1405.1580*, 2014.

[23] P. D. Grünwald and N. A. Mehta, "Fast rates for general unbounded loss functions: From ERM to generalized Bayes." *Journal of Machine Learning Research*, vol. 21, no. 56, pp. 1–80, 2020.

[24] P. Alquier, J. Ridgway, and N. Chopin, "On the properties of variational approximations of Gibbs posteriors," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, 2016.

[25] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand, "PAC-Bayesian learning of linear classifiers," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009, pp. 353–360.

[26] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien, "PAC-Bayesian theory meets Bayesian inference," in *Advances in Neural Information Processing Systems*, 2016, pp. 1884–1892.

[27] N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin, "A strongly quasiconvex PAC-Bayesian bound," in *International Conference on Algorithmic Learning Theory (ALT)*, 2017, pp. 466–492.

[28] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor, "PAC-Bayes analysis beyond the usual bounds," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[29] I. Kuzborskij, N. Cesa-Bianchi, and C. Szepesvári, "Distribution-dependent analysis of Gibbs-ERM principle," in *Conference on Learning Theory*, 2019, pp. 2028–2054.

[30] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

[31] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[32] M. Seeger, "PAC-Bayesian generalisation error bounds for Gaussian process classification," *Journal of Machine Learning Research*, vol. 3, no. Oct, pp. 233–269, 2002.

[33] F. Hellström and G. Durisi, "Generalization error bounds via mth central moments of the information density," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2741–2746.

[34] ——, "Generalization bounds via information density and conditional information density," *IEEE Journal on Selected Areas in Information Theory*, pp. 824–839, 2020.

[35] A. Blum and J. Langford, "PAC-MDL bounds," in *Learning theory and kernel machines*. Springer, 2003, pp. 344–357.

[36] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[37] G. Lever, F. Laviolette, and J. Shawe-Taylor, "Tighter PAC-Bayes bounds through distribution-dependent priors," *Theoretical Computer Science*, vol. 473, pp. 4–28, 2013.

[38] G. K. Dziugaite and D. M. Roy, "Data-dependent PAC-Bayes priors via differential privacy," in *Advances in Neural Information Processing Systems*, 2018, pp. 8430–8441.

[39] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[40] T. Zhang, "From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, 2006.

[41] J. Rissanen, *Stochastic complexity in statistical inquiry*. WS, 1989.

[42] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its applications to learning," *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1424–1439, 1998.

[43] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.

[44] J. Langford and R. Caruana, "(Not) bounding the true error," in *Advances in Neural Information Processing Systems*, 2002, pp. 809–816.

[45] G. E. Hinton and D. van Camp, "Keeping neural networks simple by minimising the description length of weights," in *Conference On Learning Theory*, 1993, pp. 5–13.

[46] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 1613–1622.

[47] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, "Entropy-SGD: Biasing gradient descent into wide valleys," in *International Conference on Learning Representations*, 2017.

[48] G. K. Dziugaite and D. Roy, "Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 1377–1386.

[49] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[50] S. L. Smith and Q. V. Le, "A Bayesian perspective on generalization and stochastic gradient descent," in *International Conference on Learning Representations*, 2018.

APPENDIX

A. Proofs for Section III-A

Lemma 23 (Donsker-Varadhan [31, Corollary 4.15])**.** *Let P, Q be probability measures on $\mathcal{W}$, and let $\mathcal{F}$ denote the set of measurable functions $f : \mathcal{W} \to \mathbb{R}$ such that $\mathbb{E}_Q[e^{f(W)}] < \infty$. If $D(P\|Q) < \infty$, then for every $f \in \mathcal{F}$, we have*

$$D(P\|Q) = \sup_{f\in\mathcal{F}} \Big\{\mathbb{E}_P[f(W)] - \ln \mathbb{E}_Q[e^{f(W)}]\Big\},$$

where the supremum is attained when $f = \ln \frac{\mathrm{d}P}{\mathrm{d}Q}$.

We include a proof of Lemma 4 and Theorem 5, since we will use the arguments.

Lemma 4 (Information exponential inequality (IEI) [19, Lemma 2.1])**.** *For any prior $Q \in \mathcal{M}(\mathcal{W})$, any real-valued loss function ℓ on $\mathcal{W} \times \mathcal{Z}$, and any posterior distribution $P \ll Q$ over $\mathcal{W}$ that depends on an i.i.d. training sample S, we have $\mathbb{E}_S \exp\big\{n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)\big\} \le 1$.*

Proof of Lemma 4. Applying the Donsker-Varadhan Lemma 23 to the function,

$$f(w) = n\beta(M_\beta(w) - L_S(w)), \tag{14}$$

we obtain,

$$n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q) \le \ln \mathbb{E}_Q[e^{n\beta(M_\beta(W)-L_S(W))}]. \tag{15}$$

Exponentiating both sides of (15) and taking expectations w.r.t. $S \sim \mu^{\otimes n}$, we have

$$\mathbb{E}_S \exp\big\{n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)\big\} \le \mathbb{E}_S\mathbb{E}_Q[e^{n\beta(M_\beta(W)-L_S(W))}]. \tag{16}$$

Since $Z_i \overset{\text{i.i.d.}}{\sim} \mu$, for any $w \in \mathcal{W}$ and $\beta > 0$, we have $e^{-n\beta M_\beta(w)} = \mathbb{E}_{S\sim\mu^{\otimes n}}\big[e^{-n\beta L_S(w)}\big]$. This observation and Fubini's theorem implies that the right hand side of (16) is equal to one. This proves the IEI. □

Theorem 5 ([19, Theorem 2.1])**.** *Let μ be a distribution over $\mathcal{Z}$, and let S be an i.i.d. training sample from μ. Let $Q \in \mathcal{M}(\mathcal{W})$ be a prior distribution that does not depend on S, and let ℓ be a real-valued loss function on $\mathcal{W} \times \mathcal{Z}$. Let $\beta > 0$, and let $\delta \in (0, 1]$. Then, with probability of at least $1 - \delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions $P \ll Q$ over $\mathcal{W}$ (even such that depend on S), we have:*

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right). \tag{3}$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{SW}[M_\beta(W)] \le \mathbb{E}_{SW}[L_S(W)] + \frac{1}{n\beta}D(P\|Q|S). \tag{4}$$

Proof of Theorem 5. Letting $R(S) := n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)$, by Lemma 4, we have $\mathbb{E}_S[e^{R(S)}] \le 1$. By Markov's inequality, we have

$$\Pr_S\left(R(S) > \ln\frac{1}{\delta}\right) = \Pr_S\left(e^{R(S)} > \frac{1}{\delta}\right) \le \mathbb{E}_S[e^{R(S)}]\delta \le \delta,$$

when (3) follows. By Jensen's inequality, we have $e^{\mathbb{E}_S[R(S)]} \le \mathbb{E}_S[e^{R(S)}] \le 1$, which implies $\mathbb{E}_S[R(S)] \le 0$, when (4) follows. □

Proposition 10. *Consider the setting in Theorem 5. If $\ell(w, Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0, 1]$, for all $\beta \in (0, v]$, with probability of at least $1 - \delta$, we have*

$$\mathbb{E}_P[\mathrm{g}(W, S)] \le \frac{\alpha}{n\beta}\left(D(P\|Q) + \ln\frac{\log_\alpha\sqrt{n} + K}{\delta}\right) + \frac{\beta\sigma^2}{2},$$

where $K = \max\{\log_\alpha\left(\frac{v\sigma}{\sqrt{2\alpha}}\right), 0\} + e$.

The proof follows that of [22, Lemma 8], extending it to sub-Gaussian losses.

Proof of Proposition 10. For $0 < u < v$, and $i = 0, \ldots, \lceil\log_\alpha\frac{v}{u}\rceil - 1$, for all i let $\beta_i = u\alpha^i$ be selected before the draw of the training sample. Then for every $\beta \in [u, v]$, there is a β_i such that $\beta_i \le \beta \le \alpha\beta_i$.

We can extend (3) by applying a union bound over the β_i's, so that for all P with probability of at least $1-\delta$ over the draw of S, the following holds simultaneously for all β_i:

$$\mathbb{E}_P[M_{\beta_i}(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{\alpha}{n\beta_i}\left(D(P\|Q) + \ln\frac{\lceil \log_\alpha \frac{v}{u}\rceil}{\delta}\right). \tag{17}$$

By Proposition 1(2), for any $w \in \mathcal{W}$, $M_\beta(w)$ is a nonincreasing of β. Thus for any $\beta \in [u,v]$ and β_i such that $\beta_i \leq \beta \leq \alpha\beta_i$, $M_\beta(w) \leq M_{\beta_i}(w)$ and $\frac{1}{\beta_i} \leq \frac{\alpha}{\beta}$. Moreover, since $\ell(w,Z)$ is σ-sub-Gaussian under μ by assumption, we have for all $w \in \mathcal{W}$ and $\beta > 0$, $L_\mu(w) \leq M_\beta(w) + \frac{\beta}{2}\sigma^2$. Hence, with probability of at least $1-\delta$ we have,

$$\mathbb{E}_P[L_\mu(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{\alpha}{n\beta}\left(D(P\|Q) + \ln\frac{\lceil \log_\alpha \frac{v}{u}\rceil}{\delta}\right) + \frac{\beta\sigma^2}{2}. \tag{18}$$

Letting $J = D(P\|Q) + \ln\frac{\log_\alpha \sqrt{n}+K}{\delta}$, we find that the value for β that optimizes the right hand side of the bound in the statement of the proposition is bounded from below by $\sqrt{\frac{2\alpha}{n\sigma^2}}$. Letting $u = \frac{1}{\sqrt{n}}\min\left\{\sqrt{\frac{2\alpha}{\sigma^2}}, v\right\}$ and plugging it in (18) completes the proof. □

B. Proofs for Section III-B

Proposition 12. *For any $[0,1]$-valued loss function ℓ, for any $\beta > 0$ and $\delta \in (0,1]$, with probability of at least $1-\delta$ over a draw of $\tilde{Z}, U$ as defined above, we have:*

$$\mathbb{E}_P[\mathrm{g}(W,\tilde{Z},U)] \leq \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\beta}{2}. \tag{8}$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{W,\tilde{Z},U}[\mathrm{g}(W,\tilde{Z},U)] \leq \sqrt{\frac{2\cdot D(P\|Q|P_{\tilde{Z},U})}{n}}. \tag{9}$$

Proof of Proposition 12. Applying the Donsker-Varadhan Lemma 23 to the function, $f(w) = n\beta\mathrm{g}(w,\tilde{Z},U)$, and following the same steps as in the proof of Lemma 4, we arrive at

$$\mathbb{E}_{\tilde{Z},U}\exp\left\{n\beta\mathbb{E}_P[\mathrm{g}(W,\tilde{Z},U)] - D(P\|Q)\right\} \leq \mathbb{E}_Q\mathbb{E}_{\tilde{Z},U}[e^{n\beta\mathrm{g}(W,\tilde{Z},U)}] = \mathbb{E}_Q\mathbb{E}_{\tilde{Z}}\mathbb{E}_U[e^{n\beta\mathrm{g}(W,\tilde{Z},U)}], \tag{19}$$

where the last equality follows since $\tilde{Z} \perp\!\!\!\perp U$. Since $\ell \in [0,1]$, $\mathrm{g}(W,\tilde{Z},U)$ is $\frac{1}{\sqrt{n}}$-sub-Gaussian. Morever, $\mathbb{E}_U[\mathrm{g}(W,\tilde{Z},U)] = 0$. By Hoeffding's lemma, we have $\mathbb{E}_{\tilde{Z}}\mathbb{E}_U[e^{n\beta\mathrm{g}(W,\tilde{Z},U)}] \leq e^{n\beta^2/2}$, and hence

$$\mathbb{E}_{\tilde{Z},U}\exp\left\{n\beta\mathbb{E}_P[\mathrm{g}(W,\tilde{Z},U)] - D(P\|Q) - \tfrac{n\beta^2}{2}\right\} \leq 1. \tag{20}$$

(8) then follows by an application of Markov's inequality.

Let $R(\tilde{Z},U) = n\beta\mathbb{E}_P[\mathrm{g}(W,\tilde{Z},U)] - D(P\|Q) - \frac{n\beta^2}{2}$. From (20), and using Jensen's inequality, we have $e^{\mathbb{E}_{\tilde{Z},U}[R(\tilde{Z},U)]} \leq \mathbb{E}_{\tilde{Z},U}[e^{R(\tilde{Z},U)}] \leq 1$, which implies

$$\mathbb{E}_{\tilde{Z},U,W}[\mathrm{g}(W,\tilde{Z},U)] \leq \inf_{\beta>0}\left(\frac{D(P_{W|\tilde{Z}U}\|Q_{W|\tilde{Z}}|P_{\tilde{Z}U})}{n\beta} + \frac{\beta}{2}\right) = \sqrt{\frac{2\cdot D(P_{W|\tilde{Z}U}\|Q_{W|\tilde{Z}}|P_{\tilde{Z}U})}{n}},$$

and we have shown (9). □

Under the oracle prior $Q_{W|\tilde{Z}} = P_{W|\tilde{Z}}$, we have $D(P_{W|\tilde{Z}U}\|P_{W|\tilde{Z}}|P_{\tilde{Z}U}) = I(W;U|\tilde{Z})$. By noting that $\mathbb{E}_{\tilde{Z},U,W}[\mathrm{g}(W,\tilde{Z},U)] = \mathbb{E}_{\tilde{Z},U,W}[\mathrm{g}(W,\tilde{Z}_U)]$, we recover [7, Theorem 2(1)]. The proof for the single-draw bound follows that of [4, Lemmas 14, 15].

C. Proofs for Section III-C

Starting from Catoni's bound in Theorem 13, using $1 \leq \beta(1-e^{-\beta})^{-1} \leq (1-\frac{\beta}{2})^{-1}$, we have

$$\begin{aligned}
\mathbb{E}_P[L_\mu(W)] &\leq \Phi_\beta^{-1}\left\{\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left[D(P\|Q) + \ln\frac{1}{\delta}\right]\right. = \frac{1-e^{-\beta\mathbb{E}_P[L_S(W)] - \frac{1}{n}\left(D(P\|Q)+\ln\frac{1}{\delta}\right)}}{1-e^{-\beta}} \\
&\leq \frac{\beta}{1-e^{-\beta}}\left[\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right)\right] \qquad (21) \\
&\leq \frac{1}{1-\frac{\beta}{2}}\left[\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right)\right]. \qquad (22)
\end{aligned}$$

(21) and (22) recover, resp., Catoni's [20, Theorem 1.2.1] and McAllester's "Linear PAC-Bayes bound" [17, Theorem 2], where for the latter we additionally require that $\beta < 2$.

We now show how inequality (3) relates to other well-known PAC-Bayesian inequalities such as the "PAC-Bayes-KL-inequality" [21], [32]. Applying the Donsker-Varadhan lemma to the function $f(w) = n\beta(L_\mu(w) - L_S(w))$, which involves the true risk $L_\mu(w)$ instead of the annealed expectation $M_\beta(w)$ (see 14), and following the same steps as in the proof of (3) in Theorem 5, we arrive at the following PAC-Bayesian bound:

$$\Pr_{S\sim\mu^{\otimes n}}\left(\mathbb{E}_P[L_\mu(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta} + \ln \mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\beta\left(L_\mu(W)-L_{S'}(W)\right)}\Big]\right) \ge 1-\delta. \tag{23}$$

For an explicit comparison of (23) with (3), we write the latter as

$$\Pr_{S\sim\mu^{\otimes n}}\left(\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta} + \underbrace{\ln \mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\beta\left(M_\beta(W)-L_{S'}(W)\right)}}_{=0}\Big]\right) \ge 1-\delta, \tag{24}$$

where the last term in the right hand side of the bound in (24) vanishes since $e^{-n\beta M_\beta(w)} = \mathbb{E}_{S'\sim\mu^{\otimes n}}\left[e^{-n\beta L_{S'}(w)}\right]$ for any $w \in \mathcal{W}$ and $\beta > 0$. In contrast, the term $\ln \mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\beta(L_\mu(W)-L_{S'}(W))}$ involving the true risk in (23) is, in general, positive.

Specializing to the case of a $\{0,1\}$-valued loss, fix $\beta = 1$, and let $\Delta : [0,1]\times[0,1] \to \mathbb{R}$ be a convex function. Applying the Donsker-Varadhan lemma to the function, $f(w) = n\Delta(L_S(w), L_\mu(w))$, following the same steps as in the proof of (3) in Theorem 5, and by noting that $\Delta(\mathbb{E}_P[L_S(W)], \mathbb{E}_P[L_\mu(W)])) \le \mathbb{E}_P[\Delta(L_S(W), L_\mu(W))]$, we arrive at the following PAC-Bayesian bound (see, e.g., [21, Lemma 3], [25, Theorem 2.1], [28, Equation 4]):

$$\Pr_{S\sim\mu^{\otimes n}}\left(\Delta(\mathbb{E}_P[L_S(W)], \mathbb{E}_P[L_\mu(W)])) \le \frac{1}{n}\left(D(P\|Q) + \ln\frac{1}{\delta} + \ln \mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\Delta(L_{S'}(W),L_\mu(W))}\right)\right) \ge 1-\delta. \tag{25}$$

For $x, y \in [0,1]$, the binary KL divergence is $kl(y\|x) = y\ln\frac{y}{x} + (1-y)\ln\frac{1-y}{1-x}$. The PAC-Bayes-KL-inequality comes about by upper-bounding the log-exponential-moment term involving the true risk in the right hand side of the bound in (25): For $\Delta(y,x) = kl(y\|x)$, Maurer [21] showed that for $n \ge 8$, $\mathbb{E}_Q\mathbb{E}_{S'}[e^{n\Delta(L_{S'}(W),L_\mu(W))}] \le 2\sqrt{n}$, when we have

$$\Pr_{S\sim\mu^{\otimes n}}\left(kl(\mathbb{E}_P[L_S(W)], \mathbb{E}_P[L_\mu(W)])) \le \frac{1}{n}\left(D(P\|Q) + \ln\frac{2\sqrt{n}}{\delta}\right)\right) \ge 1-\delta. \tag{26}$$

Letting $\Delta(y,x) = 2(y-x)^2$ leads to the bound in [16], while letting $\Delta(y,x) = (y-x)^2/(2x)$ leads to that in [27].

Under a sub-gamma loss assumption, the bounds in either (23) or (24) lead to (see Corollary 9):

$$\Pr_{S\sim\mu^{\otimes n}}\left(\mathbb{E}_P[L_\mu(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\sigma^2}{2(1-c)}\right) \ge 1-\delta. \tag{27}$$

D. Proofs for Section IV

Proposition 19. *Consider the setting in Theorem 5. Let $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{W})$ be an $(\epsilon,0)$-differentially private algorithm. Then with probability of at least $1-\delta$ over the choice of $S\sim\mu^{\otimes n}$, for all $P \in \mathcal{M}(\mathcal{W})$,*

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{D(P\|Q^0(S)) + f_n(\delta,\epsilon)}{n\beta}.$$

where $f_n(\delta,\epsilon) := \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}}$.

The proof follows closely that of [38, Theorem 4.2].

Proof of Proposition 19. For every $Q \in \mathcal{M}(\mathcal{W})$, let

$$F(Q) = \left\{S' \in \mathcal{Z}^n \colon \exists\, P \in \mathcal{M}(\mathcal{W}),\, \mathbb{E}_P[M_\beta(W)] \ge \mathbb{E}_P[L_{S'}(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta'}\right)\right\}.$$

By Theorem 5, we have $\Pr_{S'\sim\mu^{\otimes n}}\left(S' \in F(Q)\right) \le \delta'$. From (10), we have

$$\Pr_{S\sim\mu^{\otimes n}}\left(S \in F(Q^0(S))\right) \le e^{I^\alpha_{\infty,\mu}(Q^0,n)} \cdot \Pr_{(S,S')\sim\mu^{\otimes 2n}}\left(S' \in F(Q^0(S))\right) + \alpha \le e^{I^\alpha_{\infty,\mu}(Q^0,n)}\cdot\delta' + \alpha. \tag{28}$$

Letting $\delta := e^{I^{\alpha}_{\infty,\mu}(Q^0,n)} \cdot \delta' + \alpha$, for $\alpha \in (0, \delta)$ we have,

$$\Pr_{S\sim\mu^{\otimes n}}\left(\exists\, P \in \mathcal{M}(\mathcal{W}),\, \mathbb{E}_P[M_\beta(W)] \geq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q^0(S)) + \ln\frac{1}{\delta-\alpha} + I^{\alpha}_{\infty,\mu}(Q^0,n)\right)\right) \leq \delta.$$

The proof is complete by replacing $I^{\alpha}_{\infty,\mu}(Q^0,n)$ with the bound in Theorem 17, and choosing $\alpha = \frac{\delta}{2}$. □

E. Proofs for Section V

Lemma 21. *Consider a quadratic approximation of the training loss around a local minimizer w_P, $\tilde{L}_S(w) = \frac{1}{2}(w-w_P)^\top H(w-w_P)$, a fixed prior $Q = \mathcal{N}(w_Q, \lambda^{-1}I_k)$, and a posterior distribution of the form $P = \mathcal{N}(w_P, \Sigma_P)$. Then the solution to the convex optimization problem $\min_{\Sigma_P} \mathbb{E}_P[\tilde{L}_S(W)] + (n\beta)^{-1}D(P\|Q)$, is given by $\Sigma_P^\star = H_\lambda^{-1}$, where $H_\lambda := (n\beta H + \lambda I_k)$. Here we assume $\lambda > 0$ is sufficiently large so that H_λ is positive definite.*

Proof. Letting $\theta = w - w_P$, and $P' = P - w_P$, note that $\theta^\top H\theta = \operatorname{Tr}(\theta^\top H\theta) = \operatorname{Tr}(H\theta\theta^\top)$. Hence

$$\mathbb{E}_{P'}[\tfrac{1}{2}\theta^\top H\theta] = \mathbb{E}_{P'}[\tfrac{1}{2}\operatorname{Tr}(H\theta\theta^\top)] = \tfrac{1}{2}\operatorname{Tr}(H\mathbb{E}_{P'}[\theta\theta^\top]) = \tfrac{1}{2}\operatorname{Tr}(H\Sigma_P).$$

For $Q \sim \mathcal{N}(w_Q, \Sigma_Q)$ and $P \sim \mathcal{N}(w_P, \Sigma_P)$, we have

$$\begin{aligned}\mathbb{E}_{P'}[\tfrac{1}{2}\theta^\top H\theta] + (n\beta)^{-1}D(P\|Q) &= \tfrac{1}{2}\operatorname{Tr}(H\Sigma_P) + (n\beta)^{-1}D(P\|Q)\\ &= \frac{\operatorname{Tr}(H\Sigma_P)}{2} + \frac{(n\beta)^{-1}}{2}\left(\ln\frac{\det\Sigma_Q}{\det\Sigma_P} + \operatorname{Tr}\left(\Sigma_Q^{-1}\Sigma_P\right) - k + (w_Q-w_P)^\top\Sigma_Q^{-1}(w_Q-w_P)\right).\end{aligned}$$

The derivative of the RHS w.r.t. Σ_P is $\frac{1}{2}\left[H - (n\beta)^{-1}\Sigma_P^{-1} + (n\beta)^{-1}\Sigma_Q^{-1}\right]^\top$, where we have used the fact that $\nabla_A \operatorname{Tr}(AB) = B^\top$, and $\nabla_A \ln\det(A) = (A^{-1})^\top$. Setting the derivative to zero and $\Sigma_Q = \lambda^{-1}I_k$ yields the result. □

Proposition 22. *Let $\{\lambda_i\}_{i=1}^k$ be the eigenvalues of H_λ and suppose that $\lambda_i \geq \lambda > 0$ for all i. Let $Q = \mathcal{N}(w_Q, \lambda^{-1}I_k)$ be a prior, and let $P = \mathcal{N}(w_P, H_\lambda^{-1})$. Then with probability of at least $1-\delta$ over a draw of the sample S, we have*

$$\begin{aligned}\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] &+ \frac{1}{n\beta}\ln\frac{1}{\delta}\\ &+ \frac{1}{n\beta}\left(\frac{\lambda}{2}\|w_Q - w_P\|^2 + \frac{1}{2}\sum_{i=1}^k \ln\frac{\lambda_i}{\lambda}\right).\end{aligned} \tag{13}$$

Proof of Proposition 22. The proof follows from Theorem 5, and the fact that for $Q = \mathcal{N}(w_Q, \lambda^{-1}I_k)$, $P = \mathcal{N}(w_P, H_\lambda^{-1})$ such that $\lambda_i \geq \lambda > 0$ for all i, we have

$$D(P\|Q) = \frac{1}{2}\left(\lambda\|w_Q - w_P\|^2 + \sum_{i=1}^k \ln\frac{\lambda_i}{\lambda} + \sum_{i=1}^k\left(\frac{\lambda}{\lambda_i} - 1\right)\right) \leq \frac{1}{2}\left(\lambda\|w_Q - w_P\|^2 + \sum_{i=1}^k \ln\frac{\lambda_i}{\lambda}\right).$$

□