Journal of Machine Learning Research 21 (2020) 1-32 Submitted 9/19; Revised 6/20; Published 6/20

Chaining Meets Chain Rule: Multilevel Entropic Regularization and Training of Neural Networks

Amir R. Asadi AASADI@PRINCETON.EDU
Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08544, USA

Emmanuel Abbe EMMANUEL.ABBE@EPFL.CH
Mathematics Institute and School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

Editor: Gabor Lugosi

Abstract

We derive generalization and excess risk bounds for neural networks using a family of complexity measures based on a multilevel relative entropy. The bounds are obtained by introducing the notion of generated hierarchical coverings of neural networks and by using the technique of chaining mutual information introduced by Asadi et al. '18. The resulting bounds are algorithm-dependent and multiscale: they exploit the multilevel structure of neural networks. This, in turn, leads to an empirical risk minimization problem with a multilevel entropic regularization. The minimization problem is resolved by introducing a multiscale extension of the celebrated Gibbs posterior distribution, proving that the derived distribution achieves the unique minimum. This leads to a new training procedure for neural networks with performance guarantees, which exploits the chain rule of relative entropy rather than the chain rule of derivatives (as in backpropagation), and which takes into account the interactions between different scales of the hypothesis sets of neural networks corresponding to different depths of the hidden layers. To obtain an efficient implementation of the latter, we further develop a multilevel Metropolis algorithm simulating the multiscale Gibbs distribution, with an experiment for a two-layer neural network on the MNIST data set.

Keywords: neural networks, multilevel relative entropy, chaining mutual information, multiscale generalization bound, multiscale Gibbs distribution

1. Introduction

Deep neural networks have found profound applications in many areas of artificial intelligence, yet they are lacking solid theoretical grounds. Constructing a theory for understanding neural networks and for how to design their architecture and better train them is of vital interest and a main challenge in machine learning. Nowadays, deep neural networks are dominantly trained by stochastic gradient descent (SGD) or its variants. In this paper, based on ideas from high dimensional probability and information theory, we present a new perspective on designing the architecture of neural nets and propose a novel algorithm which is fundamentally different from SGD and its variants.

Chaining, originated from Kolmogorov in 1934, is a powerful multiscale method in high dimensional probability for bounding the suprema of random processes. Furthermore, one may realize at a high-level that the multiscale argument in chaining has the potential to apply naturally to the multilevel architecture of deep neural networks. We thus raise the following question:

Can we use the intrinsic power of chaining to devise learning algorithms for deep neural networks with performance guarantees?

Motivated by this question, we show that the multilevel architecture of neural networks makes them ideal for devising a training procedure based on the recent information-theoretic and algorithm-dependent extension of chaining, that is, the chaining mutual information (CMI) technique introduced by Asadi et al. (2018). The CMI is a combination of classical chaining with the mutual information bound of Russo and Zou (2016) and Xu and Raginsky (2017). We give a brief overview of this technique in Section 2.

A key tool used in both classical chaining and CMI is the notion of hierarchical coverings of index sets and hypothesis sets, with controlled diameters. We strengthen and extend the CMI technique of Asadi et al. (2018) and adapt it to the architecture of deep neural nets. Then, using our strengthened CMI technique, we obtain new *multiscale* and *algorithm-dependent* generalization bounds for neural nets. We show that these chaining-style and information-theoretic generalization bounds are capable of creating a new method of training neural nets, modeled as a conditional probability distribution. This multilevel algorithm is intrinsically different from algorithms which treat the whole net as a single block, such as the widely used SGD and its variants. It is also distinct from layer-wise training algorithms in which the layers under training are oblivious to the untrained layers at each stage. In fact, this multilevel algorithm differentiates between the different scales of the hypothesis set of deep neural nets, corresponding to different depth of its hidden layers, and takes into account the interactions between them. Crucially, using the generalization bound, we demonstrate that the excess risk of our proposed training algorithm satisfies a chaining-style multiscale bound.

These generalization and excess risk bounds introduce a family of complexity measures for the hypotheses of neural nets, based on a *multilevel relative entropy*; see Definition 8. These complexity measures take into account the multilevel and compositional structure of neural nets, as opposed to the classical relative entropy (KL-divergence) derived from the PAC-Bayesian bounds (see e.g. Catoni, 2007), or mutual information bounds (Russo and Zou, 2016; Xu and Raginsky, 2017).

More precisely, in the main results of this paper, we first demonstrate an advantage of the multilevel architecture of deep neural nets by showing how one can obtain accessible hierarchical coverings for their hypothesis sets, introducing the notion of *generated coverings* in Section 3. Since these hierarchical coverings are naturally generated from the architecture of deep neural nets, they are easily accessible and convenient for use in the training algorithm. Furthermore, we show how one can regularize the hypothesis set of neural nets to make these hierarchical sequence of generated coverings possess controlled diameters suitable for the chaining argument; see Section 4 on *multilevel regularization*. The effect of such regularization on the representation ability of neural nets has been recently studied, such as by Hardt and Ma (2016) and Bartlett et al. (2018) for the special case where layers

are restricted to nearly-identity functions similar to residual networks (He et al., 2016). Then, we derive our generalization bound for arbitrarily deep feedforward neural nets via applying our strengthened CMI technique and using their hierarchical sequence of generated coverings. Although such a sequence of coverings may not be the sequence which gives the tightest possible generalization bound, it has the major advantage of being easily accessible, and hence can be exploited in devising multilevel training algorithms. Designing training algorithms based on hierarchical coverings of hypothesis sets which achieve chaining-style excess risk (or regret) bounds has first been studied by Cesa-Bianchi and Lugosi (1999), and has recently regained traction in, for example, works by Gaillard and Gerchinovitz (2015) and Cesa-Bianchi et al. (2017), all in the context of online learning and prediction of individual sequences. With such approaches, hierarchical coverings are no longer viewed merely as methods of proof for generalization bounds: they further allow for algorithms achieving low statistical error. However, a major difficulty of using the algorithms given in the aforementioned prior works is in constructing suitable hierarchical coverings. In this paper, we show how this task is easy for the multilevel architecture of neural nets. Moreover, to the best of our knowledge, we are the first to devise chaining-based multilevel algorithms for the batch learning setting.

In our case, the derived generalization bound puts forward a multilevel relative entropy as a regularization term. We then turn to minimizing the empirical error with this induced regularization, called here the *multilevel entropic regularization*. Interestingly, we can solve this minimization problem exactly, obtaining a multiscale extension of the celebrated Gibbs algorithm (posterior distribution); see Sections 5 and 6. This target conditional distribution is obtained in a backwards manner by successive marginalization and tilting of the classical Gibbs distribution, as described in the *marginalize-tilt* algorithm introduced in Section 6. Unlike the classical Gibbs distribution which has a global temperature parameter, its multiscale counter-part possesses a *temperature vector*. We then present a multilevel training algorithm by simulating our target distribution via a multilevel Metropolis algorithm introduced for a two layer net in Section 8. In contrast to the celebrated backpropagation algorithm which exploits the chain rule of derivatives, our target distribution and its simulated version are derived from the chain rule of relative entropy, and take into account the interactions between different scales of the hypothesis sets of neural nets corresponding to different depths of the hidden layers.

This paper introduces the new concepts and main results behind this alternative approach to training neural nets. Many directions emerge from this approach, in particular for its applicability. It is worth noting that Markov chain Monte Carlo (MCMC) methods are known to often better cope with non-convexity issues than gradient descent, since they are able to backtrack from local minima (Geman and Geman, 1984). Furthermore, in contrast to gradient descent, MCMC methods take into account parameter uncertainty that helps preventing overfitting (Welling and Teh, 2011). However, compared to gradient based methods, these methods are typically computationally more demanding.

Notice that in this work, instead of endeavoring to theoretically explain the performance of currently used algorithms in practice, we take the reverse course of proposing an entirely new and different method for training neural networks and designing their architectures, *driven* by new rigorous theory. It is widely believed that neural nets learn features from data in a hierarchical manner; see e.g. LeCun et al. (2015). Supportive to this belief, the

new training algorithm proposed in this paper has a hierarchical and multilevel structure, in contrast to SGD which trains all the layers together on each pass.

1.1. Further Related Literature

Information-theoretic approaches to statistical learning have been studied in PAC-Bayesian theory; see works by McAllester (1999), Catoni (2007), Guedj (2019), Audibert and Bousquet (2004) and references therein, and with the recent mutual information generalization bound such as by Russo and Zou (2016), Xu and Raginsky (2017), Raginsky et al. (2016), Jiao et al. (2017), Pensia et al. (2018), Bassily et al. (2018) and Bu et al. (2020). PAC-Bayes generalization bounds have been specifically derived for neural networks in recent works such as by Dziugaite and Roy (2017a, 2018), Neyshabur et al. (2017) and Zhou et al. (2018). The statistical properties of the Gibbs distribution, also known as the Boltzmann distribution, or the exponential weights distribution (Rigollet and Tsybakov, 2012), have been studied in the information-theoretic works by Zhang (1999, 2006a,b), Xu and Raginsky (2017) and Raginsky et al. (2016). The Gibbs distribution has been applied in devising and analyzing training algorithms in recent studies such as by Chaudhari et al. (2016), Raginsky et al. (2017) and Dziugaite and Roy (2017b). Tilted distributions in unsupervised and semi-supervised statistical learning problems has also been studied by Asadi et al. (2017) in the context of community detection. A notion of multiscale entropy, related to our multilevel relative entropy, has been used by Bubeck et al. (2018) in the context of online algorithms and the k-server problem.

1.2. Notation

In this paper, all logarithms are in natural base and all information-theoretic measures are in nats. Let $\imath_{P\|Q}$, $D(P\|Q)$ and $D_\lambda(P\|Q)$ denote the relative information, the relative entropy, and the Rényi divergence of order λ between probability measures P and Q, and let $D(P_{Y|X}\|Q_{Y|X}|P_X) \triangleq \int D(P_{Y|X=\omega}\|Q_{Y|X=\omega})\mathrm{d}P_X(\omega)$ denote conditional relative entropy (see Appendix A for precise definitions). In the framework of supervised statistical batch learning, $\mathcal{X}$ denotes the instances domain, $\mathcal{Y}$ is the labels domain, $\mathsf{Z} = \mathcal{X} \times \mathcal{Y}$ denotes the examples domain and $\mathcal{H} = \{h_w : w \in \mathcal{W}\}$ is the hypothesis set, where the hypotheses are indexed by an index set $\mathcal{W}$. Let $\ell : \mathcal{W} \times \mathsf{Z} \to \mathbb{R}^+$ be the loss function. A learning algorithm receives the training set $S = (Z_1, Z_2, ..., Z_n)$ of n examples with i.i.d. random elements drawn from Z with an unknown distribution μ, thus $P_S = \mu^{\otimes n}$. Then it picks an element $h_W \in \mathcal{H}$ as the output hypothesis according to a random transformation $P_{W|S}$. For any $w \in \mathcal{W}$, let $L_\mu(w) \triangleq \mathbb{E}[\ell(w, Z)]$ denote the statistical (or population) risk of hypothesis h_w, where $Z \sim \mu$. For a given training set S, the empirical risk of hypothesis h_w is defined as $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$, and the generalization error of hypothesis h_w (dependent on the training set) is defined as $\mathrm{gen}(w) \triangleq L_\mu(w) - L_S(w)$. Averaging with respect to the joint distribution $P_{S,W} = \mu^{\otimes n} P_{W|S}$, we denote the expected generalization error by $\mathrm{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}[\mathrm{gen}(W)]$, and the expected statistical risk by $\mathrm{risk}(\mu, P_{W|S}) \triangleq \mathbb{E}[L_\mu(W)]$. Throughout the paper, $\|A\|_2$ denotes the spectral norm of matrix A and $|b|_2$ denotes the Euclidean norm of vector b. Let δ_w denote the Dirac measure centered at w.

2. Preliminary: The CMI Technique

Chaining, originated from Kolmogorov and developed by Dudley, Talagrand, Fernique and others, is a powerful technique in high dimensional probability for bounding the expected suprema of random processes while taking into account the dependencies between their random variables in a multiscale manner using maximal inequalities. Here we emphasize the core idea of the chaining technique: performing refined approximations of the random variables of a process by using a telescoping sum, named as *chaining sum*. If T is an arbitrary index set and $\{X_t\}_{t\in T}$ is a random process, then for any $t \in T$ one can write

$$X_t = X_{\pi_1(t)} + \left(X_{\pi_2(t)} - X_{\pi_1(t)}\right) + \cdots + \left(X_{\pi_d(t)} - X_{\pi_{d-1}(t)}\right) + \left(X_t - X_{\pi_d(t)}\right),$$

where $\pi_1(t), \pi_2(t), \ldots, \pi_d(t)$ are finer and finer approximations of the index t. Each of the differences $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$, $k = 1, 2, \ldots, d$, is called a *link* of the chaining sum. Informally speaking, if the approximations $\pi_k(t)$, $k = 1, 2, \ldots, d$, are close enough to each other and $\pi_d(t)$ is close to t, then, in many important applications, controlling the expected supremum of each of the links with union bounds and summing them up will give a much tighter bound than bounding the supremum of X_t upfront with a union bound.[1] For instance, the approximations may be the projections of t on an increasing sequence of partitions of T, which are partitions of T at different *scales*. For more information, see van Handel (2016), Vershynin (2018), Talagrand (2014) and references therein.

The technique of chaining mutual information, recently introduced by Asadi et al. (2018), can be interpreted as an algorithm-dependent version of the above, extending a result of Fernique (1976) by further taking into account such dependencies, and adjusting chaining for statistical learning problems. Assume that, for the given random process $\{X_t\}_{t\in T}$, the goal is to obtain an upper bound on the expected bias $\mathbb{E}[X_W]$, where W is the output of an algorithm which takes values on the index set T. In brief, Asadi et al. (2018) assert that one can replace the metric entropy in chaining with the mutual information between the input $\{X_t\}_{t\in T}$ and the discretized output $\pi_k(W)$. By writing the chaining sum with random index W and after taking expectations, we obtain:

$$\mathbb{E}\left[X_W\right] = \mathbb{E}\left[X_{\pi_1(W)}\right] + \mathbb{E}\left[X_{\pi_2(W)} - X_{\pi_1(W)}\right] + \cdots + \mathbb{E}\left[X_W - X_{\pi_d(W)}\right]. \tag{1}$$

With this technique, rather than bounding $\mathbb{E}\left[X_W\right]$ with a single mutual information term (Russo and Zou, 2016; Xu and Raginsky, 2017), one bounds each link $\mathbb{E}\left[X_{\pi_k(W)} - X_{\pi_{k-1}(W)}\right]$, $k = 1, 2, \ldots, d$, and then sums them up. This gives a multiscale and algorithm-dependent upper bound on $\mathbb{E}\left[X_W\right]$.

Remark 1 The notion of metric entropy is similar to Hartley entropy in the information theory literature. To deal with the effect of noise in communication systems, Hartley entropy was generalized and replaced by mutual information by Shannon (see Verdú, 1998).

In this paper, first we note that unlike the classical chaining method in which we require finite size partitions whose cardinalities appear in the bounds,[2] that requirement is unnecessary for the CMI technique. Therefore one may use a hierarchical sequence of

1. The idea is that the increments may capture more efficiently the dependencies.
2. Finite partitions is not required in the theory of majorizing measures (generic chaining).

coverings of the index set which includes covers of possibly uncountably infinite size. This fact will be useful for analyzing neural networks with continuous weight values in the next sections. For details, see Appendix B.

The second important contribution is to design the coverings to meet the multilayer structure of neural nets. In the classical chaining and the CMI of Asadi et al. (2018), these are applied on an arbitrary infinite sequence of 2^{-k}-partitions. In this paper, we take a different and new approach and use the hierarchical sequences of generated coverings associated with multilevel architectures, as defined in the next section.

Remark 2 Using Theorem 2 of Bu et al. (2020), we also show that for empirical processes, one can replace the mutual information between the whole input set and the discretized output with mutual informations between individual examples and the discretized output to obtain a tighter CMI bound. For details, see Appendix B.

3. Multilevel Architectures and Their Generated Coverings

Assume that in a statistical learning problem, the hypothesis set $\mathcal{H} = \{h_w : w \in \mathcal{W}\}$ consists of multilevel functions, that is, the index set $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ can be written as a Cartesian product and consists of elements $w \in \mathcal{W}$ representable with $d \geq 2$ components as $w = (\mathbf{W}_1, \ldots, \mathbf{W}_d)$. Examples for neural nets can be: 1. When the components are the layers. 2. When the components are stacks of layers plus skip connections, such as in residual networks. For all $1 \leq k \leq d$, let $\mathcal{G}_k$ be the exact covering of $\mathcal{W}$ determined by all possible values of the first k components, that is, any two indices are in the same set if and only their first k components match:

$$\mathcal{G}_k \triangleq \{\{\mathbf{W}_1\} \times \cdots \times \{\mathbf{W}_k\} \times \mathcal{W}_{k+1} \times \cdots \times \mathcal{W}_d : (\mathbf{W}_1, \ldots, \mathbf{W}_k) \in \mathcal{W}_1 \times \cdots \times \mathcal{W}_k\}\,.$$

Notice that $\{\mathcal{G}_k\}_{k=1}^d$ is a hierarchical sequence of exact coverings of the index set $\mathcal{W}$, and the projection set of any $w \in \mathcal{W}$ in $\mathcal{G}_k$, that is, the unique set in $\mathcal{G}_k$ which includes w, is determined only by the values of the first k components of w. We call $\{\mathcal{G}_k\}_{k=1}^d$ the hierarchical sequence of *generated coverings* of the index set $\mathcal{W}$, and will use CMI with this sequence in the next sections.[3]

Remark 3 The notion of generated coverings of $\mathcal{W}$ is akin in nature to the notion of *generated filtrations* of random processes in probability theory (for a definition, see Çınlar 2011, p. 171) and applying the CMI technique on this sequence is akin to the *martingale method* for concentration bounds.

We provide the following simple yet useful example by revisiting Example 1 of Asadi et al. (2018):

Example 1 Consider a canonical Gaussian process $X_t \triangleq \langle t, G^n \rangle, t \in T$ where $n = 2$, $G^2 = (G_1, G_2)$ has independent standard normal components and $T \triangleq \{t \in \mathbb{R}^2 : |t|_2 = 1\}$. The process $\{X_t\}_{t \in T}$ can also be expressed according to the phase of each point $t \in T$, i.e. the unique number $\phi \in [0, 2\pi)$ such that $t = (\sin\phi, \cos\phi)$. Assume that the indices are in

3. Notice that for a given architecture, one can re-parameterize the components with different permutations of $\{1, 2, \ldots, d\}$ to give different generated coverings.

the phase form and define the following dyadic sequence of partitions of T: For all integers $k \geq 1$,

$$\mathcal{P}_k \triangleq \left\{ \left[0, \frac{2\pi}{2^k}\right), \left[\frac{2\pi}{2^k}, 2 \times \frac{2\pi}{2^k}\right), ..., \left[\left(2^k - 1\right)\frac{2\pi}{2^k}, 2\pi\right) \right\}.$$

Can T and the sequence $\{\mathcal{P}_k\}_{k=1}^{\infty}$ be related to the hypothesis set of a multilevel architecture and its generated coverings? For all integers $i \geq 1$, let $\mathcal{W}_i \triangleq \left\{ \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \middle| \theta \in \left\{-\frac{\pi}{2^i}, \frac{\pi}{2^i}\right\} \right\}$. Notice that for each $t = [t_1, t_2] \in T$, one can write

$$\begin{aligned} X_t &= \begin{bmatrix} t_1 & t_2 \end{bmatrix} G^2 \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix} (\cdots W_2 W_1)\, G^2, \end{aligned}$$

where each $W_i \in \mathcal{W}_i$ is uniquely determined by t. For all $k \geq 1$, fixing the values of $W_1, \ldots, W_k$ and allowing the rest of the matrices to take arbitrary values in their corresponding $\mathcal{W}_i$ gives one of the elements of $\mathcal{P}_k$. Therefore, the sequence of generated coverings associated with the index set of the infinite-depth linear neural net

$$f_W(G^2) = \begin{bmatrix} 1 & 0 \end{bmatrix} (\cdots W_2 W_1)\, G^2$$

is $\{\mathcal{P}_k\}_{k=1}^{\infty}$.

4. Multilevel Regularization

The purpose of multilevel regularization is to control the diameters of the generated coverings[4] and the links of its corresponding chaining sum. Consider a d layer feedforward neural net with parameters $w \triangleq (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_d) \in \mathcal{W}$, where for all $1 \leq k \leq d$, $\mathbf{W}_k \in \mathbb{R}^{\tau_k \times \tau_{k-1}}$ is a matrix between hidden layers $k-1$ and k. Let ϕ denote any non-linearity which is 1-Lipschitz[5] and satisfies $\phi(0) = 0$, such as the entry-wise ReLU activation function, and let ϕ_o either be the soft-max function, or the identity function. For a given $R > 0$, assume that the instances domain is $\mathcal{X} \triangleq \{\mathbf{x} \in \mathbb{R}^m : |\mathbf{x}|_2 \leq R\}$. The feedforward neural net with parameters w is a function $h_w : \mathcal{X} \to \mathbb{R}^{\tau_d}$ defined as $h_w(\mathbf{x}) \triangleq \phi_o(\mathbf{W}_d(\phi(\cdots \phi(\mathbf{W}_1(\mathbf{x})) \cdots)))$. For all $1 \leq k \leq d$, let $\mathbf{M}_k \in \mathbb{R}^{\tau_k \times \tau_{k-1}}$ be a fixed matrix such that $\|\mathbf{M}_k\|_2 > 0$, and for $\alpha_k > 0$, define the following set of matrices:

$$\mathcal{S}_k \triangleq \{\mathbf{W} \in \mathbb{R}^{\tau_k \times \tau_{k-1}} : \|\mathbf{W} - \mathbf{M}_k\|_2 \leq \alpha_k \|\mathbf{M}_k\|_2\}. \tag{2}$$

We assume that the domain of $\mathbf{W}_k$, that is $\mathcal{W}_k$, is a subset of $\mathcal{S}_k$. We are regularizing $\mathbf{W}_k$ with $\mathbf{M}_k$ and α_k, for all $1 \leq k \leq d$, to constrain the links of the chaining sum, as we will see in Lemma 4. We name $\mathbf{M}_k$ and α_k as the *reference*[6] and *radius* of $\mathcal{W}_k$, respectively. A common example used in practice is to let the references be identity matrices, such as for residual nets; e.g. by Hardt and Ma (2016) and Bartlett et al. (2018, 2017). For instance, for

4. The diameter of a covering for a metric space is defined as the supremum of the diameters of its blocks.
5. One can readily replace the ReLU activation function with any other ρ-Lipschitz activation function which maps the origin to origin. Our bounds in the next section will then depend on ρ.
6. This is similar to the terminology of "reference matrices" by Bartlett et al. (2017).

the linear neural net in Example 1, one can take $\mathbf{M}_k = I_{2\times 2}$ and $\alpha_k = \pi 2^{-k}$, for all $k \geq 1$. We define the projection of w on the generated covering $\mathcal{G}_k$ as

$$(\mathbf{W}_1, \ldots, \mathbf{W}_k, \mathbf{M}_{k+1}, \ldots, \mathbf{M}_d).$$

Let $M \triangleq \prod_{j=1}^{d} \|\mathbf{M}_j\|_2$.

Lemma 4 *Let $2 \leq k \leq d$. Let $w_1 = (\mathbf{W}_1, \ldots, \mathbf{W}_{k-1}, \mathbf{W}_k, \mathbf{M}_{k+1}, \ldots, \mathbf{M}_d)$ and $w_2 = (\mathbf{W}_1, \ldots, \mathbf{W}_{k-1}, \mathbf{M}_k, \mathbf{M}_{k+1}, \ldots, \mathbf{M}_d)$ be the projections of w on $\mathcal{G}_k$ and $\mathcal{G}_{k-1}$, respectively. Then, for all $\mathbf{x} \in \mathcal{X}$,*

$$|h_{w_1}(\mathbf{x}) - h_{w_2}(\mathbf{x})|_2 \leq \alpha_k \exp\left(\sum_{i=1}^{k-1} \alpha_i\right) M|\mathbf{x}|_2.$$

For a proof, see Appendix C.

Notice that for any $w \in \mathcal{W}$ and any $\mathbf{x} \in \mathcal{X}$, if ϕ_o is the soft-max function, then $|h_w(\mathbf{x})|_2 \leq 1$, and if ϕ_o is the identity function, then from (2) and the triangle inequality, we derive $|h_w(\mathbf{x})|_2 \leq \exp\left(\sum_{i=1}^{d} \alpha_i\right) MR$. Let the loss function ℓ be chosen such that there exists[7] $L > 0$ for which for any $w_1, w_2 \in \mathcal{W}$ and any $z = (\mathbf{x}, y) \in \mathsf{Z}$ we have $|\ell(w_1, z) - \ell(w_2, z)| \leq L|h_{w_1}(\mathbf{x}) - h_{w_2}(\mathbf{x})|_2$. A commonly used example is the squared ℓ_2 loss, that is, for the net with parameters w and for any example $z = (\mathbf{x}, y) \in \mathsf{Z}$, define $\ell(w, z) \triangleq |h_w(\mathbf{x}) - y|_2^2$. For classification problems, assume that the labels y are one-hot vectors, otherwise, let $|y|_2 \leq 1$. Note that for this loss function, based on triangle inequality, if ϕ_o is the soft-max function, then one can assume $L = 4$, and if ϕ_o is the identity function, then one can take $L = 2 + 2\exp\left(\sum_{i=1}^{d} \alpha_i\right) MR$.

Notice that if $\alpha_i = \frac{1}{d}$ and $\mathbf{M}_i = I_{\tau\times\tau}$ for all $1 \leq i \leq d$ and $R = O(1)$ as assumed by Hardt and Ma (2016), then $L = O(1)$. If $\alpha_i = \frac{\log d}{d}$ and $\mathbf{M}_i = I_{\tau\times\tau}$ for all $1 \leq i \leq d$ and $R = O(1)$ as assumed by Bartlett et al. (2018), then $L = O(d)$. The choice of the radii α_k, $k = 1, \ldots, d$, depends on the representation ability that we require from the neural net.

5. Generalization and Excess Risk Bounds

For simplicity in the notation, we first give the following definition:

Definition 5 *For all $1 \leq k \leq d$, let*

$$\begin{aligned} h_{[\mathbf{W}_1,\ldots,\mathbf{W}_k]} &\triangleq h_{[\mathbf{W}_1,\ldots,\mathbf{W}_k,\mathbf{M}_{k+1},\ldots,\mathbf{M}_d]}, \\ \ell([\mathbf{W}_1, \ldots, \mathbf{W}_k], z) &\triangleq \ell([\mathbf{W}_1, \ldots, \mathbf{W}_k, \mathbf{M}_{k+1}, \ldots, \mathbf{M}_d], z), \end{aligned}$$

and

$$\mathrm{gen}(\mathbf{W}_1, \ldots, \mathbf{W}_k) \triangleq \mathrm{gen}([\mathbf{W}_1, \ldots, \mathbf{W}_k, \mathbf{M}_{k+1}, \ldots, \mathbf{M}_d]).$$

For all $1 \leq k \leq d$, let $\beta_k \triangleq \alpha_k \exp\left(\sum_{i=1}^{k-1} \alpha_i\right)$ and assume that W_k denotes a *random* matrix. We can now state the following multiscale and algorithm-dependent generalization bound derived from CMI using the sequence of generated coverings, in which mutual informations between the training set S and the first k layers appear:

7. This assumption is similar to the assumption of Lemma 17.6 of Anthony and Bartlett (2009).

Theorem 6 *Given the assumptions in the previous section, we have*

$$\operatorname{gen}(\mu, P_{W|S}) \leq \frac{LMR\sqrt{2}}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \sqrt{I(S; W_1, \ldots, W_k)}. \tag{3}$$

Proof According to (1), one can write the chaining sum with respect to the sequence of generated coverings as

$$\begin{aligned}\operatorname{gen}(\mu, P_{W|S}) = \mathbb{E}[\operatorname{gen}(W)] = \mathbb{E}[\operatorname{gen}(W_1)] + \mathbb{E}[\operatorname{gen}(W_1, W_2) - \operatorname{gen}(W_1)] + \ldots \\ + \mathbb{E}[\operatorname{gen}(W) - \operatorname{gen}(W_1, \ldots, W_{d-1})].\end{aligned} \tag{4}$$

Based on the Azuma–Hoeffding inequality, $\{\operatorname{gen}(w)\}_{w\in\mathcal{W}}$ is a subgaussian process with the metric

$$d(w, w') \triangleq \frac{\|\ell(w, \cdot) - \ell(w', \cdot)\|_\infty}{\sqrt{n}},$$

regardless of the choice of distribution μ on Z. For any example $z = (\mathbf{x}, y) \in \mathsf{Z}$, we have

$$|\ell(w, z) - \ell(w', z)| \leq L \left|h_w(\mathbf{x}) - h_{w'}(\mathbf{x})\right|_2.$$

Therefore

$$\|\ell(w, \cdot) - \ell(w', \cdot)\|_\infty \leq L \sup_{\mathbf{x}\in\mathcal{X}} |h_w(\mathbf{x}) - h_{w'}(\mathbf{x})|_2. \tag{5}$$

Based on Lemma 4, for all $1 \leq k \leq d$, we have

$$\sup_{\mathbf{x}\in\mathcal{X}} |h_{[\mathbf{W}_1,\ldots,\mathbf{W}_k]}(\mathbf{x}) - h_{[\mathbf{W}_1,\ldots,\mathbf{W}_{k-1}]}(\mathbf{x})|_2 \leq \beta_k MR. \tag{6}$$

Using (5), we deduce

$$\|\ell([\mathbf{W}_1, \ldots, \mathbf{W}_k], \cdot) - \ell([\mathbf{W}_1, \ldots, \mathbf{W}_{k-1}], \cdot)\|_\infty \leq L\beta_k MR. \tag{7}$$

Notice that knowing the value of $(W_1, \ldots, W_k)$ is enough to determine which one of the random variables $\{\operatorname{gen}(\mathbf{W}_1, \ldots, \mathbf{W}_k) - \operatorname{gen}(\mathbf{W}_1, \ldots, \mathbf{W}_{k-1})\}_{w\in\mathcal{W}}$ is chosen according to W. Therefore $(W_1, \ldots, W_k)$ is playing the role of the random index, and since

$$\operatorname{gen}(\mathbf{W}_1, \ldots, \mathbf{W}_k) - \operatorname{gen}(\mathbf{W}_1, \ldots, \mathbf{W}_{k-1})$$

is $d^2\left([\mathbf{W}_1, \ldots, \mathbf{W}_k], [\mathbf{W}_1, \ldots, \mathbf{W}_{k-1}]\right)$-subgaussian, based on (7), Theorem 2 of Xu and Raginsky (2017) and an application of the data processing inequality on the Markov chain $\{\operatorname{gen}(w)\}_{w\in\mathcal{W}} \leftrightarrow S \leftrightarrow W \leftrightarrow (W_1, \ldots, W_k)$, we obtain

$$\mathbb{E}[\operatorname{gen}(W_1, \ldots, W_k) - \operatorname{gen}(W_1, \ldots, W_{k-1})] \leq \frac{LMR\sqrt{2}\beta_k}{\sqrt{n}} \sqrt{I(S; W_1, \ldots, W_k)}. \tag{8}$$

From (4) and (8) we deduce

$$\operatorname{gen}(\mu, P_{W|S}) = \mathbb{E}[\operatorname{gen}(W)] \leq \frac{LMR\sqrt{2}}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \sqrt{I(S; W_1, \ldots, W_k)}.$$

■

Notice that we can rewrite (3) as

$$\text{risk}\left(\mu, P_{W|S}\right) = \mathbb{E}[L_\mu(W)] \leq \mathbb{E}[L_S(W)] + \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \sqrt{I(S; W_1, \ldots, W_k)}, \tag{9}$$

where $C \triangleq LMR\sqrt{2}$. The goal in statistical learning is to find an algorithm $P_{W|S}$ which minimizes $\text{risk}\left(\mu, P_{W|S}\right) = \mathbb{E}[L_\mu(W)]$. To that end, we derive an upper bound on $\mathbb{E}[L_\mu(W)]$ from inequality (9) whose minimization over $P_{W|S}$ is algorithmically feasible. If for each $k = 1, 2, \ldots, d$, we define $Q^{(k)}_{W_1 \ldots W_k}$ to be a fixed distribution on $\mathcal{W}_1 \times \cdots \times \mathcal{W}_k$ that does not depend on the training set S, which we name as *prior distribution*,[8] then from (9) we deduce

$$\text{risk}\left(\mu, P_{W|S}\right) \leq \mathbb{E}[L_S(W)] + \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k I(S; W_1, \ldots, W_k) + \frac{1}{4\gamma_k} \right) \tag{10}$$

$$\leq \mathbb{E}[L_S(W)] + \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(P_{W_1 \ldots W_k|S} \middle\| Q^{(k)}_{W_1 \ldots W_k} \middle| P_S \right) + \frac{1}{4\gamma_k} \right), \tag{11}$$

where (10) follows from the inequality $\sqrt{x} \leq cx + \frac{1}{4c}$ for all $x, c > 0$, which is upper bounding the concave function $\sqrt{x}$ with a tangent line, and (11) follows from the difference decomposition of mutual information: $I(X;Y) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y)$; see Lemma 19 in Appendix A.[9] Given fixed parameters γ_k, $k = 1, 2, \ldots, d$, and for any fixed n, let $P^\star_{W|S}$ be the conditional distribution that minimizes the right side of (11), that is,

$$P^\star_{W|S} \triangleq \arg\min_{P_{W|S}} \left\{ \mathbb{E}[L_S(W)] + \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \gamma_k D\left(P_{W_1 \ldots W_k|S} \middle\| Q^{(k)}_{W_1 \ldots W_k} \middle| P_S \right) \right\}. \tag{12}$$

Note that we made the expression in (12) linear in P_S. This is crucial, since, in turn, it implies that the algorithm $P^\star_{W|S}$ does not depend on the unknown input distribution μ (recall that $P_S = \mu^{\otimes n}$), which is a desired property of $P^\star_{W|S}$. The excess risk of $P^\star_{W|S}$ satisfies the following multiscale bound:

Theorem 7 *Let $\widehat{w}(\mu)$ denote the index of a hypothesis which achieves the minimum statistical risk among $\mathcal{W}$. For $\epsilon \geq 0$, let $B^{(\epsilon)}_{W_1 \ldots W_d}$ denote the uniform distribution over a neighborhood U_ϵ of $\widehat{w}(\mu)$ for which all $w \in U_\epsilon$ satisfy $L_\mu(w) \leq \inf_{w \in \mathcal{W}} L_\mu(w) + \epsilon$. Then*

$$\text{risk}\left(\mu, P^\star_{W|S}\right) - \inf_{w \in \mathcal{W}} L_\mu(w) \leq \epsilon + \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(B^{(\epsilon)}_{W_1 \ldots W_k} \middle\| Q^{(k)}_{W_1 \ldots W_k} \right) + \frac{1}{4\gamma_k} \right). \tag{13}$$

Proof By plugging in $P_{W_1 \ldots W_k|S} \leftarrow B^{(\epsilon)}_{W_1 \ldots W_k}$ in the right side of (11), and by noting that $P^\star_{W|S}$ is defined as the conditional distribution which minimizes that expression, we obtain

8. This is similar to the terminology in PAC-Bayes theory (see e.g. Catoni, 2007).
9. Bassily et al. (2018) show a relation between the difference decomposition of mutual information with PAC-Bayesian bounds.

(13). ■

In particular, for discrete $\mathcal{W}$, one may choose $\epsilon = 0$ and use the Dirac measure on $\widehat{w}$ to obtain

$$\operatorname{risk}\left(\mu, P^{\star}_{W|S}\right) - \inf_{w\in\mathcal{W}} L_\mu(w) \le \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{4\gamma_k}\right), \tag{14}$$

where, for all $1 \le k \le d$,

$$D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) = \log \frac{1}{Q^{(k)}_{W_1\ldots W_k}(\widehat{w}_1, \ldots, \widehat{w}_k)}.$$

For a high-probability version of Theorem 7, see Appendix C. A case of special and practical interest is when the prior distributions are consistent, that is, when there exists a single distribution $Q_{W_1\ldots W_d}$ such that $Q^{(k)}_{W_1\ldots W_k} = Q_{W_1\ldots W_k}$ for all $1 \le k \le d$. In this case, both (12) and (13) can be expressed with the following divergence:

Definition 8 (Multilevel relative entropy) *For probability measures $P_{X_1\ldots X_n}$ and $Q_{X_1\ldots X_n}$, and a vector $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n_+$, define the* multilevel relative entropy *as*

$$D_{(\mathbf{a})}\left(P_{X_1\ldots X_n} \| Q_{X_1\ldots X_n}\right) \triangleq \sum_{i=1}^{n} a_i D\left(P_{X_1\ldots X_i} \| Q_{X_1\ldots X_i}\right). \tag{15}$$

The prior distributions $Q^{(k)}_{W_1\ldots W_k}$ may be given by Gaussian matrices truncated on bounded-norm sets.

It is shown by Xu and Raginsky (2017) (with a related result by Zhang, 2006b) that the Gibbs posterior distribution $P^{\gamma,Q}_{W|S} \propto e^{-\gamma L_s(w)} Q$, as defined precisely in Definition 29 in Appendix D, is the unique solution to

$$\arg\min_{P_{W|S}} \left\{ \mathbb{E}[L_S(W)] + \frac{1}{\gamma} D(P_{W|S} \| Q | P_S) \right\},$$

where γ is called the *inverse temperature*. Thus, based on (12), the desired distribution $P^{\star}_{W|S}$ is a multiscale extension of the Gibbs distribution. In the next section, we obtain the functional form of $P^{\star}_{W|S}$. Inspired from the terminology for the Gibbs distribution, we call the vector of coefficients $\left(\frac{C\beta_1\gamma_1}{\sqrt{n}}, \ldots, \frac{C\beta_d\gamma_d}{\sqrt{n}}\right)$ in (12) the *temperature vector* of $P^{\star}_{W|S}$. Note that for minimizing the excess risk bound (13), the optimal value for γ_k, for all $1 \le k \le d$, is

$$\gamma^{\star}_k = \frac{1}{2\sqrt{D\left(B^{(\epsilon)}_{W_1\ldots W_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right)}}.$$

Furthermore, as a byproduct of the above analysis, we give new excess risk bounds for the Gibbs distribution in Propositions 31 and 33 in Appendix D (a related result has recently been obtained by Kuzborskij et al., 2019, though using stability arguments). These results generalize Corollaries 2 and 3 of Xu and Raginsky (2017) to arbitrary subgaussian losses, and unlike their proof which is based on stability arguments of Raginsky et al. (2016), merely uses the mutual information bound (Russo and Zou (2016); Xu and Raginsky (2017)).

6. The Marginalize-Tilt (MT) Algorithm

The optimization problem (12), which was derived by *chaining* mutual information, can be solved via the *chain rule* of relative entropy, and based on a key property of conditional relative entropy (Lemma 36 in Appendix E), can be shown to have a unique solution, as we illustrate in this section. Assume that we know the solution to the following more general relative entropy sum minimization:

$$\arg\min_{P_{X_1\dots X_d}} \left\{ a_1 D\left(P_{X_1}\middle\|R^{(1)}_{X_1}\right) + a_2 D\left(P_{X_1X_2}\middle\|R^{(2)}_{X_1X_2}\right) + \cdots + a_d D\left(P_{X_1\dots X_d}\middle\|R^{(d)}_{X_1\dots X_d}\right)\right\}, \tag{16}$$

where $a_i > 0$ and distributions $R^{(i)}_{X_1\dots X_i}$ are given for all $1 \le i \le d$. Then, we can use that to solve for $P^\star_{W|S=s}$ in (12) for any $s \in \mathsf{Z}^n$, by assuming the following: $X_i \triangleq W_i$ and $a_i \leftarrow \frac{C\beta_i\gamma_i}{\sqrt{n}}$ for all $1 \le i \le d$, $R^{(i)} \leftarrow Q^{(i)}$ for all $1 \le i \le d-1$, and

$$R^{(d)}(\mathrm{d}x) \leftarrow \frac{e^{-\frac{\sqrt{n}}{C\beta_d\gamma_d}L_s(x)}Q^{(d)}(\mathrm{d}x)}{\mathbb{E}\left[e^{-\frac{\sqrt{n}}{C\beta_d\gamma_d}L_s(\widetilde{X})}\right]}, \quad \widetilde{X}\sim Q^{(d)},$$

where we combined the expected empirical risk with the last relative entropy in (12) and ignored the resulting term which does not depend of $P_{X_1\dots X_n}$ (such combination is similarly performed in Section IV of the work by Zhang, 2006b, for proving the optimality of the single-scale Gibbs distribution). The solution to (16), denoted as $P^\star_{X_1\dots X_d}$, is the output of Algorithm 1. If P and Q are distributions on a set $\mathcal{A}$, then let the relative information $\imath_{P\|Q}(a) = \log\frac{\mathrm{d}P}{\mathrm{d}Q}(a)$ denote the logarithm of the Radon–Nikodym derivative of P with respect to Q for all $a \in \mathcal{A}$. The algorithm uses the following notion, which is basically the geometric mixture between distributions:

Definition 9 (Tilted distribution) *Given distributions P and Q defined on a set $\mathcal{A}$, let R be a dominating measure such that $R \gg P$ and $R \gg Q$. The tilted distribution $(P,Q)_\lambda \ll R$ for $\lambda \in [0,1]$ is defined with*

$$\imath_{(P,Q)_\lambda\|R}(a) = \lambda\imath_{P\|R}(a) + (1-\lambda)\imath_{Q\|R}(a) + (1-\lambda)D_\lambda(P\|Q),$$

for all $a \in \mathcal{A}$. If $P \perp Q$, then $(P,Q)_\lambda$ is not defined for $\lambda \in (0,1)$.

Remark 10 In the special case that P and Q are distributions on a discrete set $\mathcal{A}$, for all $a \in \mathcal{A}$, we have

$$(P,Q)_\lambda(a) = \frac{P^\lambda(a)Q^{1-\lambda}(a)}{\sum_{x\in\mathcal{A}}P^\lambda(x)Q^{1-\lambda}(x)}.$$

In the case that P and Q are distributions of real-valued absolutely continuous random variables with probability density functions f_0 and f_1, the tilted random variable has probability density function

$$f_\lambda(x) = \frac{e^{\lambda\log f_0(x)+(1-\lambda)\log f_1(x)}}{\int_{-\infty}^{\infty}e^{\lambda\log f_0(t)+(1-\lambda)\log f_1(t)}\mathrm{d}t}.$$

Notice that $(P,Q)_\lambda$ traverses between Q and P as λ traverses between 0 and 1.

Remark 11 The tilted distribution is known as the *generalized escort distribution* in the statistical physics and the statistics literatures (see e.g. Bercher, 2012).

The following shows the useful role of tilted distributions in linearly combining relative entropies. For a proof, see Theorem 30 of Van Erven and Harremos (2014).

Lemma 12 *Let $\lambda \in [0,1]$. For any $P \ll Q$ and $P \ll R$,*

$$\lambda D(P\|Q) + (1-\lambda)D(P\|R) = D\left(P\|(Q,R)_\lambda\right) + (1-\lambda)D_\lambda(Q\|R).$$

Algorithm 1 Marginalize-tilt (MT)

Input: Distributions $R^{(i)}_{X_1\ldots X_i}$ and coefficients a_i, for all $1 \le i \le d$.
Output: Solution $P^\star_{X_1\ldots X_d}$ to the minimization problem (16).

$U^{(d)}_{X_1\ldots X_d} \leftarrow R^{(d)}_{X_1\ldots X_d}$
for $k = d-1$ `to` 1 **do**
  $M_{X_1\ldots X_k} \leftarrow U^{(k+1)}_{X_1\ldots X_k}$ ▷ The marginalization step
  $U^{(k)}_{X_1\ldots X_k} \leftarrow \left(R^{(k)}_{X_1\ldots X_k}, M_{X_1\ldots X_k}\right)_{\frac{a_k}{a_k+\cdots+a_d}}$ ▷ The tilting step
return $P^\star_{X_1\ldots X_d} = U^{(1)}_{X_1} U^{(2)}_{X_2|X_1} \cdots U^{(d)}_{X_d|X_1\ldots X_{d-1}}$ ▷ The unique solution to (16)

Theorem 13 *The output of Algorithm 1 is the unique solution to (16).*

Proof Note that we can rewrite the expression in (16) as follows:

$$\begin{aligned}
&\sum_{i=1}^{d} a_i D\left(P_{X_1\ldots X_i}\|R^{(i)}_{X_1\ldots X_i}\right)\\
&= \sum_{i=1}^{d-1} a_i D\left(P_{X_1\ldots X_i}\middle\|R^{(i)}_{X_1\ldots X_i}\right)\\
&\quad + a_d\left(D\left(P_{X_1\ldots X_{d-1}}\middle\|R^{(d)}_{X_1\ldots X_{d-1}}\right) + D\left(P_{X_d|X_1\ldots X_{d-1}}\middle\|R^{(d)}_{X_d|X_1\ldots X_{d-1}}\middle|P_{X_1\ldots X_{d-1}}\right)\right) \quad (17)\\
&= \sum_{i=1}^{d-2} a_i D\left(P_{X_1\ldots X_i}\middle\|R^{(i)}_{X_1\ldots X_i}\right)\\
&\quad + (a_{d-1}+a_d) D\left(P_{X_1\ldots X_{d-1}}\middle\|\left(R^{(d-1)}_{X_1\ldots X_{d-1}}, R^{(d)}_{X_1\ldots X_{d-1}}\right)_{\frac{a_{d-1}}{a_{d-1}+a_d}}\right)\\
&\quad + a_d D_{\frac{a_{d-1}}{a_{d-1}+a_d}}\left(R^{(d-1)}_{X_1\ldots X_{d-1}}\middle\|R^{(d)}_{X_1\ldots X_{d-1}}\right)\\
&\quad + a_d D\left(P_{X_d|X_1\ldots X_{d-1}}\middle\|R^{(d)}_{X_d|X_1\ldots X_{d-1}}\middle|P_{X_1\ldots X_{d-1}}\right), \quad (18)
\end{aligned}$$

where (17) follows from the chain rule of relative entropy (see Lemma 18 in Appendix A) and (18) follows from Lemma 12. Notice that we can set

$$P^\star_{X_d|X_1\ldots X_{d-1}} \leftarrow R^{(d)}_{X_d|X_1\ldots X_{d-1}} = U^{(d)}_{X_d|X_1\ldots X_{d-1}},$$

to make the last conditional relative entropy in the right side of (18) vanish (and hence minimized, due to Lemma 36 in Appendix E), regardless of any choice for $P_{X_1 \dots X_{d-1}}$ that we may take later on. Since the Rényi divergence in (18) does not depend on $P_{X_1 \dots X_d}$, we can ignore that term, and repeat this process to the sum of the remaining terms iteratively to obtain $P^\star_{X_i|X_1 \dots X_{i-1}} = U^{(i)}_{X_i|X_1 \dots X_{i-1}}$ for all $1 \le i \le d-1$, where the *intermediate distributions* $U^{(i)}_{X_1 \dots X_i}$ are defined as in Algorithm 1. In view of the fact that

$$P^\star_{X_1 \dots X_d} = P^\star_{X_1} P^\star_{X_2|X_1} \dots P^\star_{X_d|X_1 \dots X_{d-1}},$$

we have obtained the desired distribution $P^\star_{X_1 \dots X_d}$, up to almost sure equality, as

$$P^\star_{X_1 \dots X_d} = U^{(1)}_{X_1} U^{(2)}_{X_2|X_1} \dots U^{(d)}_{X_d|X_1 \dots X_{d-1}}.$$

■

The key point of the previous proof is to rewrite the expression in (16) as the sum of some Rényi divergences which do not depend on $P_{X_1 \dots X_d}$, and some conditional relative entropies which can all be set equal to zero, *simultaneously*. This shows the uniqueness of the solution as well. The proof also implies that the minimum value of the expression in (16) is a summation of Rényi divergences between functions of distributions $R^{(i)}_{X_1 \dots X_i}$, $1 \le i \le d$.

7. Discussion on the Multiscale Gibbs Algorithm

Using the MT algorithm, we can find the functional form of the "twisted" distribution $P^\star_{W|S}$. Notice that what makes $P^\star_{W|S}$ different from the classical Gibbs distribution is the repetitive tilting steps in Algorithm 1. In fact, when the temperature vector has 0 as its first $d-1$ entries, then the multiscale Gibbs distribution $P^\star_{W|S}$ has the same exponential form of the Gibbs distribution, as (12) will have the same form of (16). In this case, Algorithm 1 only performs marginalization of the Gibbs distribution, and by definition of conditional distribution, reverses those marginalizations and simply outputs the Gibbs distribution. However, to assuredly achieve the chaining-style multiscale excess risk of Theorem 7, those tilting operations are essential.

Note that the MT algorithm describes how the intermediate distributions can be derived from each other in a back-wards manner. Once these distributions are obtained, based on line 5, samples from the multilevel Gibbs distribution can be obtained from a forward pass on these intermediate distributions. This demonstrates the forwards and backwards interactions between the different scales of the neural net while simulating $P^\star_{W|S}$.

8. Multilevel Entropic Training

We now seek an efficient sampling implementation of the multiscale Gibbs distribution. We have defined multilevel entropic training as simulating $P^\star_{W|S=s}$, given the training set $S = s$. For a two layer net, we implement this with Algorithm 2. Let $f(w_1, w_2) \triangleq e^{-L_s(w_1, w_2)}$, where w_1 and w_2 are the matrices of the first and second layer, respectively.[10] In the important

10. In this section, we are denoting matrices with lower case for clarity.

case of having consistent product priors, i.e., when we can write $Q^{(1)}(w_1) = \tilde{Q}^{(1)}(w_1)$ and $Q^{(2)}(w_1, w_2) = \tilde{Q}^{(1)}(w_1)\tilde{Q}^{(2)}(w_2)$, assuming temperature vector (a_1, a_2), we have

$$P^{\star}_{W|S=s}(w_1, w_2) = \frac{\left(\int_{v_2} f(w_1, v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2\right)^{\frac{a_2}{a_1+a_2}} \tilde{Q}^{(1)}(w_1)}{\int_{v_1}\left(\int_{v_2} f(v_1, v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2\right)^{\frac{a_2}{a_1+a_2}} \tilde{Q}^{(1)}(v_1)\mathrm{d}v_1} \times \frac{f(w_1, w_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(w_2)}{\int_{v_2} f(w_1, v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}, \quad (19)$$

see Appendix F for more details. Algorithm 2 consists of two Metropolis algorithms, one

Algorithm 2 Two-level Metropolis

Input: Distributions $\tilde{Q}^{(1)}$ and $\tilde{Q}^{(2)}$, temperature vector $\mathbf{a} = (a_1, a_2)$, proposals q_1 and q_2, inner level running time T', and initializations $(w_1^{(1)}, w_2^{(0)})$.

Output: A sequence $(w_1^{(t)}, w_2^{(t)})_{t=1}^{T}$ drawn from $P^{\star}_{W|S=s}$ in (19).

1: **for** $t = 1$ `to` T **do**

2: $\quad \widehat{w}_1 \sim q_1\left(w_1^{(t)}\right)$ ▷ Symmetric proposal

3: $\quad$ Initialize $v_2^{(0)} \leftarrow w_2^{(t-1)}$, generate sequence $\{v_2^{(i)}\}_{i=0}^{T'}$ drawn from distribution $\frac{f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)}{\int_{v_2} f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}$, and let $w_2^{(t)} \leftarrow v_2^{(T')}$. ▷ Inner level Metropolis algorithm

4: $\quad$ Approximate $\frac{\int_{v_2} f(\widehat{w}_1, v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}{\int_{v_2} f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2} \approx \frac{1}{T'}\sum_{i=1}^{T'}\left(\frac{f\left(\widehat{w}_1, v_2^{(i)}\right)}{f\left(w_1^{(t)}, v_2^{(i)}\right)}\right)^{\frac{1}{a_2}} \triangleq A.$

5: $\quad \alpha \leftarrow A^{\frac{a_2}{a_1+a_2}} \times \frac{\tilde{Q}^{(1)}(\widehat{w}_1)}{\tilde{Q}^{(1)}\left(w_1^{(t)}\right)}$ ▷ Acceptance ratio

6: $\quad U \sim \mathrm{Unif}[0, 1]$ ▷ Uniform distribution

7: $\quad$ **if** $U \le \alpha$ **then**

8: $\quad\quad w_1^{(t+1)} \leftarrow \widehat{w}_1$ ▷ Accept proposal

9: $\quad$ **else** $w_1^{(t+1)} \leftarrow w_1^{(t)}$ ▷ Reject proposal and keep current state

in an outer level to sample $\{w_1^{(t)}\}_{t=1}^{T}$ with distribution as the first fraction in (19), and the other in the inner level at line 3 to sample $\{w_2^{(i)}\}_{i=1}^{T'}$ given $w_1^{(t)}$ with conditional distribution equal to second fraction in (19). Line 4, which can be run concurrently with line 3, shows how the inner level sampling is used in the outer level algorithm: Note that to compute the acceptance ratio of the outer level algorithm, we can write

$$\frac{\int_{v_2} f\left(\widehat{w}_1, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}{\int_{v_2} f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2} = \frac{\int_{v_2}\left(\frac{f(\widehat{w}_1, v_2)}{f\left(w_1^{(t)}, v_2\right)}\right)^{\frac{1}{a_2}} f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}{\int_{v_2} f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2},$$

$$= \mathbb{E}\left[\left(\frac{f\left(\widehat{w}_1, V_2\right)}{f\left(w_1^{(t)}, V_2\right)}\right)^{\frac{1}{a_2}}\right],$$

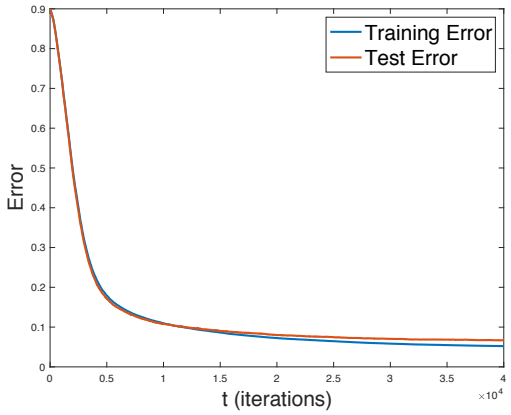

Figure 1: Training and test errors in Example 2

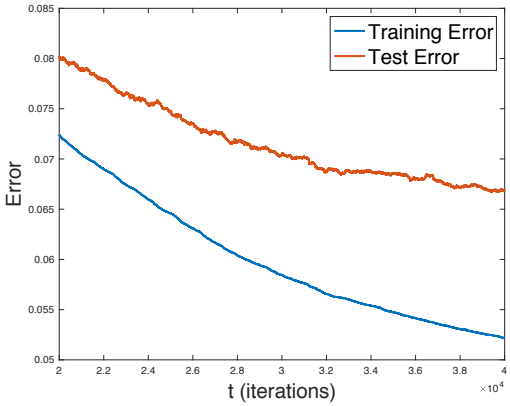

Figure 2: Training and test errors in Example 2

where for any fixed $w_1^{(t)}$,

$$V_2 \sim \frac{f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}{\int_{v_2} f\left(w_1^{(t)}, v_2\right)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}.$$

This justifies the Monte Carlo approximation in line 4. The initialization at line 3 is chosen to let the inner level algorithm mix faster along with the mixing of the outer level algorithm. Algorithm 2 reduces the dimensionality of the proposal distributions, which is a desired property, compared to simulating the Gibbs distribution when w_1 and w_2 are sampled jointly. For more details and explanations about Algorithm 2, see Appendix F.

Example 2 We tested a basic implementation of Algorithm 2 with random walk Gaussian proposals on the MNIST data set (as a proof of concept). We used a two-layer net of size $784 - 100 - 10$ with ReLU activation function for the hidden layer, soft-max activation function for the output layer, and with squared ℓ_2 loss function. We let $\mathbf{a} = (2 \times 10^{-6}, 10^{-6})$, $T' = 10$ and ran the outer level algorithm for $T = 40000$ iterations; see Figures 1 and 2. This number of iterations is large, in part due to the fact that we did not use any tricks to speed up the algorithm, such as tuning the proposals variances during the burn-in period, or lowering the temperatures gradually as in simulated annealing. For more details about this experiment, see Appendix G. The code is available at `https://github.com/ARAsadi/Multilevel-Metropolis`.

Tuning the temperature parameter for simulating the Gibbs distribution is usually done with cross-validation (see Catoni, 2007; Guedj, 2019). We leave for future work the problem of tuning the temperature vector for achieving low test error while having low mixing time.

To simulate $P^{\star}_{W|S}$ for more than two layers, similar to line 4 of Algorithm 2, one can compute Monte Carlo approximations to the acceptance ratio of each layer, based on the samples from the next layers and the inner level algorithms. However, this will make the

algorithm computationally expensive to scale up on larger neural networks and it would be interesting to make this algorithm faster.

Various ideas could be used to decrease the running time of simulating the multiscale Gibbs distribution $P^{\star}_{W|S}$. In particular, one may use gradients as in Hamiltonian Monte Carlo (Neal, 1992; Chen et al., 2014) and stochastic gradient Langevin dynamics (Welling and Teh, 2011), divide the training set into mini-batches with divide-and-conquer approaches, use subsampling methods (Bardenet et al., 2017), or simulate a variational Bayes approximation to the multiscale Gibbs distribution (Alquier et al., 2016, discuss approximating the single-scale Gibbs distribution). However, significant work is required in order to obtain algorithms with efficiency comparable to SGD. We leave this for future work.

Remark 14 As a side result, in Appendix H, we show how to alternatively achieve the excess risk bound of Theorem 7 with an *average predictor* for the special case of binary classification with ℓ_1 loss, based on an idea of Cesa-Bianchi and Lugosi (1999).

Acknowledgments

We are grateful to Ramon van Handel for his generous time and for the many discussions on chaining.

Appendix A. Information-Theoretic Tools

In this section, we present some information-theoretic tools.

Definition 15 (Relative information) *Given probability measures P and Q defined on a measurable space $(\mathcal{A}, \mathscr{F})$, such that $P \ll Q$, the relative information between P and Q in $a \in \mathcal{A}$ is the logarithm of the Radon–Nikodym derivative of P with respect to Q:*

$$\imath_{P\|Q}(a) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a).$$

Definition 16 (Relative entropy) *The relative entropy between distributions P and Q defined on the same measurable space $(\mathcal{A}, \mathscr{F})$, if $P \ll Q$, is*

$$D(P\|Q) = \mathbb{E}[\imath_{P\|Q}(X)], \qquad X \sim P,$$

otherwise, we define $D(P\|Q) = \infty$.

Definition 17 (Conditional relative entropy) *The conditional relative entropy is defined as*

$$\begin{aligned} D(P_{Y|X}\|Q_{Y|X}|P_X) &= \int D(P_{Y|X=\omega}\|Q_{Y|X=\omega})\mathrm{d}P_X(\omega) \\ &= \mathbb{E}[D(P_{Y|X}(\cdot|X)\|Q_{Y|X}(\cdot|X))], \quad X \sim P_X. \end{aligned}$$

The following lemma is known as the chain rule of relative entropy. For a proof of this property of relative entropy, see e.g. Theorem 2.5.3 of Cover and Thomas (2012):

Lemma 18 (Chain rule of relative entropy) *We have*

$$D(P_{XY}\|Q_{XY}) = D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X).$$

More generally,

$$D\left(P_{X_1\ldots X_n}\|Q_{X_1\ldots X_n}\right) = \sum_{i=1}^{n} D\left(P_{X_i|X_1\ldots X_{i-1}}\|Q_{X_i|X_1\ldots X_{i-1}}|P_{X_1\ldots X_{i-1}}\right).$$

The following is a well-known and important property of mutual information:

Lemma 19 (Difference decomposition of mutual information) *For any Q_Y such that $D(P_Y\|Q_Y) < \infty$, we have*

$$I(X;Y) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y).$$

We give the general definition of Rényi divergence from Verdú (2015):

Definition 20 (Rényi divergence) *Given distributions P and Q defined on the same probability space, let probability measure R be such that $P \ll R$ and $Q \ll R$, and let $Z \sim R$. Then, the Rényi divergence of order $\alpha \in (0,1)\cup(1,\infty)$ between P and Q is defined as*

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log \mathbb{E}\left[\exp\left(\alpha\imath_{P\|R}(Z) + (1-\alpha)\imath_{Q\|R}(Z)\right)\right].$$

Due to its limiting behavior, for $\alpha = 1$ we define $D_1(P\|Q) = D(P\|Q)$.

For instance, for discrete distributions P and Q defined on a set $\mathcal{A}$ and for any $\alpha \in (0,1)\cup(1,\infty)$, we have

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\left(\sum_{a\in\mathcal{A}} P^\alpha(a)Q^{1-\alpha}(a)\right).$$

Appendix B. Chaining Mutual Information

In this section, we strengthen the results of Asadi et al. (2018). First we give the necessary definitions:

Definition 21 (Subgaussian process) *The random process $\{X_t\}_{t\in T}$ on the metric space (T,d) is called* subgaussian *if $\mathbb{E}[X_t] = 0$ for all $t \in T$ and*

$$\mathbb{E}[e^{\lambda(X_t - X_s)}] \le e^{\frac{1}{2}\lambda^2 d^2(t,s)} \quad \text{for all} \quad t,s \in T, \lambda \ge 0.$$

The following is a technical assumption which holds in almost all cases of interest:

Definition 22 (Separable process) *The random process $\{X_t\}_{t\in T}$ is called* separable *if there is a countable set $T_0 \subseteq T$ such that $X_t \in \lim_{\substack{s\to t\\ s\in T_0}} X_s$ for all $t \in T$ a.s., where $x \in \lim_{\substack{s\to t\\ s\in T_0}} x_s$ means that there is a sequence (s_n) in T_0 such that $s_n \to t$ and $x_{s_n} \to x$.*

For instance, if $t \to X_t$ is continuous almost surely, then X_t is a separable process (see e.g. van Handel, 2016).

Notice that, unlike a partition, an exact cover $\mathcal{P} = \{A_i : i \in M\}$ of the set T may have countably or uncountably infinite number of blocks, that is, M may have countably or uncountably infinite size.

Definition 23 (ϵ-cover) *We call a cover $\mathcal{P} = \{A_i : i \in M\}$ of the set T an* ϵ-cover *of the metric space (T, d) if for all $i \in M$, A_i can be contained within a ball of radius ϵ.*

Definition 24 (Hierarchical sequence of covers) *A sequence of covers $\{\mathcal{P}_k\}_{k=m}^{\infty}$ of a set T is called a* hierarchical sequence *(or an* increasing sequence*) if for all $k \geq m$ and each $A \in \mathcal{P}_{k+1}$, there exists $B \in \mathcal{P}_k$ such that $A \subseteq B$. For any such sequence of exact covers and any $t \in T$, let $[t]_k$ denote the unique set $A \in \mathcal{P}_k$ such that $t \in A$.*

If $\mathcal{N}$ is a set, let $X_{\mathcal{N}} \triangleq \{X_i : i \in N\}$ denote a random process indexed by the elements of $\mathcal{N}$. For any bounded metric space (T, d), let $k_1(T)$ be an integer such that $2^{-(k_1(T)-1)} \geq \operatorname{diam}(T)$.

Theorem 25 *Assume that $\{\operatorname{gen}(w)\}_{w\in\mathcal{W}}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W}, d)$. Let $\{\mathcal{P}_k\}_{k=k_1(\mathcal{W})}^{\infty}$ be a hierarchical sequence of exact coverings of $\mathcal{W}$, where for each $k \geq k_1(\mathcal{W})$, $\mathcal{P}_k$ is a 2^{-k}-cover of $(\mathcal{W}, d)$.*

(a) Then,

$$\operatorname{gen}(\mu, P_{W|S}) \leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I([W]_k; S)},$$

(b) If $\mathbf{0} \in \{\ell(h_w, \cdot) : w \in \mathcal{W}\}$, then

$$\operatorname{gen}^+(\mu, P_{W|S}) \leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I([W]_k; S) + \log 2},$$

where $\mathbf{0}$ is a function identically equal to zero and $\operatorname{gen}^+(\mu, P_{W|S}) \triangleq \mathbb{E}\left[|L_\mu(W) - L_S(W)|\right]$.

Theorem 25 is in the context of statistical learning. The more general counterpart in the context of random processes is Theorem 26:

Theorem 26 *Assume that $\{X_t\}_{t\in T}$ is a separable subgaussian process on the bounded metric space (T, d). Let $\{\mathcal{P}_k\}_{k=k_1(T)}^{\infty}$ be a hierarchical sequence of exact coverings of T, where for each $k \geq k_1(T)$, $\mathcal{P}_k$ is a 2^{-k}-cover of (T, d). Let W be a random variable taking values from T.*

(a) Then,

$$\mathbb{E}[X_W] \leq 3\sqrt{2} \sum_{k=k_1(T)}^{\infty} 2^{-k} \sqrt{I([W]_k; X_T)}.$$

(b) For any arbitrary $t_0 \in T$,

$$\mathbb{E}[|X_W - X_{t_0}|] \leq 3\sqrt{2} \sum_{k=k_1(T)}^{\infty} 2^{-k} \sqrt{I([W]_k; X_T) + \log 2}.$$

Proof For an arbitrary $k \geq k_1(T)$, consider $\mathcal{P}_k = \{A_i^{(k)} : i \in M_k\}$. Since $\mathcal{P}_k$ is a 2^{-k}-cover of (T, d), based on Definition 23, there exists a multi-set $\mathcal{N}_k \triangleq \{a_i : i \in M_k\} \subseteq T$ and a mapping $\pi_{\mathcal{N}_k} : T \to \mathcal{N}_k$ such that $\pi_{\mathcal{N}_k}(t) = a_i$ if $t \in A_i^{(k)}$ for all $i \in M_k$, and $d\left(t, \pi_{\mathcal{N}_k}(t)\right) \leq 2^{-k}$ for all $t \in T$. For an arbitrary $t_0 \in T$, let $\mathcal{N}_{k_0} \triangleq \{t_0\}$. For any integer $n \geq k_1(T)$, we can write

$$X_W = X_{t_0} + \sum_{k=k_1(T)}^{n} \left(X_{\pi_{\mathcal{N}_k}(W)} - X_{\pi_{\mathcal{N}_{k-1}}(W)}\right) + \left(X_W - X_{\pi_{\mathcal{N}_n}(W)}\right).$$

Based on the definition of subgaussian processes, the process is centered, thus $\mathbb{E}[X_{t_0}] = 0$. Therefore

$$\mathbb{E}[X_W] - \mathbb{E}\left[X_W - X_{\pi_{\mathcal{N}_n}(W)}\right] = \sum_{k=k_1(T)}^{n} \mathbb{E}\left[X_{\pi_{\mathcal{N}_k}(W)} - X_{\pi_{\mathcal{N}_{k-1}}(W)}\right].$$

For every $k \geq k_1(T)$ and $t \in T$, based on the triangle inequality,

$$\begin{aligned} d\left(\pi_{\mathcal{N}_k}(t), \pi_{\mathcal{N}_{k-1}}(t)\right) &\leq d\left(t, \pi_{\mathcal{N}_k}(t)\right) + d\left(t, \pi_{\mathcal{N}_{k-1}}(t)\right) \\ &\leq 3 \times 2^{-k}. \end{aligned}$$

Knowing the value of $\left(\pi_{\mathcal{N}_k}(W), \pi_{\mathcal{N}_{k-1}}(W)\right)$ is sufficient to determine which one of the random variables $\left\{X_{\pi_{\mathcal{N}_k}(t)} - X_{\pi_{\mathcal{N}_{k-1}}(t)}\right\}_{t \in T}$ is chosen according to W. Therefore $\left(\pi_{\mathcal{N}_k}(W), \pi_{\mathcal{N}_{k-1}}(W)\right)$ is playing the role of the random index, and since $X_{\pi_{\mathcal{N}_k}(t)} - X_{\pi_{\mathcal{N}_{k-1}}(t)}$ is $d^2\left(\pi_{\mathcal{N}_k}(t), \pi_{\mathcal{N}_{k-1}}(t)\right)$-subgaussian, based on Theorem 2 of Xu and Raginsky (2017), an application of the data processing inequality and by summation, we have

$$\sum_{k=k_1(T)}^{n} \mathbb{E}\left[X_{\pi_{\mathcal{N}_k}(W)} - X_{\pi_{\mathcal{N}_{k-1}}(W)}\right] \leq \sum_{k=k_1(T)}^{n} 3\sqrt{2} \times 2^{-k} \sqrt{I(\pi_{\mathcal{N}_k}(W), \pi_{\mathcal{N}_{k-1}}(W); X_T)}.$$

Since $\{\mathcal{P}_k\}_{k=k_1(T)}^{\infty}$ is a hierarchical sequence of coverings, for any $t \in T$, knowing $\mathcal{N}_k(t)$ will uniquely determine $\mathcal{N}_{k-1}(t)$. Therefore

$$\begin{aligned} I\left(\pi_{\mathcal{N}_k}(W), \pi_{\mathcal{N}_{k-1}}(W); X_T\right) &= I\left(\pi_{\mathcal{N}_k}(W); X_T\right) \\ &= I\left([W]_k; X_T\right). \end{aligned}$$

The rest of the proof follows from the definition of separable processes and the fact that

$$\lim_{n \to \infty} \mathbb{E}\left[X_W - X_{\pi_{\mathcal{N}_n}(W)}\right] = 0.$$

■

If in Theorem 26, we let $T \triangleq \mathcal{W}$ and $X_w \triangleq \text{gen}(w)$ for all $w \in \mathcal{W}$, then for each $k \geq k_1(T)$, due to the Markov chain

$$X_T = \{\text{gen}(w)\}_{w\in\mathcal{W}} \leftrightarrow S \leftrightarrow W \leftrightarrow [W]_k \tag{20}$$

and the data processing inequality, we deduce $I([W]_k; X_T) \leq I([W]_k; S)$. Therefore Theorem 25 follows from Theorem 26.

If we use Theorem 2 of Bu et al. (2020) instead of Theorem 2 of Xu and Raginsky (2017), then we can tighten the bound of Theorem 25 to the following result. Recall that $S = (Z_1, \ldots, Z_n)$ denotes the training set.

Proposition 27 *Assume that $\{\text{gen}(w)\}_{w\in\mathcal{W}}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W}, d)$. Let $\{\mathcal{P}_k\}_{k=k_1(\mathcal{W})}^{\infty}$ be an increasing sequence of partitions of $\mathcal{W}$, where for each $k \geq k_1(\mathcal{W})$, $\mathcal{P}_k$ is a 2^{-k}-partition of $(\mathcal{W}, d)$. Then*

$$\text{gen}(\mu, P_{W|S}) \leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \left(\sum_{i=1}^{n} \sqrt{I([W]_k; Z_i)} \right), \tag{21}$$

Appendix C. Proofs for Generalization and Excess Risk Bounds

Proof [of Lemma 4] Since ϕ is 1-Lipschitz and $\phi(0) = 0$, for all vectors $\mathbf{x}$ we have $|\phi(\mathbf{x})|_2 \leq |\mathbf{x}|_2$. Based on the triangle inequality, for all $1 \leq i \leq k-1$, we can write

$$\begin{aligned} \|\mathbf{W}_i\|_2 &\leq \|\mathbf{W}_i - \mathbf{M}_i\|_2 + \|\mathbf{M}_i\|_2 \\ &\leq (\alpha_i + 1)\|\mathbf{M}_i\|_2 \\ &\leq \exp(\alpha_i)\|\mathbf{M}_i\|_2. \end{aligned}$$

Thus, for all $\mathbf{x} \in \mathcal{X}$,

$$|\sigma(\mathbf{W}_{k-1}(\ldots\sigma(\mathbf{W}_1(\mathbf{x}))\ldots))|_2 \leq \exp\left(\sum_{i=1}^{k-1}\alpha_i\right)\left(\prod_{i=1}^{k-1}\|\mathbf{M}_i\|_2\right)|\mathbf{x}|_2.$$

This yields

$$\begin{aligned} |\sigma(\mathbf{W}_k(\ldots(\sigma(\mathbf{W}_1(\mathbf{x}))\ldots)))-\sigma(\mathbf{M}_k(\ldots(\sigma(\mathbf{W}_1(\mathbf{x}))\ldots)))|_2 & \\ \leq \exp\left(\sum_{i=1}^{k-1}\alpha_i\right)\left(\prod_{i=1}^{k-1}\|\mathbf{M}_i\|_2\right)\|\mathbf{W}_k - \mathbf{M}_k\|_2|\mathbf{x}|_2 & \\ \leq \alpha_k \exp\left(\sum_{i=1}^{k-1}\alpha_i\right)\left(\prod_{i=1}^{k}\|\mathbf{M}_i\|_2\right)|\mathbf{x}|_2. & \end{aligned}$$

Since $\mathbf{M}_i$ is $\|\mathbf{M}_i\|_2$-Lipschitz for all $k+1 \leq i \leq d$, and soft-max is 1-Lipschitz with respect to the Euclidean norm (see e.g. Gao and Pavel, 2017), we conclude that

$$|h_{w_1}(\mathbf{x}) - h_{w_2}(\mathbf{x})|_2 \leq \alpha_k \exp\left(\sum_{i=1}^{k-1}\alpha_i\right) M|\mathbf{x}|_2.$$

■

In the following, the notation $P_X \to Q_{Y|X} \to P_Y$ indicates that the joint distribution of X and Y is $P_{XY} = P_X Q_{Y|X}$. We state a high-probability result for the case of discrete $\mathcal{W}$, the more general case is similar:

Corollary 28 *For a given μ, let $\widehat{w}(\mu)$ denote the index of a hypothesis which achieves the minimum statistical risk among $\mathcal{W}$. If $P_S \to P^{\star}_{W|S} \to P_W$, then*

$$\mathbb{P}\left[L_\mu(W) \leq \inf_{w\in\mathcal{W}} L_\mu(w) + \epsilon\right] \geq 1 - \frac{C}{\epsilon\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{4\gamma_k}\right). \quad (22)$$

Proof Based on Theorem 7, we have

$$\mathbb{E}[L_\mu(W)] - \inf_{w\in\mathcal{W}} L_\mu(w) \leq \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{4\gamma_k}\right).$$

Thus

$$\mathbb{E}\left[L_\mu(W) - \inf_{w\in\mathcal{W}} L_\mu(w)\right] \leq \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{4\gamma_k}\right).$$

Since $L_\mu(W) - \inf_{w\in\mathcal{W}} L_\mu(w)$ is a positive random variable, by Markov's inequality we obtain

$$\mathbb{P}\left[L_\mu(W) - \inf_{w\in\mathcal{W}} L_\mu(w) > \epsilon\right] \leq \frac{C}{\epsilon\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{4\gamma_k}\right),$$

which yields

$$\mathbb{P}\left[L_\mu(W) \leq \inf_{w\in\mathcal{W}} L_\mu(w) + \epsilon\right] \geq 1 - \frac{C}{\epsilon\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{4\gamma_k}\right).$$

■

Appendix D. Gibbs Distribution Results

Definition 29 (Gibbs distribution) *The Gibbs (posterior) distribution associated to parameter γ and prior distribution Q, is denoted with $P^{\gamma,Q}_{W|S}$ and defined as follows:*

$$P^{\gamma,Q}_{W|S=s}(\mathrm{d}w) \triangleq \frac{e^{-\gamma L_s(w)} Q(\mathrm{d}w)}{\mathbb{E}[e^{-\gamma L_s(\widetilde{W})}]}, \qquad \widetilde{W} \sim Q.$$

Lemma 30 *[Xu and Raginsky (2017)] The Gibbs distribution $P^{\gamma,Q}_{W|S}$ is the unique solution to the optimization problem*

$$\arg\min_{P_{W|S}} \left\{\mathbb{E}[L_S(W)] + \frac{1}{\gamma} D(P_{W|S} \| Q | P_S)\right\}.$$

The next results are new excess risk bounds for the Gibbs distribution:

Proposition 31 *Assume that $\mathcal{W}$ is a countable set. For any input distribution μ, let $\widehat{w}(\mu)$ denote the index of a hypothesis which achieves the minimum statistical risk among $\mathcal{W}$. If for all $w \in \mathcal{W}$, $\ell(w, Z)$ is σ^2-subgaussian where $Z \sim \mu$, then for any $\gamma > 0$,*

$$\operatorname{risk}\left(\mu, P_{W|S}^{\gamma,Q}\right) \leq \inf_{w\in\mathcal{W}} L_\mu(w) + \frac{1}{\gamma} D\left(\delta_{\widehat{w}(\mu)} \middle\| Q\right) + \frac{\gamma\sigma^2}{2n}. \tag{23}$$

Proof Assuming $\mu^{\otimes n} = P_S \to P_{W|S}^{\gamma,Q} \to P_W$, we can write

$$\begin{aligned}
\operatorname{risk}\left(\mu, P_{W|S}^{\gamma,Q}\right) &= \mathbb{E}[L_\mu(W)] \\
&\leq \mathbb{E}[L_S(W)] + \sqrt{\frac{2\sigma^2}{n}}.\sqrt{I(S;W)} \\
&\leq \mathbb{E}[L_S(W)] + \sqrt{\frac{2\sigma^2}{n}}\left(\frac{1}{\gamma}\sqrt{\frac{n}{2\sigma^2}} I(S;W) + \frac{1}{4\frac{1}{\gamma}\sqrt{\frac{n}{2\sigma^2}}}\right) \qquad (24) \\
&= \mathbb{E}[L_S(W)] + \frac{1}{\gamma} I(S;W) + \frac{\gamma\sigma^2}{2n} \\
&\leq \mathbb{E}[L_S(W)] + \frac{1}{\gamma} D\left(P_{W|S}^{\gamma,Q} \middle\| Q \middle| P_S\right) + \frac{\gamma\sigma^2}{2n} \\
&\leq \inf_{w\in\mathcal{W}} L_\mu(w) + \frac{1}{\gamma} D\left(\delta_{\widehat{w}(\mu)} \middle\| Q\right) + \frac{\gamma\sigma^2}{2n}, \qquad (25)
\end{aligned}$$

where (24) follows from the inequality

$$\sqrt{x} \leq cx + \frac{1}{4c} \iff 0 \leq \left(\sqrt{cx} - \frac{1}{2\sqrt{c}}\right)^2 \qquad \text{for all } x, c > 0, \tag{26}$$

which is upper bounding $\sqrt{x}$ with a tangent line, and (25) follows from Lemma 30 and by plugging $P_{W|S} \leftarrow \delta_{\widehat{w}(\mu)}$ into

$$\mathbb{E}[L_S(W)] + \frac{1}{\gamma} D(P_{W|S} \| Q | P_S).$$

■

Corollary 32 *If we set $\gamma \leftarrow \gamma^\star \triangleq \frac{1}{\sigma}\sqrt{2nD\left(\delta_{\widehat{w}(\mu)} \middle\| Q\right)}$, then we minimize the right side of (23) to obtain*

$$\operatorname{risk}\left(\mu, P_{W|S}^{Q,\gamma^\star}\right) \leq \inf_{w\in\mathcal{W}} L_\mu(w) + \sigma\sqrt{\frac{D\left(\delta_{\widehat{w}(\mu)} \middle\| Q\right)}{2n}}.$$

Proposition 33 *Assume that $\mathcal{W}$ is an uncountable set. For any input distribution μ, let $\widehat{w}(\mu)$ denote the index of a hypothesis which achieves the minimum statistical risk among*

$\mathcal{W}$. If for all $w \in \mathcal{W}$, $\ell(w, Z)$ is σ^2-subgaussian where $Z \sim \mu$ and $\ell(\cdot, z)$ is ρ-Lipschitz for all $z \in \mathsf{Z}$, then for any $\gamma > 0$,

$$\text{risk}\left(\mu, P_{W|S}^{\gamma,Q}\right) \leq \inf_{w\in\mathcal{W}} L_\mu(w) + \inf_{a>0}\left(a\rho\sqrt{d} + \frac{1}{\gamma}D\left(\mathcal{N}\left(\widehat{w}(\mu), a^2 I_d\right)\middle\|Q\right)\right) + \frac{\gamma\sigma^2}{2n},$$

where $\mathcal{N}\left(\widehat{w}(\mu), a^2 I_d\right)$ denotes the Gaussian distribution centered at $\widehat{w}(\mu)$ with covariance matrix $a^2 I_d$.

Proof Assuming $\mu^{\otimes n} = P_S \to P_{W|S}^{\gamma,Q} \to P_W$, we can write

$$\begin{aligned}
\text{risk}\left(\mu, P_{W|S}^{\gamma,Q}\right) &= \mathbb{E}[L_\mu(W)] \\
&\leq \mathbb{E}[L_S(W)] + \sqrt{\frac{2\sigma^2}{n}}.\sqrt{I(S;W)} \\
&\leq \mathbb{E}[L_S(W)] + \sqrt{\frac{2\sigma^2}{n}}\left(\frac{1}{\gamma}\sqrt{\frac{n}{2\sigma^2}}I(S;W) + \frac{1}{4\frac{1}{\gamma}\sqrt{\frac{n}{2\sigma^2}}}\right) \qquad (27)\\
&= \mathbb{E}[L_S(W)] + \frac{1}{\gamma}I(S;W) + \frac{\gamma\sigma^2}{2n} \\
&\leq \mathbb{E}[L_S(W)] + \frac{1}{\gamma}D\left(P_{W|S}^{\gamma,Q}\middle\|Q\middle|P_S\right) + \frac{\gamma\sigma^2}{2n} \\
&\leq \inf_{w\in\mathcal{W}} L_\mu(w) + \inf_{a>0}\left(a\rho\sqrt{d} + \frac{1}{\gamma}D\left(\mathcal{N}(\widehat{w}(\mu), a^2 I_d)\|Q\right)\right) + \frac{\gamma\sigma^2}{2n}, \qquad (28)
\end{aligned}$$

where (27) follows from the inequality (26), and (28) follows from Lemma 30 and by plugging $P_{W|S} \leftarrow \mathcal{N}\left(\widehat{w}(\mu), a^2 I_d\right)$ into

$$\mathbb{E}[L_S(W)] + \frac{1}{\gamma}D(P_{W|S}\|Q|P_S),$$

while writing

$$\mathbb{E}[L_S(W)] \leq \inf_{w\in\mathcal{W}} L_\mu(w) + a\rho\sqrt{d} \qquad (29)$$

and taking infimum over $a > 0$. Inequality (29) is based on the proof of Corollary 3 of Xu and Raginsky (2017). ■

More generally, in the context of empirical processes, let $\mathcal{F} = \{f_w : w \in \mathcal{W}\}$ be a collection of measurable functions from a set Z to $\mathbb{R}$, indexed by the set $\mathcal{W}$. Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d elements drawn from Z with distribution μ, and define $S = (Z_1, \ldots, Z_n)$. For each $w \in \mathcal{W}$, define the empirical mean of function f_w as

$$\mu_n(f_w) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_w(Z_i),$$

and its true mean as

$$\mu(f_w) \triangleq \mathbb{E}\left[f_w(Z)\right], \quad Z \sim \mu.$$

One can prove the following proposition, analogous to the poof of Proposition 31:

Proposition 34 *Assume that $\mathcal{W}$ is a countable set. For any input distribution μ, let $\widehat{w}(\mu)$ denote the index of a function which has the minimum true mean among functions in $\mathcal{F}$. If $f_w(Z),\ Z \sim \mu$ is σ^2-subgaussian for all $w \in \mathcal{W}$, then for any $\gamma > 0$,*

$$\mathbb{E}\left[\mu(f_W)\right] \leq \inf_{w \in \mathcal{W}} \mu(f_w) + \frac{1}{\gamma} D(\delta_{\widehat{w}(\mu)} \| Q) + \frac{\gamma \sigma^2}{2n},$$

where $\mu^{\otimes n} = P_S \to P_{W|S}^{\gamma,Q} \to P_W$.

Appendix E. Tools for the MT Algorithm

We first state the following lemmas. Lemma 35 shows the useful role of tilted distributions in linearly combining relative entropies. For a proof, see Theorem 30 of Van Erven and Harremos (2014).

Lemma 35 *Let $\lambda \in [0,1]$. For any $P \ll Q$ and $P \ll R$,*

$$\lambda D(P\|Q) + (1-\lambda) D(P\|R) = D\left(P\|(Q,R)_\lambda\right) + (1-\lambda) D_\lambda(Q\|R),$$

where $(Q,R)_\lambda$ denotes the tilted distribution. Therefore

$$\arg\min_P \left\{\lambda D(P\|Q) + (1-\lambda) D(P\|R)\right\} = (Q,R)_\lambda,$$

and

$$\min_P \left\{\lambda D(P\|Q) + (1-\lambda) D(P\|R)\right\} = (1-\lambda) D_\lambda(Q\|R).$$

The next lemma is a crucial property of conditional relative entropy:

Lemma 36 *Given distribution P_X defined on a set $\mathcal{A}$ and conditional distributions $P_{Y|X}$ and $Q_{Y|X}$, we have*

$$D(P_{Y|X}\|Q_{Y|X}|P_X) \geq 0, \tag{30}$$

with equality if and only if $P_{Y|X} = Q_{Y|X}$ holds on a set $\mathcal{A}' \subseteq \mathcal{A}$ of conditioning values with $P_X(\mathcal{A}') = 1$.

The simplest case of (16) is when $d = 2$, whose solution, characterized by the following result, is useful for obtaining the solution to the general case:

Proposition 37 *Let Q_X and R_{XY} be two arbitrary distributions. For any $a_1, a_2 > 0$, we have*

$$\arg\min_{P_{XY}} \left(a_1 D(P_X\|Q_X) + a_2 D(P_{XY}\|R_{XY})\right) = P_{XY}^\star, \tag{31}$$

where

$$\begin{cases} P_X^\star = (Q_X, R_X)_{\frac{a_1}{a_1+a_2}}, \\ P_{Y|X}^\star = R_{Y|X}. \end{cases}$$

Proof Based on the chain rule of relative entropy, we have

$$D(P_{XY}\|R_{XY}) = D(P_X\|R_X) + D(P_{Y|X}\|R_{Y|X}|P_X).$$

Therefore

$$\begin{aligned}
&a_1 D(P_X\|Q_X) + a_2 D(P_{XY}\|R_{XY}) \\
&= a_1 D(P_X\|Q_X) + a_2(D(P_X\|R_X) + D(P_{Y|X}\|R_{Y|X}|P_X)) \\
&= (a_1 D(P_X\|Q_X) + a_2 D(P_X\|R_X)) + a_2 D(P_{Y|X}\|R_{Y|X}|P_X) \\
&= (a_1+a_2) D\left(P_X \middle\| (Q_X, R_X)_{\frac{a_1}{a_1+a_2}}\right) + a_2 D_{\frac{a_1}{a_1+a_2}}(Q_X\|R_X) + a_2 D(P_{Y|X}\|R_{Y|X}|P_X), \quad (32)
\end{aligned}$$

where (32) is based on Lemma 35. Note that, due to Lemma 36, distribution $P^\star_{XY}$ is the unique distribution for which both relative entropies vanish simultaneously, and since the Rényi divergence does not depend on P_{XY}, Equation (31) is proven. ■

Appendix F. The Two-level Metropolis Algorithm

Using the MT algorithm, we derive the twisted distribution $P^\star_{W|S}$ for a two-layer net with prior distribution $Q^{(1)}_{W_1}$ and $Q^{(2)}_{W_1W_2}$, and temperature vector (a_1, a_2), as

$$\begin{aligned}
P^\star_{W|S=s}(w_1,w_2) = &\frac{\left(\int_{v_2} f(w_1,v_2)^{\frac{1}{a_2}} Q^{(2)}(w_1,v_2)\mathrm{d}v_2\right)^{\frac{a_2}{a_1+a_2}} Q^{(1)}(w_1)^{\frac{a_1}{a_1+a_2}}}{\int_{v_1}\left(\int_{v_2} f(v_1,v_2)^{\frac{1}{a_2}} Q^{(2)}(v_1,v_2)\mathrm{d}v_2\right)^{\frac{a_2}{a_1+a_2}} Q^{(1)}(v_1)^{\frac{a_1}{a_1+a_2}}\mathrm{d}v_1} \\
&\times \frac{f(w_1,w_2)^{\frac{1}{a_2}} Q^{(2)}(w_1,w_2)}{\int_{v_2} f(w_1,v_2)^{\frac{1}{a_2}} Q^{(2)}(w_1,v_2)\mathrm{d}v_2}. \quad (33)
\end{aligned}$$

In the case of having consistent product prior distributions $Q^{(1)}(w_1) = \tilde{Q}^{(1)}(w_1)$ and $Q^{(2)}(w_1,w_2) = \tilde{Q}^{(1)}(w_1)\tilde{Q}^{(2)}(w_2)$, Equation (33) simplifies to

$$\begin{aligned}
&P^\star_{W|S=s}(w_1,w_2) \\
&= \frac{\left(\int_{v_2} f(w_1,v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2\right)^{\frac{a_2}{a_1+a_2}} \tilde{Q}^{(1)}(w_1)}{\int_{v_1}\left(\int_{v_2} f(v_1,v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2\right)^{\frac{a_2}{a_1+a_2}} \tilde{Q}^{(1)}(v_1)\mathrm{d}v_1} \times \frac{f(w_1,w_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(w_2)}{\int_{v_2} f(w_1,v_2)^{\frac{1}{a_2}} \tilde{Q}^{(2)}(v_2)\mathrm{d}v_2}.
\end{aligned}$$

Notice that we can run line 3 and line 4 of Algorithm 2 concurrently, that is, each time we sample $v_2^{(i)}$, we can compute the next term in the sum in line 4, hence the required space is a constant times the required space for storing matrices w_1 and w_2 and does not depend on the number of iterations. The computational complexity of the algorithm depends on the proposal distributions. The algorithm performs $T \times T'$ total iterations and at each of these iterations, the algorithm computes the empirical error over the entire training set.

Appendix G. Experiment

The MNIST data set is available at `http://yann.lecun.com/exdb/mnist/`. This benchmark data set has 60000 training examples and 10000 test examples consisting of images with 28×28 gray pixels and with 10 classes. We flattened the images into vectors of length 784 and normalized their values to between 0 and 1. Let $I_{m\times l}$ denote the $m \times l$ matrix with entries equal to 1 on its main diagonal and zero elsewhere. We initialized the training algorithm at the reference matrices $M_1 = I_{100\times 784}$ and $M_2 = I_{10\times 100}$. For simplicity, we let the distributions $\tilde{Q}^{(1)}$ and $\tilde{Q}^{(2)}$ to be flat distributions, and we chose the temperature vector to be $\mathbf{a} = (2 \times 10^{-6}, 10^{-6})$. The proposal distributions q_1 and q_2 are centered Gaussian distributions with independent entries having variances 0.001 and 0.0005, respectively. The training error at iteration $t = 40000$ reached 0.052154361878265 and the test error reached 0.066840303697749.

The computing infrastructure had the following specifications: 4.2 GHz Intel Core i7-7700K, 16 GB 2400 MHz DDR4 Memory, and Radeon Pro 575 4096 MB Graphics.

Appendix H. Average Predictors

Definition 38 (Gibbs average predictor) *We define the Gibbs average predictor as*

$$h_s^{\gamma,Q}(x) \triangleq \mathbb{E}[h_W(x)], \qquad W \sim P_{W|S=s}^{\gamma,Q}.$$

for all $s \in \mathsf{Z}^n$ and $x \in \mathcal{X}$, where $P_{W|S=s}^{\gamma,Q}$ is the Gibbs posterior distribution defined in Definition 29.

Notice that the Gibbs average predictor is a deterministic function from $\mathcal{X}$ to $\mathcal{Y}$. If $\ell(h, z)$ is convex in h, then based on Jensen's inequality,

$$\ell(h_s^{\gamma,Q}, z) \leq \mathbb{E}[\ell(h_W, z)], \qquad W \sim P_{W|S=s}^{\gamma,Q}. \tag{34}$$

Averaging both sides of (34) with respect to $Z \sim \mu$ and swapping the expectations on the right side gives

$$L_\mu\left(h_s^{\gamma,Q}\right) \leq \mathbb{E}[L_\mu(W)], \qquad W \sim P_{W|S=s}^{\gamma,Q}.$$

Taking expectations with respect to P_S yields

$$\begin{aligned}\mathbb{E}\left[L_\mu\left(h_S^{\gamma,Q}\right)\right] &\leq \mathbb{E}[L_\mu(W)], \qquad P_S \to P_{W|S}^{\gamma,Q} \to P_W \\ &= \text{risk}\left(\mu, P_{W|S}^{\gamma,Q}\right).\end{aligned} \tag{35}$$

Assume that $\mathcal{Y} = \{0, 1\}$ and that the loss function is the ℓ_1 loss. Based on the key idea of Equation (4.3) of Cesa-Bianchi and Lugosi (1999), since y can only take values 0 or 1, we have the following lemma:

Lemma 39 *If $\{h_w^{(k)}\}_{k=1}^d$ and $\{h_{w'}^{(k)}\}_{k=1}^d$ are collections of functions which take values from $\mathcal{X}$ to $[0, 1]$, and $\xi_k > 0$, $1 \leq k \leq d$ are such that $\sum_{k=1}^d \xi_k = 1$, then*

$$\left|\sum_{k=1}^d \xi_k h_w^{(k)}(x) - y\right| - \left|\sum_{k=1}^d \xi_k h_{w'}^{(k)}(x) - y\right| = \sum_{k=1}^d \xi_k \left[\left|h_w^{(k)}(x) - y\right| - \left|h_{w'}^{(k)}(x) - y\right|\right]. \tag{36}$$

Corollary 40 *Averaging both sides of (36) with respect to $z = (x, y) \sim \mu$ yields*

$$L_\mu\left(\sum_{k=1}^{d} \xi_k h_w^{(k)}\right) - L_\mu\left(\sum_{k=1}^{d} \xi_k h_{w'}^{(k)}\right) = \sum_{k=1}^{d} \xi_k \left(L_\mu\left(h_w^{(k)}\right) - L_\mu\left(h_{w'}^{(k)}\right)\right). \tag{37}$$

Assume that $\mathcal{W}$ is a discrete set. We now construct an average predictor which achieves the excess risk bound of Theorem 7. For all $1 \leq k \leq d$, let

$$p_k \triangleq \frac{\beta_k}{\sum_{i=1}^{d} \beta_i}.$$

Note that $\sum_{k=1}^{d} p_k = 1$. For all $1 \leq k \leq d$, let

$$\mathcal{H}_k \triangleq \left\{\frac{1}{2} + \frac{h_{[\mathbf{W}_1,\ldots,\mathbf{W}_k]} - h_{[\mathbf{W}_1,\ldots,\mathbf{W}_{k-1}]}}{2\beta_k MR} : w \in \mathcal{W}\right\}.$$

Based on inequality (6), the domain of all $h \in \mathcal{H}_k$ is $[0, 1]$. Given training set S, let $h_S^{(k)}$ be the Gibbs average predictor obtained from $\mathcal{H}_k$ with prior $Q^{(k)}$ and inverse temperature

$$\zeta_k \triangleq \frac{\sqrt{n}}{\gamma_k(\sum_{i=1}^{d} \beta_i)LMR}.$$

Based on (35), the proof of Proposition 31, and after taking average from both sides of (37) with respect to P_S, we get:

$$\begin{aligned}
\mathbb{E}\left[L_\mu\left(\sum_{k=1}^{d} p_k h_S^{(k)}\right)\right] - \inf_{w\in\mathcal{W}} L_\mu(w) &= \sum_{k=1}^{d} p_k \left(\mathbb{E}\left[L_\mu\left(h_S^{(k)}\right)\right] - \mathbb{E}\left[L_\mu\left(h_{\widehat{w}}^{(k)}\right)\right]\right) \\
&\leq LMR \sum_{k=1}^{d} p_k \left(\frac{\gamma_k(\sum_{i=1}^{d}\beta_i)}{\sqrt{n}} D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{\sum_{i=1}^{d}\beta_i}{2\sqrt{n}\gamma_k}\right) \\
&= \frac{C}{\sqrt{n}} \sum_{k=1}^{d} \beta_k \left(\gamma_k D\left(\delta_{\widehat{w}_1\ldots\widehat{w}_k} \middle\| Q^{(k)}_{W_1\ldots W_k}\right) + \frac{1}{2\gamma_k}\right).
\end{aligned}$$

Remark 41 The results of this section can be viewed as the "dual" of the results of Cesa-Bianchi and Lugosi (1999) in the batch learning context.

References

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

Amir R Asadi, Emmanuel Abbe, and Sergio Verdú. Compressing data on graphs with clusters. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1583–1587, 2017.

Amir R Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

Jean-Yves Audibert and Olivier Bousquet. PAC-Bayesian generic chaining. In *Advances in Neural Information Processing Systems*, pages 1125–1132, 2004.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Peter L Bartlett, Steven N Evans, and Philip M Long. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012*, 2018.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*, pages 25–55, 2018.

Jean-Francois Bercher. A simple probabilistic construction yielding generalized entropies and divergences, escort distributions and q-gaussians. *Physica A: Statistical Mechanics and its Applications*, 391(19):4460–4469, 2012.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, James R Lee, and Aleksander Mądry. k-server via multiscale entropic regularization. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 3–16, 2018.

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

Nicoló Cesa-Bianchi and Gábor Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 27(6):1865–1895, 1999.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. *arXiv preprint arXiv:1702.08211*, 2017.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

Erhan Çınlar. *Probability and Stochastics*, volume 261. Springer Science & Business Media, 2011.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017a.

Gintare Karolina Dziugaite and Daniel M Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: generalization properties of entropy-SGD and data-dependent priors. *arXiv preprint arXiv:1712.09376*, 2017b.

Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, pages 8430–8441, 2018.

Xavier Fernique. Evaluations de processus Gaussiens composes. In *Probability in Banach Spaces*, pages 67–83. Springer, 1976.

Pierre Gaillard and Sébastien Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory*, pages 764–796, 2015.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Benjamin Guedj. A primer on PAC-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479, 2017.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of Gibbs-ERM principle. *arXiv preprint arXiv:1902.01846*, 2019.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170, 1999.

Radford M Neal. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical report, Citeseer, 1992.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550, 2018.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30, 2016.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Philippe Rigollet and Alexandre B Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240, 2016.

Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Ramon van Handel. Probability in high dimension. *[Online]. Available:* `https://www.princeton.edu/~rvan/APC550.pdf`, Dec. 21 2016.

Sergio Verdú. Fifty years of Shannon theory. *IEEE Transactions on Information Theory*, 44 (6):2057–2078, 1998.

Sergio Verdú. α-mutual information. In *2015 Information Theory and Applications Workshop (ITA)*, pages 1–6, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

Tong Zhang. Theoretical analysis of a class of randomized regularization methods. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 156–163, 1999.

Tong Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006a.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. *arXiv preprint arXiv:1804.05862*, 2018.