Proceedings of Machine Learning Research vol 99:1–27, 2019 32nd Annual Conference on Learning Theory

Distribution-Dependent Analysis of Gibbs-ERM Principle

Ilja Kuzborskij* ILJA.KUZBORSKIJ@GMAIL.COM
DeepMind, London

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Dip. di Informatica & DSRC, Università degli Studi di Milano, Italy

Csaba Szepesvári SZEPI@GOOGLE.COM
DeepMind, London

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Gibbs-ERM is a natural idealized model of learning with stochastic optimization algorithms (such as Stochastic Gradient Langevin Dynamics and —to some extent— Stochastic Gradient Descent) which also appears in other contexts, including PAC-Bayesian theory, and sampling mechanisms. In this work we study the excess risk suffered by the Gibbs-ERM learner with non-convex, regularized empirical risk. Our goal is to understand the interplay between the data-generating distribution and the problem of learning in large hypothesis spaces. Our main results are *distribution-dependent* upper bounds on several notions of excess risk. We show that, in all cases, the distribution-dependent excess risk is essentially controlled by the *effective dimension* $\operatorname{tr}\left(\boldsymbol{H}^\star(\boldsymbol{H}^\star + \lambda \boldsymbol{I})^{-1}\right)$ of the problem, where $\boldsymbol{H}^\star$ is the Hessian matrix of the risk at a local minimum. This is a well-established notion of effective dimension appearing in the analyses of several previous algorithms, including SGD and ridge regression. Ours is the first work that brings this notion of dimension to the analysis of learning via Gibbs densities. The distribution-dependent view we advocate here improves upon earlier results of Raginsky et al. (2017), and can yield much tighter bounds depending on the interplay between the data-generating distribution and the loss function. The first part of our analysis focuses on the *localized* excess risk in the vicinity of a fixed local minimizer. This result is then extended to bounds on the *global* excess risk, by characterizing probabilities of local minima (and their complement) under Gibbs densities, a result which might be of independent interest.

1. Introduction

In the parametric setting of statistical learning, the learner is given a tuple $S = (z_1, \dots, z_m)$ of *training examples*, drawn independently of each other from a fixed and unknown probability distribution $\mathcal{D}$ supported on an *example space* $\mathcal{Z}$. Based on the training examples S the learner selects a model $\boldsymbol{w}$ from a parameter space $\mathbb{R}^d$. The learner's goal is to minimize the *statistical risk* $R(\boldsymbol{w}) = \mathbb{E}_z[\ell(\boldsymbol{w}, z)]$ of the selected model, where z is drawn from $\mathcal{D}$ and $\ell : \mathbb{R}^d \times \mathcal{Z} \to [0, M]$ is a known *loss function*, which we assume to be non-negative, bounded, and twice differentiable.
A learner following the Empirical Risk Minimization (ERM) principle selects a model with the smallest empirical risk. Often, learners also incorporate a penalty, leading to the selection of a model from the set

$$\operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ \widehat{R}_S(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|^2 \right\}, \qquad \lambda > 0, \tag{1}$$

[*] Work partly done while at the University of Milan, Italy.

where $\widehat{R}_S(\boldsymbol{w})$ is the *empirical risk* of $\boldsymbol{w}$, defined by

$$\widehat{R}_S(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \ell(\boldsymbol{w}, z_i)\,.$$

We abbreviate the regularized empirical risk by $\widehat{R}_{S,\lambda}$ and its population counterpart, $\mathbb{E}[\widehat{R}_{S,\lambda}]$, by R_λ. In this paper we study a randomized version of ERM known as *Gibbs-ERM*. A Gibbs algorithm outputs a model $\boldsymbol{w} \in \mathbb{R}^d$ sampled from the *Gibbs density*

$$\widehat{p}_{S,\gamma}(\boldsymbol{w}) = \frac{1}{Z}\, e^{-\gamma\left(\widehat{R}_S(\boldsymbol{w})+\lambda\|\boldsymbol{w}\|^2\right)}\,, \qquad \gamma > 0\,, \tag{2}$$

where $Z = \int_{\mathbb{R}^d} e^{-\gamma\left(\widehat{R}_S(\boldsymbol{w})+\lambda\|\boldsymbol{w}\|^2\right)}\,\mathrm{d}\boldsymbol{w}$ is the normalization constant and $\widehat{R}_S$ is assumed to be such that $Z < \infty$ (for instance, this is the case when $\|\cdot\|$ is any norm and $\widehat{R}_S$ is nonnegative). It is not hard to see that we obtain ERM as a special case of (2) for $\gamma \to \infty$. In the following expectations $\mathbb{E}[\cdot]$ are taken with respect to the joint distribution over the sample space $\mathcal{Z}^m \times \mathbb{R}^d$ (i.e., the product of the example space and the parameter space) unless explicitly stated otherwise, for instance $\mathbb{E}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}[\cdot]$. Gibbs-ERM reveals its usefulness when the regularized empirical risk is non-convex and (1) becomes intractable. This scenario brings out the connections between Gibbs-ERM and stochastic optimization algorithms, for instance Stochastic Gradient Langevin Dynamics (SGLD) —see below, along with a number of other settings in which Gibbs-ERM arises naturally. One tantalizing related line of research lies in understanding theoretical properties of learning in overparameterized problems, such as deep neural networks, through the prism of stochastic optimization, since in these settings Stochastic Gradient Descent (SGD) and its variants become de facto the method of choice. We believe that Gibbs-ERM principle provides an opportunity for explaining some of the learning-theoretic phenomena in this area.
In this paper we focus on the statistical properties of the Gibbs-ERM by analyzing distribution-dependent excess risk bounds. In particular, we give upper bounds on the excess risk that can be much smaller compared to the previous literature, for instance SGLD (Raginsky et al., 2017), depending on the interplay between the data-generating distribution and the loss function.

SGLD algorithm. The recent interest in stochastic gradient descent algorithms for non-convex optimization led to the study of a variant called SGLD. Apart from its simplicity, SGLD has amenable theoretical properties, such as asymptotic convergence to global minima and polynomial saddle-point escape times (Ge et al., 2015). The update rule of plain SGLD is

$$\widehat{\boldsymbol{w}}_{t+1} = \widehat{\boldsymbol{w}}_t - \eta\nabla\widehat{R}_{S,\lambda}(\widehat{\boldsymbol{w}}_t) + \sqrt{\frac{2\eta}{\gamma}}\,\boldsymbol{\xi}_t \qquad t = 0, 1, 2, \ldots$$

where $\widehat{\boldsymbol{w}}_1$ is sampled from a fixed distribution, $\boldsymbol{\xi}_t$ is a standard Gaussian "noise" vector (independent from the choice of $\widehat{\boldsymbol{w}}_1$), and η is a step size. The SGLD algorithm is known to approximate the continuous-time Langevin diffusion equation

$$\mathrm{d}\boldsymbol{w}(t) = -\nabla\widehat{R}_{S,\lambda}(\boldsymbol{w}(t))\,\mathrm{d}t + \sqrt{\frac{2}{\gamma}}\,\mathrm{d}\boldsymbol{b}(t), \quad t \geq 0 \tag{3}$$

where $\boldsymbol{b}(t)$ is the standard Brownian motion. Indeed, under appropriate assumptions on the empirical risk, one can show that the solution to (3) admits (2) as a stationary distribution (Raginsky et al., 2017; Tzen et al., 2018). While convergence in the limit is reassuring, the best known bound on the mixing time is of order $\mathrm{polylog}(1/\varepsilon)\, e^{\widetilde{\mathcal{O}}(d)}$ for non-convex empirical risks assuming that objective is smooth and dissipative (roughly speaking, assuring that the process (3) on average moves towards the origin) (Raginsky et al., 2017; Xu et al., 2018), and it is not clear whether the exponential dependence on the parameters can be eliminated.

Information Risk Minimization. Gibbs-ERM naturally arises when introducing a relative entropy regularization in the so-called Information Risk Minimization framework (Zhang, 2006; Xu and Raginsky, 2017). Indeed $\widehat{p}_{S,\gamma}$ in (2) can be equivalently defined as the solution to the following convex optimization problem

$$\operatorname*{arg\,inf}_{\widehat{p} \in \mathcal{M}_1} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{w} \sim \widehat{p}} \left[\widehat{R}_S(\boldsymbol{w}) \right] + \frac{1}{\gamma} \mathrm{KL}\left(\widehat{p} \,||\, \mathcal{N}(\boldsymbol{0}, \lambda^{-1} \boldsymbol{I}) \right) \right\} , \tag{4}$$

where $\mathcal{M}_1$ is the set of all sample-dependent probability densities on $\mathbb{R}^d$, with sample drawn from $\mathcal{D}$, and KL-divergence is defined between densities that are absolutely continuous with respect to some measure over $\mathbb{R}^d$. Problem (4) can be also motivated from the perspective of PAC-Bayesian analysis (McAllester, 1998; Seeger, 2002), where (2) is the density minimizing the bound on the expected risk. Another instance of (4) is the well-known Maximum Entropy Discrimination framework of Jaakkola et al. (1999).

***Sampling from* (2).** Markov chain Monte Carlo (MCMC) algorithms can be used to sample directly from (2). Unfortunately, this is often known to be computationally inefficient (Andrieu et al., 2003). On the other hand, there is a number of cases where MCMC demonstrates amenable computational properties, for instance when sampling from log-concave densities (Andrieu et al., 2003). Recent works have also showed that for a particular class of densities (such as smooth and strongly concave densities) variations of MCMC algorithms can sometimes achieve linear convergence (Cheng et al., 2018b). For certain non-log-concave densities MCMC variants can achieve polynomial convergence in the dimension (Cheng et al., 2018a), or ever achieve faster convergence than optimization algorithms (Ma et al., 2018). Another popular line of research is a variational approximations of the Gibbs density, such as Variational Bayes (Wang and Blei, 2018) where one resorts to the variational approximation of a target density.

1.1. Our Contribution

The algorithms discussed above perform randomized empirical risk minimization. However, minimizing empirical risk does not always lead to minimization of the risk. Hence, the quality of the solution $A(S)$ generated by the randomized algorithm A given the training set S is typically analyzed through the notion of *excess risk* $\mathbb{E}_{S,A}\left[R(A(S))\right] - R(\boldsymbol{w}^\star)$, where $\boldsymbol{w}^\star$ is one of the minimizers of the risk. This is decomposed into the generalization error $R(A(S)) - \widehat{R}_S(A(S))$ and the term $\widehat{R}_S(A(S)) - R(\boldsymbol{w}^\star)$. Similarly to Raginsky et al. (2017), we follow instead a Gibbs-centric decomposition of the excess risk

$$\mathop{\mathbb{E}}_{S,A}\left[R(A(S))\right] - R(\boldsymbol{w}^\star) = \underbrace{\mathop{\mathbb{E}}_{S,A,\boldsymbol{w} \sim \widehat{p}_{S,\gamma}} \left[R(A(S)) - R(\boldsymbol{w})\right]}_{\text{Computational excess risk}} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \widehat{p}_{S,\gamma}} \left[R(\boldsymbol{w})\right] - R(\boldsymbol{w}^\star)}_{\text{Statistical excess risk}} .$$

The first term is due solely to the dynamics of the algorithm, be it SGLD or a sampling procedure, while the second one is a purely learning-theoretic quantity. Raginsky et al. (2017) mainly focused on the finite-time analysis of the first term for SGLD while showing convergence for non-convex objective functions. Their analysis of the second term —the statistical excess risk— provides a bound of order (ignoring logarithmic factors)

$$\frac{(\gamma+d)^2}{\lambda_\star m} + \frac{d}{\gamma} \tag{5}$$

where $\lambda_\star$ is a positive spectral gap characterizing the exponential convergence rate of the Langevin diffusion to the stationary point. They conservatively bounded the reciprocal of $\lambda_\star$ as

$$\frac{1}{\lambda_\star} = \tilde{\mathcal{O}}\left(\frac{1}{\gamma(d+\gamma)}\right) + \left(1 + \frac{d}{\gamma}\right) e^{\tilde{\mathcal{O}}(\gamma+d)} \ . \tag{6}$$

This results in a statistical excess risk bound with a rather pessimistic exponential dependence on the ambient dimension d. Therefore, a natural question to ask is whether the dependence on d can be improved by taking into account specific properties of the learning problem, and whether the dependence on $\lambda_\star$ can be avoided altogether (since the *statistical* excess risk does not really depend on the convergence properties of SGLD). We believe that $\lambda_\star$ (which has exponential dependence on the dimension) can be avoided in the analysis of Raginsky et al. (2017), although in their case this would not have improved the final result due to the contribution of the computational excess risk.
In this paper we consider the statistical excess risk, while we forego computational aspects of concrete algorithms. In particular, we focus on the *distribution-dependent* analysis of *statistical excess risk* (or, simply, excess risk). In the following we show upper bounds on the statistical excess risk that can be much smaller than (5) depending on the interplay between the data-generating distribution and the loss function.
The notion of excess risk considered in this paper is defined with respect to the *regularized* minimizer of the risk

$$\boldsymbol{w}_\lambda^\star \in \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \left\{R(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2\right\} \ . \tag{7}$$

Note that this is not a limitation because we can always recover the regularizer-free analysis by looking at the asymptotic behavior $\lambda \to 0$. In the following, we assume that both risk and empirical risk are bounded and twice differentiable, and the Hessian matrix of the risk is locally-Lipschitz (in a sense precisely defined later on). Therefore, the objective function of (7) (as well as the one of (1)) can have more than one minimum. However we assume that local minima are *isolated*, meaning that a sufficiently small neighborhood of $\boldsymbol{w}_\lambda^\star$ contains a unique minimum. One compelling example is a large family of non-convex *strict-saddle* ERM problems (Ge et al., 2015; Gonen and Shalev-Shwartz, 2017), such as matrix completion, tensor decomposition, PCA, ICA, and others. Another example of such is the empirical risk of a *ReLU neural network* with weight decay, or L2 regularization, (Milne, 2018, Theorem 1) where minima resulting in a sufficiently small empirical risk are locally strongly-convex. Even though the theorem holds for empirical measures, we suspect that it could be extended to the population risk through the uniform convergence argument.

Localized excess risk. Before delving into the global analysis of the excess risk, we look at the *local* approximation properties of the Gibbs-ERM principle, which will also be instrumental in the forthcoming global analysis. We begin by looking at the *localized* excess risk with respect to a fixed minimizer $\boldsymbol{w}_\lambda^\star$. Specifically, we consider the excess risk of a parameter $\boldsymbol{w}$ generated by Gibbs-ERM within a certain neighborhood around $\boldsymbol{w}_\lambda^\star$. This is defined as

$$\Delta(\boldsymbol{w}_\lambda^\star) = \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[R(\boldsymbol{w}) - R(\boldsymbol{w}_\lambda^\star) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right]\right] ,$$

where conditioning is taken on the event that $\boldsymbol{w}$ lies in the ellipsoid $\mathcal{E}^\star(r)$ of radius r centered at the minimizer $\boldsymbol{w}_\lambda^\star$ and aligned with the curvature of the risk at that minimum. We prove (Theorem 4) that the local excess risk behaves as[1]

$$\Delta(\boldsymbol{w}_\lambda^\star) \lesssim \frac{1}{\gamma}\operatorname{tr}\left(\boldsymbol{H}^\star\boldsymbol{H}_\lambda^{\star-1}\right) + \varepsilon(r) + \sqrt{\gamma\varepsilon(r)} + \frac{\gamma}{\sqrt{m}} \tag{8}$$

where $\boldsymbol{H}^\star$ is the Hessian matrix $\nabla^2 R(\boldsymbol{w}_\lambda^\star)$, $\boldsymbol{H}_\lambda^\star = \boldsymbol{H}^\star + 2\lambda\boldsymbol{I}$, and $\varepsilon(r)$ is a local approximation error that vanishes as $r \to 0$ (defined precisely in Section 4). The trace term in (8), a distribution-dependent quantity known as the *effective dimension* of minimizer $\boldsymbol{w}_\lambda^\star$, can be also expressed as $\lambda_1/(\lambda_1+\lambda) + \cdots + \lambda_d/(\lambda_d+\lambda)$ where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\boldsymbol{H}^\star$. This can be viewed as a "soft" version of the rank of $\boldsymbol{H}^\star$. Note that $\operatorname{tr}\left(\boldsymbol{H}^\star\boldsymbol{H}_\lambda^{\star-1}\right) \le d$ always, and $\operatorname{tr}\left(\boldsymbol{H}^\star\boldsymbol{H}_\lambda^{\star-1}\right) \ll d$ whenever the spectrum of the Hessian matrix is light-tailed. This notion of effective dimension occasionally appears in the analysis of ridge regression (Audibert and Catoni, 2011; Neu and Rosasco, 2018).

Next, to get a sense of the strength of the bound and as a sanity check, one may look at limiting cases with respect to parameters λ and γ. When $\lambda \to 0$, corresponding to the unregularized Gibbs-ERM principle, our bound becomes

$$\Delta(\boldsymbol{w}^\star) \lesssim \frac{1}{\gamma}\operatorname{rank}(\boldsymbol{H}^\star) + \operatorname{poly}(m, \gamma, r, \lambda_{\min}(\boldsymbol{H}^\star))$$

where $\lambda_{\min}(\boldsymbol{H}^\star)$ denotes the smallest non-zero eigenvalue of $\boldsymbol{H}^\star$. Assuming the radius is set to $r = \gamma^{-\frac{1+p}{3}}$ for some $p > 0$, the polynomial term in the right-hand side of the above bound vanishes as $\gamma, m \to \infty$, even for singular $\boldsymbol{H}^\star$. On the other hand, for $\lambda > 0$ and $\gamma, m \to \infty$ the right-hand side of (8) tends to zero, and the bound backs up the intuition that the Gibbs-ERM principle should exactly recover the ERM solution. This observation also serves as a sanity check that the bound is reasonably tight, at least with respect to γ.

Finally, in cases when the Hessian matrix of the risk is constant, for instance in Regularized Least Squares (RLS) problems, $\varepsilon(r) = 0$ and our bound specializes to

$$\Delta(\boldsymbol{w}_\lambda^\star) \lesssim \frac{1}{\gamma}\operatorname{tr}\left(\boldsymbol{H}^\star\boldsymbol{H}_\lambda^{\star-1}\right) + \frac{\gamma}{\sqrt{m}} .$$

When γ is tuned optimally the above bound becomes $\Delta(\boldsymbol{w}_\lambda^\star) \lesssim m^{-\frac{1}{4}}\sqrt{\operatorname{tr}\left(\boldsymbol{H}^\star\boldsymbol{H}_\lambda^{\star-1}\right)}$. Note that, for the square loss, the best known dependence on m that can be achieved is $m^{-\frac{1}{2}}$. The worse exponent in our bound is the price we pay for the generality of our approach. Although our results are never worse in terms of the dimension, the dependence on the sample size in our bounds is worse than in those of (Raginsky et al., 2017, (3.27)). This is because we are stating bounds in terms of the distribution-dependent effective dimension.

1. Throughout this paper, we use $f \lesssim g$ to say that there exists a universal constant $C > 0$ such that $f \le Cg$ holds uniformly over all arguments.

Global excess risk. Next, we consider a *global* notion of excess risk,

$$\Delta(\pi) = \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}[R(\boldsymbol{w})]\right] - \mathop{\mathbb{E}}_{I\sim\pi}\left[R(\boldsymbol{w}^\star_{\lambda,I})\right] .$$

Here, in the second term π is a distribution over the countable set $\mathcal{I}$ of all minima (recall that minima are isolated). In this setting, all minima, Hessian matrices, approximation errors, and ellipsoids gain a corresponding subscript $i \in \mathcal{I}$: $\boldsymbol{w}^\star_{\lambda,i}$, $\boldsymbol{H}^\star_{\lambda,i}$, $\varepsilon_i(r)$, and $\mathcal{E}^\star_i(r)$.
We first focus on the finite-temperature distribution over minima of the regularized risk

$$\pi_{\gamma,r}(i) = \frac{\mathbb{P}_\gamma(\mathcal{E}^\star_i(r))}{\sum_{j\in\mathcal{I}}\mathbb{P}_\gamma(\mathcal{E}^\star_j(r))} , \qquad i \in \mathcal{I} , \tag{9}$$

where probabilities are taken with respect to the population Gibbs density $p_\gamma(\boldsymbol{w}) \propto e^{-\gamma R_\lambda(\boldsymbol{w})}$. For this distribution we prove that

$$\Delta(\pi_{\gamma,r}) \lesssim \frac{1}{\gamma}\,\mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{H}^\star_I \boldsymbol{H}^{\star-1}_{\lambda,I}\right)\right] + \frac{\gamma}{\sqrt{m}} + \mathbb{E}[\varepsilon_I(r)] + \sqrt{\gamma\,\mathbb{E}[\varepsilon_I(r)]} + \mathbb{P}_\gamma(\mathcal{C}^\star(r)) , \tag{10}$$

for any $r < r_0$ where the radius r_0 is chosen such that all ellipsoids in the set $\left\{\mathcal{E}^\star_i(r) \,:\, i \in \mathcal{I}\right\}$ are disjoint, $\mathcal{C}^\star(r)$ is the *complement* of the union of the ellipsoids in this set (i.e., the volume outside of minima), and the expectation is taken with respect to $I \sim \pi_{\gamma,r}$.
Note that there is a trade-off in (10) between the first term, which is essentially a bound on the expected excess risk in the neighborhood of a minimum drawn according to $\pi_{\gamma,r}$, and the last term, which is the probability of sampling outside of the neighborhood of any minimum. This means that we can obtain an *oracle inequality* by choosing $r \in [0, r_0]$ such that it minimizes the right-hand side of (10).
Now we focus on the probability of the complement, which behaves as

$$\mathbb{P}_\gamma(\mathcal{C}^\star(r)) \le 1 - \left(1 - de^{-r^2\gamma\alpha_{d/2}}\right)\sum_{i\in\mathcal{I}} e^{-\frac{1}{3}\gamma\varepsilon_i(r)} , \tag{11}$$

where $\alpha_{d/2}$ depends only on d. So, as long as $r^2\gamma$ is increasing and $\gamma\varepsilon_i(r)$ is non-increasing, the probability of generating a solution outside of the minima decreases. For example, when $r = \gamma^{\frac{p-1}{2}}$ for $p \in (0, 1/3]$ (as discussed in Section 4.2) the right-hand side of (11) vanishes as $\gamma \to \infty$.

Asymptotic pseudo excess risk. It is also natural to ask what happens in the zero-temperature regime $\gamma \to \infty$, when the Gibbs-ERM principle reduces to a rule for selecting empirical risk minimizers. We can study this by observing that (10) vanishes when the radius is set to $r = \gamma^{\frac{p-1}{2}}$ —as we previously discussed— and γ is set to $m^{\frac{1}{4}}$, which is a meaningful result. Indeed, whenever $m = \infty$, then $\gamma = \infty$ and the risk of Gibbs-ERM should not differ from the risk of a minimum drawn from the limiting distribution $\pi_\infty = \lim_{\gamma\to\infty}\pi_{\gamma,r}$. Interestingly, the distribution π_∞ has the following analytic form (this is shown in Lemma 5 assuming the tuning $r = \gamma^{\frac{p-1}{2}}$):

$$\pi_\infty(i) = \frac{1}{\sum_{j\in\mathcal{I}^{\text{GLOB}}}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} , \qquad i \in \mathcal{I}^{\text{GLOB}} \tag{12}$$

where $\mathcal{I}^{\text{GLOB}}$ is a countable set enumerating global minima of the regularized risk. Hence, the probability of a minimum $\boldsymbol{w}^\star_{\lambda,i}$ is proportional to the reciprocal of the normalized volume of the ellipsoid defined by the eigenvalues of the Hessian at that minimum. In particular, this implies that the probability of choosing a global minimum with larger volume is higher. Note that all suboptimal minima have zero probability under π_∞. At the same time, in this asymptotic regime it is also rather clear that Gibbs-ERM generates models outside of the neighborhoods of the minima with zero probability. These two observations show how to strike a middle ground between the nonasymptotic bound of (10) and the asymptotic distribution (12). This is captured by the global asymptotic *pseudo* excess risk

$$\Delta_r^\infty = \mathop{\mathbb{E}}_{I\sim\pi_\infty}\left[\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[R(\boldsymbol{w}) \mid \boldsymbol{w}\in\mathcal{E}_I^\star(r)\right]\right] - R(\boldsymbol{w}^\star_{\lambda,I})\right]\,, \qquad r>0\,,$$

which bounds the localized excess risk at finite temperature $\gamma > 0$ when minima are drawn from the global limiting distribution π_∞. For any $r \ge 0$ we have

$$\Delta_r^\infty \lesssim \frac{1}{\gamma}\,\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{H}_I^\star \boldsymbol{H}_{\lambda,I}^{\star-1}\right)\right] + \mathbb{E}\left[\varepsilon_I(r)\right] + \sqrt{\gamma\,\mathbb{E}\left[\varepsilon_I(r)\right]} + \frac{\gamma}{\sqrt{m}}\,. \tag{13}$$

Observe that whereas the local excess risk (8) is essentially controlled by the soft rank of the minimum, the bound (13) implies that globally this is not necessarily the case, since low-rank minima have smaller probability under distribution π_∞.

1.2. Additional Related Work

Generalization bounds for the Gibbs-ERM principle have been extensively studied in a number of works over the past years. One prolific thread of research in this direction is PAC-Bayesian analysis, starting from the seminal works of McAllester (1998); Langford and Shawe-Taylor (2003) —see, for instance, (Germain et al., 2015; Grünwald and Mehta, 2017) for the latest developments. PAC-Bayesian analysis follows from uniform convergence arguments (over the class of densities), where —as pointed out earlier— Gibbs density is minimizing the bound on the risk (Information Risk Minimization). In this paper we focus on excess risk bounds (rather than generalization bounds) that manifest distribution-dependent properties of a potentially non-convex risk function. PAC-Bayesian excess risk bounds have also been studied in a number of contexts (Alquier et al., 2016; Audibert and Catoni, 2011; Grünwald and Mehta, 2017). However, these typically assume convex risk (e.g. least-squares), or focus on the properties of the hypothesis class (sometimes distribution-dependent) rather than those of the objective function (Grünwald and Mehta, 2017). Distribution-dependent arguments have been exploited to develop sharper generalization bounds (Lever et al., 2013), and data-dependent PAC-Bayesian bounds were shown to be numerically non-vacuous as shown by Dziugaite and Roy (2018).

A number of works have also analyzed generalization and approximation properties of Gibbs-ERM from the algorithmic point of view. A heuristic approach to the analysis of SGLD algorithm was given by Welling and Teh (2011); Mandt et al. (2016) while recent works have also argued that its generalization ability is controlled by the "width" (or the notion of pseudo-rank) at the minimum of the empirical risk (Keskar et al., 2017; Chaudhari and Soatto, 2018; Liang et al., 2019), which is reminiscent of the effective dimension studied in this paper. Mou et al. (2018) developed generalization bounds for SGLD from PAC-Bayesian and algorithmic stability point of view (Bousquet and Elisseeff, 2002). Apart from excess risk bounds, Raginsky et al. (2017) also showed Gibbs-ERM specific generalization bounds through the algorithmic stability framework. In this paper we also analyze generalization and stability of the Gibbs-ERM principle, however we present a simpler proof technique similar in spirit to Xu and Raginsky (2017). Finally, Sheth and Khardon (2017) analyzed a slightly different notion of excess risk bounds for variational inference assuming the use of latent Gaussian models (such as generalized linear models and Gaussian processes).

2. Preliminaries

Throughout this paper, we use $f \lesssim g$ to indicate that there exists a universal constant $C > 0$ such that $f \leq Cg$ holds uniformly over all arguments. Let $\mathcal{B}_r(\boldsymbol{z}) \subset \mathbb{R}^d$ be the ball of center $\boldsymbol{z}$ and radius $r > 0$ and let $\mathcal{B}_r = \mathcal{B}_r(\boldsymbol{0})$. Given a positive definite $d \times d$ matrix $\boldsymbol{M}$, define $\|\boldsymbol{x}\|^2_{\boldsymbol{M}} = \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}$ for $\boldsymbol{x} \in \mathbb{R}^d$. Then, for any positive semi-definite $d \times d$ matrix $\boldsymbol{A}$ and $r > 0$ the corresponding ellipsoid centered at $\boldsymbol{x}_0 \in \mathbb{R}^d$ is defined as $\mathcal{E}(\boldsymbol{x}_0, \boldsymbol{A}, r) \equiv \left\{\boldsymbol{x} \in \mathbb{R}^d \ : \ \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{A}} \leq r\right\}$.
If p and q are densities that are absolutely continuous with respect to a measure μ over $\mathbb{R}^d$, the Kullback-Liebler (KL) divergence between p and q is defined as

$$\mathrm{KL}(p, q) = \mathop{\mathbb{E}}_{\boldsymbol{w} \sim p}\left[\ln\big(p(\boldsymbol{w})\big) - \ln\big(q(\boldsymbol{w})\big)\right] .$$

3. Sketch of the Analysis

In this section we briefly explain the main arguments at the basis of our analysis. We start from the analysis of the localized excess risk, which is decomposed into the generalization error (i.e., the difference between the risk and empirical risk) and the gap between the empirical risk and the risk of the minimizer $\boldsymbol{w}^\star_\lambda$,

$$\begin{aligned}
\mathbb{E}\left[R(\boldsymbol{w}) - R(\boldsymbol{w}^\star_\lambda) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right] &= \mathbb{E}\left[R(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right] && (14)\\
&\quad + \mathbb{E}\left[\widehat{R}_S(\boldsymbol{w}) - R(\boldsymbol{w}^\star_\lambda) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right] && (15)
\end{aligned}$$

where the expectation is taken with respect to the empirical Gibbs density (2). This decomposition is similar to the one of (Raginsky et al., 2017, last two terms in (1.5)), however, in our case it is done with respect to $\boldsymbol{w}^\star_\lambda$ instead of $\boldsymbol{w}^\star$. For brevity we omit the indices corresponding to the minima, as in the localized setting we consider a local minimum at the time. The generalization error of Gibbs-ERM is captured by Theorem 1 below (whose proof can be found in Section A.2).

Theorem 1 (Generalization bound) *Consider any loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ that is σ-sub-Gaussian in the first argument with respect to the Gibbs density*

$$\widehat{p}_{S,\gamma}(\boldsymbol{w}) \propto e^{-\frac{\gamma}{m}\sum_{i=1}^m f(\boldsymbol{w}, z_i)} \qquad \gamma > 0 \tag{16}$$

conditioned on a measurable $A \subseteq \mathbb{R}^d$. Then the generalization error of Gibbs-ERM satisfies

$$\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \widehat{p}_{S,\gamma}}\left[R(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w}) \,\middle|\, \boldsymbol{w} \in A\right]\right] \leq \frac{4\sigma^2\gamma}{m} .$$

Assuming that f is everywhere bounded by M, Hoeffding Lemma implies the bound $\frac{M^2\gamma}{2m}$. Under the same boundedness assumption, a bound with similar rates was also shown by Xu and Raginsky (2017) within the mutual information framework. Our proof works by showing that the Gibbs density is on-average replace-one stable in the sense of (Shalev-Shwartz and Ben-David, 2014, Section 13) (we include it here for completeness).

The second quantity (15) is less straightforward to control in a distribution-dependent setting. Raginsky et al. (2017) give a global upper bound $\widehat{R}_S(\boldsymbol{w}) - R(\boldsymbol{w}_\lambda^\star) = \tilde{\mathcal{O}}\big(d/\gamma\big)$ by further decomposing (15) as follows

$$\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}[\widehat{R}_S(\boldsymbol{w})] - \min_{\boldsymbol{u}\in\mathbb{R}^d} \widehat{R}_S(\boldsymbol{u}) + \min_{\boldsymbol{u}\in\mathbb{R}^d} \widehat{R}_S(\boldsymbol{u}) - \mathop{\mathbb{E}}_{S}[\widehat{R}(\boldsymbol{w}_\lambda^\star)] \,.$$

The second term is bounded trivially, while the analysis of the first term follows the so called "almost ERM" argument. In other words, understanding how "close" the solutions generated by Gibbs-ERM are to the solutions of ERM.

In our distribution-dependent setting we follow a different route: consider the case of RLS, where the empirical risk is an average of square-regularized losses. A rather straightforward argument (based on Gaussian integration) gives the following *identity* in an "almost ERM" style:

$$\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[\widehat{R}_S(\boldsymbol{w})\right] - \min_{\boldsymbol{u}\in\mathbb{R}^d} \widehat{R}_S(\boldsymbol{u}) = \frac{1}{\gamma}\operatorname{tr}\left(\nabla^2\widehat{R}_S(\widehat{\boldsymbol{w}}_\lambda)\left(\nabla^2\widehat{R}_S(\widehat{\boldsymbol{w}}_\lambda) + \lambda\boldsymbol{I}\right)^{-1}\right)$$

where $\widehat{\boldsymbol{w}}_\lambda$ is a minimizer of the RLS problem. Observe that in the above identity we obtain an empirical counterpart of the effective dimension introduced in (8). Since our goal is a distribution-dependent result, one possibility is to consider the concentration of Hessian eigenvalues. However, we follow a more direct approach. As we said earlier, the "almost ERM" analysis of RLS is relatively easy using Gaussian integration (since $\widehat{p}_{S,\gamma}$ is a Gaussian density). Since we deal instead with general smooth densities, our idea is to quantify the gap between the density at hand and the Gaussian density. This is nicely handled by the *transportation lemma* (Boucheron et al., 2013, Lemma 4.18), characterizing the difference between the expectations of different densities in terms of the KL-divergence between them. As a comparison we choose the Gaussian density $q_\gamma \propto \exp\big(-\frac{\gamma}{2}\|\boldsymbol{w}-\boldsymbol{w}_\lambda^\star\|^2_{\boldsymbol{H}_\lambda^\star}\big)$, which only depends on distribution-related quantities. An "almost ERM" style analysis applied to q_γ will give us exactly a distribution-dependent effective dimension. So, the only issue is the actual gap between densities. By expanding the KL-divergence we observe that the critical terms are

$$\begin{aligned}
&-\gamma\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S,\gamma}}\left[\widehat{R}_{S,\lambda}(\boldsymbol{w}) \mid \boldsymbol{w}\in\mathcal{E}^\star(r)\right]\right] + \frac{\gamma}{2}\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S,\gamma}}\left[\|\boldsymbol{w}-\boldsymbol{w}_\lambda^\star\|^2_{\boldsymbol{H}_\lambda^\star} \mid \boldsymbol{w}\in\mathcal{E}^\star(r)\right]\right]\\
&\quad\lesssim -\gamma\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S,\gamma}}\left[R_\lambda(\boldsymbol{w}) \mid \boldsymbol{w}\in\mathcal{E}^\star(r)\right]\right] + \frac{\gamma^2}{m} + \frac{\gamma}{2}\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S,\gamma}}\left[\|\boldsymbol{w}-\boldsymbol{w}_\lambda^\star\|^2_{\boldsymbol{H}_\lambda^\star} \mid \boldsymbol{w}\in\mathcal{E}^\star(r)\right]\right] \qquad (17)\\
&\quad\lesssim \gamma\varepsilon(r) + \frac{\gamma^2}{m}\,. \qquad (18)
\end{aligned}$$

To obtain (17), instead of using concentration, we resort to the generalization bound of Theorem 1, whereas (18) is obtained by Taylor expansion of the regularized risk around its minimizer. This is formally shown in Lemma 3, while the decomposition (14) is bounded in Theorem 4. Hence, the gap is quantified by the approximation error at the radius r plus a sample-dependent term due to the use of empirical Gibbs density. These terms appear in excess risk bounds (8), (10), and (13).

3.1. Global Analysis

Starting from the conditional local excess risk in the form given by the left-hand side of (14), we analyze a notion of global risk by bounding the probability of individual ellipsoids (i.e., neighborhoods of minima) and that of the complement of their union. Since for $\gamma \to \infty$ the probability of a complement approaches zero (as discussed in Remark 7), in order to obtain an asymptotic bound it is enough to focus on the relative probability of ellipsoids. In Lemma 5 we derive upper and lower bounds on this probability via Laplace approximation (Lemma 12), and then analyze their limit for $\gamma \to \infty$. Combining the local excess risk bound and the bound on the asymptotic relative probability of ellipsoids allows us to control the asymptotic global pseudo excess risk (13) —see Corollary 1. Finally, using a nonasymptotic bound on the probability of a complement —see the proof of Theorem 6— we can apply the law of total expectation to get also a nonasymptotic bound on the global excess risk.

4. Main results

4.1. Local analysis

We first turn our attention to the local analysis considering a fixed minimizer[2]

$$\boldsymbol{w}_\lambda^\star \in \arg\min \left\{ R(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|^2 \right\} \ . \tag{19}$$

Specifically, we prove that the risk of Gibbs-ERM in a neighborhood of $\boldsymbol{w}_\lambda^\star$ is controlled by the *local* effective dimension $\operatorname{tr}\left(\boldsymbol{H}^\star \boldsymbol{H}_\lambda^{\star-1}\right)$, defined in terms of the Hessian $\boldsymbol{H}^\star = \nabla^2 R(\boldsymbol{w}_\lambda^\star)$ of the risk of the minimizer, where $\boldsymbol{H}_\lambda^\star = \boldsymbol{H}^\star + 2\lambda \boldsymbol{I}$. We require that Hessians do not change "too quickly" by assuming that Hessian of the risk is Lipschitz in an ellipsoid $\mathcal{E}^\star(r) = \mathcal{E}(\boldsymbol{w}_\lambda^\star, \boldsymbol{H}_\lambda^\star, r)$ centered at the minimizer and aligned with the local curvature. Formally the *local Lipschitzness* of the Hessian is defined as follows.

Definition 2 (Locally-Lipschitz Hessian) *The Hessian $\nabla^2 R$ is locally Lipschitz around a minimizer $\boldsymbol{w}_\lambda^\star$ if there exists a function $L^\star : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\|\nabla^2 R(\boldsymbol{w}_\lambda^\star) - \nabla^2 R(\boldsymbol{w})\|_2 \le L^\star(r) \|\boldsymbol{w}_\lambda^\star - \boldsymbol{w}\| \qquad \textit{for all} \quad \boldsymbol{w} \in \mathcal{E}^\star(r) \ . \tag{20}$$

Note that local Lipschitzness of the risk Hessian implies local Lipschitzness of the regularized risk Hessian for the same function $L^\star$.

Local Lipschitzness of Hessians plays an important role in bounding the gap between the Gibbs density and the Gaussian density, as discussed in Section 3. In particular, the approximation error introduced by taking a Taylor expansion of the regularized risk up to the third term is

$$\varepsilon(r) = L^\star(r) \left(\frac{r}{\sqrt{\lambda_{\min} + \lambda}} \right)^3 \qquad r \ge 0 \tag{21}$$

where $\lambda_{\min}$ is the smallest non-zero eigenvalue of $\boldsymbol{H}^\star$. Observe that $\varepsilon(r) = 0$ for any constant Hessian matrices (e.g., in the case of RLS). Lemma 3 below here establishes a result needed to prove our bound on the local excess risk.

2. We will drop subscript indexing of minima in this section.

Lemma 3 *Recall that* $\boldsymbol{H}^\star = \nabla^2 R(\boldsymbol{w}_\lambda^\star)$, $\boldsymbol{H}_\lambda^\star = \boldsymbol{H}^\star + 2\lambda \boldsymbol{I}$, *and* $\varepsilon(r)$ *is a local approximation error defined in* (21). *For any minimizer* $\boldsymbol{w}_\lambda^\star$ *of the regularized risk we have*

$$\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[\widehat{R}_S(\boldsymbol{w}) \,\middle|\, \boldsymbol{w} \in \mathcal{E}^\star(r)\right]\right] - R(\boldsymbol{w}_\lambda^\star) \le \frac{1}{\gamma}\operatorname{tr}\left(\boldsymbol{H}^\star \boldsymbol{H}_\lambda^{\star-1}\right) + \frac{\varepsilon(r)}{6} + \frac{M}{2}\sqrt{\frac{\gamma\varepsilon(r)}{3} + \frac{M^2\gamma^2}{2m}} \,.$$

(The proof of Lemma 3, along with that of all remaining statements in this section, can be found in Section A.3.) This lemma, combined with the bound on the generalization error (Theorem 1), gives the desired result.

Theorem 4 (Localized Excess Risk Bound) *Assume the same as in Lemma 3. Then,*

$$\Delta(\boldsymbol{w}_\lambda^\star) \le \frac{1}{\gamma}\operatorname{tr}\left(\boldsymbol{H}^\star \boldsymbol{H}_\lambda^{\star-1}\right) + \frac{\varepsilon(r)}{6} + \frac{M}{2}\sqrt{\frac{\gamma\varepsilon(r)}{3} + \frac{M^2\gamma^2}{2m}} + \frac{M^2\gamma}{2m} \,.$$

4.2. Global analysis

We now turn our attention to the global analysis of the excess risk. Since we deal with a countable set of local minima (indexed by $\mathcal{I}$), we add a subscript to all minima-dependent quantities, such as $\boldsymbol{w}_{\lambda,i}^\star$, $\boldsymbol{H}_{\lambda,i}^\star$, $\mathcal{E}_i^\star(r)$. In particular, the approximation error is now defined as

$$\varepsilon_i(r) = L_i^\star(r)\left(\frac{r}{\sqrt{\lambda_{\min,i} + \lambda}}\right)^3 \qquad r > 0, i \in \mathcal{I} \tag{22}$$

where $L_i^\star$ is the local Lipschitz constant with respect to the minimum $\boldsymbol{w}_{\lambda,i}^\star$, and $\lambda_{\min,i}$ is the smallest non-zero eigenvalue of the Hessian matrix $\boldsymbol{H}_{\lambda,i}^\star$.
Next, we introduce an important assumption on the geometry of the regularized risk around its minimizers.

Assumption 1 *All local minima* $\boldsymbol{w}_\lambda^\star \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} R_\lambda(\boldsymbol{w})$ *satisfy* $\nabla R_\lambda(\boldsymbol{w}_\lambda^\star) = \boldsymbol{0}$ *and are such that* $\nabla^2 R_\lambda(\boldsymbol{w}_\lambda^\star)$ *is positive definite. In other words all local minima are* isolated.

The above assumption implies that there exists a number $r_0 > 0$ such that

$$r_0 = \max\left\{r > 0 \;:\; \bigcap_{i\in\mathcal{I}} \mathcal{E}_i^\star(r) \equiv \varnothing\right\} \,.$$

In other words, ellipsoids centered at minimizers and aligned with the local curvature of R_λ are non-overlapping. In addition to the set $\mathcal{I}$, indexing minima of the regularized risk, let $\mathcal{I}^{\text{GLOB}} \subseteq \mathcal{I}$ index the global minima and denote its complement by $\mathcal{I}^{\text{SUBOPT}} \equiv \mathcal{I} \setminus \mathcal{I}^{\text{GLOB}}$. Finally, introduce the complement of the ellipsoids centered at the minima (later called, with some abuse of terminology, *complement of the minima*),

$$\mathcal{C}^\star(r) \equiv \mathbb{R}^d \setminus \bigcup_{i\in\mathcal{I}} \mathcal{E}_i^\star(r) \qquad r \le r_0 \,.$$

The first result in this section concerns the distribution of local minima. In particular, we give an upper bound on the relative probability $\pi_{\gamma,r}(i)$ of the i-th minimum, and then derive the analytic form of the asymptotic distribution π_∞.

Lemma 5 (Distribution of Minima) *For all $r > 0$,*

$$\pi_{\gamma,r}(i) \leq \frac{e^{\frac{\gamma}{3}\max_{k\in\mathcal{I}}\varepsilon_k(r)}}{\sum_{j\in\mathcal{I}} e^{\gamma\left(R_\lambda(\boldsymbol{w}^\star_{\lambda,i}) - R_\lambda(\boldsymbol{w}^\star_{\lambda,j})\right)}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} \qquad i \in \mathcal{I}\,.$$

Moreover, assuming without loss of generality that $R_\lambda(\boldsymbol{w}^\star_{\lambda,i}) = 0$ for all $i \in \mathcal{I}^{\text{GLOB}}$, and setting $r = \gamma^{-p}$ for $p > 0$, we have

$$\lim_{\gamma\to\infty} \pi_{\gamma,r}(i) = \begin{cases} \frac{1}{\sum_{j\in\mathcal{I}^{\text{GLOB}}}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} & i \in \mathcal{I}^{\text{GLOB}} \\ 0 & i \in \mathcal{I}^{\text{SUBOPT}}\,. \end{cases}$$

The non-asymptotic bound in Lemma (5) follows by upper and lower Taylor expansion around minima. The asymptotic result (a zero-temperature distribution) is a direct consequence of the non-asymptotic bound. We are now ready to state the main result of this section.

Theorem 6 (Global Excess Risk Bound) *Assume the same as in Lemma 3. Then for any $r \in [0, r_0]$ the global excess risk satisfies*

$$\Delta(\pi_{\gamma,r}) \lesssim \frac{1}{\gamma}\,\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{H}^\star_I \boldsymbol{H}^{\star-1}_{\lambda,I}\right)\right] + \frac{\gamma}{\sqrt{m}} + \mathbb{E}[\varepsilon_I(r)] + \sqrt{\gamma\,\mathbb{E}[\varepsilon_I(r)]} + \mathbb{P}_\gamma(\mathcal{C}^\star(r))$$

where the expectation is taken with respect to $I \sim \pi_{\gamma,r}$ and the probability of the complement of the minima is bounded as

$$\mathbb{P}_\gamma(\mathcal{C}^\star(r)) \leq 1 - \left(1 - de^{-r^2\gamma\alpha_{d/2}}\right)\sum_{i\in\mathcal{I}} e^{-\frac{1}{3}\gamma\varepsilon_i(r)} \tag{23}$$

with

$$\alpha_{d/2} = \begin{cases} 1 & d = 1 \\ \Gamma\left(1 + \frac{d}{2}\right)^{-\frac{2}{d}} & \textit{otherwise}\,. \end{cases} \tag{24}$$

Remark 7 *We compute the value of r approximately minimizing the right-hand side in Theorem 6. Note that the probability of the complement of the minima decreases if we ensure that $r^2\gamma$ increases in γ and $\gamma\varepsilon_i(r) \propto r^3\gamma$ is non-increasing. For instance we may set $r^2\gamma = \gamma^p$ for $p > 0$ so that $r^3\gamma = \gamma^{1+\frac{3}{2}(p-1)}$. Hence we require $1 + \frac{3}{2}(p-1) \leq 0$ which is satisfied for any $p \in (0, 1/3]$. This implies that when $r = \gamma^{\frac{p-1}{2}}$ and $p \in (0, 1/3]$ the probability of the complement of the minima and the approximation terms $\gamma\varepsilon_i(r), \varepsilon_i(r)$ all vanish as $\gamma \to \infty$.*

Finally, combining the localized excess risk bound in Theorem 4 with Lemma 5 allows us to prove the following result about the asymptotic pseudo excess risk.

Corollary 1 *Assume the same as in Lemma 3. Then, for any $r > 0$, the global asymptotic pseudo-excess risk satisfies*

$$\Delta^\infty_r \lesssim \frac{1}{\gamma}\,\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{H}^\star_I \boldsymbol{H}^{\star-1}_{\lambda,I}\right)\right] + \mathbb{E}\left[\varepsilon_I(r)\right] + \sqrt{\gamma\,\mathbb{E}\left[\varepsilon_I(r)\right] + \frac{\gamma^2}{m}} + \frac{\gamma}{m}$$

where I is distributed according to

$$\pi_\infty(i) = \frac{1}{\sum_{j\in\mathcal{I}^{\text{GLOB}}}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}}\,.$$

ACKNOWLEDGEMENTS

The authors would like to thank Olivier Bousquet, Sébastien Gerchinovitz, Abbas Mehrabian, and Massimiliano Pontil for stimulating discussions on this work.

References

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

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5): 2766–2794, 2011.

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

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations (ICLR)*, 2018.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Computational Learning Theory (COLT)*, 2018b.

G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *International Conference on Machine Learing (ICML)*, 2018.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Computational Learning Theory (COLT)*, pages 797–842, 2015.

P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J.-F. Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(1):787–860, 2015.

A. Gonen and S. Shalev-Shwartz. Fast rates for empirical risk minimization of strict saddle problems. In *Conference on Computational Learning Theory (COLT)*, 2017.

P. D. Grünwald and N. A. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. *arXiv preprint arXiv:1710.07732*, 2017.

T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In *Conference on Neural Information Processing Systems (NIPS)*, 1999.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017.

J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Conference on Neural Information Processing Systems (NIPS)*, 2003.

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

T. Liang, T. Poggio, A. Rakhlin, and J. Stokes. Fisher-Rao Metric, Geometry, and Complexity of Neural Networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learing (ICML)*, pages 354–363, 2016.

D. A. McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory (COLT)*, 1998.

T. Milne. Piecewise strong convexity of neural networks. *arXiv preprint arXiv:1810.12805*, 2018.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Computational Learning Theory (COLT)*, 2018.

G. Neu and L. Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference on Computational Learning Theory (COLT)*, 2018.

NIST. Incomplete Gamma functions. https://dlmf.nist.gov/8.10#Px1.p3, 2018. Accessed: 2018-12-12.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *Conference on Computational Learning Theory (COLT)*, 2017.

M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269, 2002.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

R. Sheth and R. Khardon. Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

G.M. Tallis. Elliptical and radial truncation in normal populations. *The Annals of Mathematical Statistics*, 34(3):940–944, 1963.

B. Tzen, T. Liang, and M. Raginsky. Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. *arXiv preprint arXiv:1802.06439*, 2018.

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, pages 1–85, 2018.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learing (ICML)*, 2011.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Conference on Neural Information Processing Systems (NIPS)*, pages 2524–2533, 2017.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

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

Appendix A. Proofs

A.1. Common Tools

We compute the Taylor expansion of $R_\lambda(\boldsymbol{w})$ for $\boldsymbol{w} \in \mathcal{E}(\boldsymbol{w}_\lambda^\star, \nabla^2 R_\lambda(\boldsymbol{w}_\lambda^\star), r)$, where $\boldsymbol{w}_\lambda^\star$ is a minimizer of the regularized risk,

$$\begin{aligned} R_\lambda(\boldsymbol{w}) &\geq R_\lambda(\boldsymbol{w}_\lambda^\star) + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_\lambda^\star\|_{\boldsymbol{H}_\lambda^\star}^2 - \frac{L^\star(r)}{6}\|\boldsymbol{w} - \boldsymbol{w}_\lambda^\star\|^3 \\ &\geq R_\lambda(\boldsymbol{w}_\lambda^\star) + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_\lambda^\star\|_{\boldsymbol{H}_\lambda^\star}^2 - \frac{L^\star(r)}{6}\left(\frac{r}{\sqrt{\lambda_{\min} + \lambda}}\right)^3 \qquad (25) \\ &= R_\lambda(\boldsymbol{w}_\lambda^\star) + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_\lambda^\star\|_{\boldsymbol{H}_\lambda^\star}^2 - \frac{1}{6}\varepsilon(r) \qquad (26) \end{aligned}$$

where $\varepsilon(r)$ is defined in (21), and (25) follows because $\sqrt{\lambda_{\min} + \lambda}\|\boldsymbol{w} - \boldsymbol{w}_\lambda^\star\| \leq r$ where $\lambda_{\min}$ is the smallest non-zero eigenvalue of $\nabla^2 R(\boldsymbol{w}_\lambda^\star)$. In a similar way we have the upper expansion

$$R_\lambda(\boldsymbol{w}) \leq R_\lambda(\boldsymbol{w}_\lambda^\star) + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_\lambda^\star\|_{\boldsymbol{H}_\lambda^\star}^2 + \frac{1}{6}\varepsilon(r)\ . \qquad (27)$$

We now introduce a crucial *transportation lemma* which is instrumental in the following proofs.

Lemma 8 ((Boucheron et al., 2013, Lemma 4.18)) *Let Z be a real-valued integrable random variable with distribution P such that*

$$\ln \mathbb{E}\left[e^{\alpha(Z - \mathbb{E}[Z])}\right] \leq \frac{\alpha^2\sigma^2}{2} \qquad \alpha > 0$$

for some $\sigma > 0$ and let Z' be another random variable with distribution Q. If Q is absolutely continuous with respect to P and such that $\mathrm{KL}(Q \mid\mid P) < \infty$, then $\mathbb{E}[Z'] - \mathbb{E}[Z] \leq \sqrt{2\sigma^2 \mathrm{KL}\,(Q \mid\mid P)}$.

Next, we prove a helpful lemma about the log-ratio of Gibbs integrals.

Lemma 9 *Let $f_A, f_B : \mathcal{X} \to \mathbb{R}$ such that*

$$Z_A = \int_{\mathcal{B}} e^{-\gamma f_A(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}$$

is finite for all $\gamma > 0, \mathcal{B} \subseteq \mathcal{X}$ and let

$$p_A(\boldsymbol{x}) = \frac{1}{Z_A} e^{-\gamma f_A(\boldsymbol{x})} \quad \gamma > 0, \ \boldsymbol{x} \in \mathcal{B}$$

where f_B is similarly defined. Whenever $Z_A > 0$ we have that

$$-\ln\left(\frac{Z_A}{Z_B}\right) \le \gamma \int_{\mathcal{B}} p_B(\boldsymbol{x}) \left(f_A(\boldsymbol{x}) - f_B(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \ .$$

Proof Observe that

$$\frac{Z_A}{Z_B} = \frac{\int_{\mathcal{B}} e^{-\gamma f_A(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}}{\int_{\mathcal{B}} e^{-\gamma f_B(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}} = \frac{\int_{\mathcal{B}} e^{-\gamma f_A(\boldsymbol{x})} e^{\gamma(f_B(\boldsymbol{x}) - f_B(\boldsymbol{x}))} \, \mathrm{d}\boldsymbol{x}}{\int_{\mathcal{B}} e^{-\gamma f_B(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}} = \int_{\mathcal{B}} p_B(\boldsymbol{x}) e^{\gamma(f_B(\boldsymbol{x}) - f_A(\boldsymbol{x}))} \, \mathrm{d}\boldsymbol{x} \ .$$

Since $-\ln()$ is a convex function, by Jensen's inequality we obtain the desired result. ■

A.2. Generalization Bound for Gibbs-ERM

We start by proving a generalization bound for Gibbs-ERM.

Theorem 1 (restated) Consider any loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ that is σ-sub-Gaussian in the first argument with respect to the Gibbs density

$$\widehat{p}_{S,\gamma}(\boldsymbol{w}) \propto e^{-\frac{\gamma}{m} \sum_{i=1}^m f(\boldsymbol{w}, z_i)} \qquad \gamma > 0$$

conditioned on a measurable $A \subseteq \mathbb{R}^d$. Then the generalization error of Gibbs-ERM satisfies

$$\mathop{\mathbb{E}}_{S} \left[\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \widehat{p}_{S,\gamma}} \left[R(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w}) \,\middle|\, \boldsymbol{w} \in A \right] \right] \le \frac{4\sigma^2 \gamma}{m} \ .$$

Proof Consider the training examples S drawn i.i.d. from $\mathcal{D}$ and, for $i = 1, \dots, m$, denote by $S^{(i)} = \{z_1, \dots, z_{i-1}, z, z_{i+1}, \dots, z_m\}$ a replace-one training data, where z is independently drawn from $\mathcal{D}$. Throughout the proof, we drop γ from the notation for the Gibbs density $\widehat{p}_{S,\gamma}$. Introduce the conditional Gibbs densities $\widehat{p}_{S|A}(\boldsymbol{w})$ and $\widehat{p}_{S^{(i)}|A}(\boldsymbol{w})$. We denote by $\mathbb{E}_{\widehat{p}_{S|A}}[\cdot]$ and $\mathbb{E}_{\widehat{p}_{S^{(i)}|A}}[\cdot]$ expectations with respect to $\widehat{p}_{S|A}$ and $\widehat{p}_{S^{(i)}|A}$. We start by rewriting the expected generalization error as

$$\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[R(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w}) \mid \boldsymbol{w} \in A\right] = \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}\left[R(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w})\right]\right] \tag{28}$$

$$= \mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z)]\right] - \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z_i)]\right] \tag{29}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z_i)]\right] - \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z_i)]\right]\right)$$

(switch z and z_i in the first term)

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z_i)]\right]\right) . \tag{30}$$

Now we bound (30) showing the average replace-one stability of Gibbs distribution. We use the transportation Lemma 8 with $Q = \widehat{p}_{S|A}$ and $P = \widehat{p}_{S^{(i)}|A}$ we get that

$$\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z_i)] \le \sqrt{2\sigma^2 \mathrm{KL}\left(\widehat{p}_{S^{(i)}|A} \,||\, \widehat{p}_{S|A}\right)} . \tag{31}$$

Next, we focus on KL-divergence,

$$\mathrm{KL}\left(\widehat{p}_{S^{(i)}|A} \,||\, \widehat{p}_{S|A}\right) = \gamma \mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}\left[\widehat{R}_S(\boldsymbol{w}) - \widehat{R}_{S^{(i)}}(\boldsymbol{w})\right] - \ln\left(\frac{Z_{S^{(i)}}}{Z_S}\frac{\mathbb{P}_{\widehat{p}_{S^{(i)}}}(A)}{\mathbb{P}_{\widehat{p}_S}(A)}\right) \tag{32}$$

$$= \gamma \mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}\left[\widehat{R}_S(\boldsymbol{w}) - \widehat{R}_{S^{(i)}}(\boldsymbol{w})\right] - \ln\left(\frac{\int_A e^{-\gamma \widehat{R}_{S^{(i)}}(\boldsymbol{w})}\,\mathrm{d}\boldsymbol{w}}{\int_A e^{-\gamma \widehat{R}_S(\boldsymbol{w})}\,\mathrm{d}\boldsymbol{w}}\right) \tag{33}$$

$$\le \gamma \mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}\left[\widehat{R}_S(\boldsymbol{w}) - \widehat{R}_{S^{(i)}}(\boldsymbol{w})\right] + \gamma \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}\left[\widehat{R}_{S^{(i)}}(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w})\right]$$

(by Lemma 9)

$$= \frac{\gamma}{m}\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z_i) - f(\boldsymbol{w}, z)] + \frac{\gamma}{m}\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z) - f(\boldsymbol{w}, z_i)] \tag{34}$$

$$= \frac{\gamma}{m}\left(\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z_i)]\right) \tag{35}$$

$$+ \frac{\gamma}{m}\left(\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z)] - \mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z)]\right) .$$

By taking expectation with respect to S and z on both sides, we get that the term (35) can be expressed as

$$\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z_i)]\right] = \mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w}, z)] - \mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w}, z)]\right] \tag{36}$$

where we could switch z_i and z on the right-hand side because their are both independently drawn from $\mathcal{D}$. Thus, the expectation of (31) with respect to S and z is upper-bounded as

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w},z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w},z_i)]\right] &\le \mathop{\mathbb{E}}_{S,z}\sqrt{2\sigma^2 \mathrm{KL}\left(\widehat{p}_{S^{(i)}|A} \,||\, \widehat{p}_{S|A}\right)} && (37)\\
&\le \sqrt{2\sigma^2 \mathop{\mathbb{E}}_{S,z}\left[\mathrm{KL}\left(\widehat{p}_{S^{(i)}|A} \,||\, \widehat{p}_{S|A}\right)\right]} && \text{(Jensen's inequality)}\\
&\le 2\sqrt{\frac{\sigma^2\gamma}{m}\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w},z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w},z_i)]\right]}\,. && (38)
\end{aligned}$$

Solving the above with respect to the term on the left-hand side we get that for any $i = 1, \ldots, m$,

$$\mathop{\mathbb{E}}_{S,z}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S^{(i)}|A}}[f(\boldsymbol{w},z_i)] - \mathop{\mathbb{E}}_{\widehat{p}_{S|A}}[f(\boldsymbol{w},z_i)]\right] \le \frac{4\sigma^2\gamma}{m}\,. \tag{39}$$

Substituting the above into (29) gives the desired generalization bound. ■

A.3. Localized Excess Risk Bounds

First, we prove a key lemma about the conditional expectation of quadratic forms.

Lemma 10 *Suppose that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{M})$. Then for the ellipsoid*

$$\mathcal{E}(r) \equiv \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \|\boldsymbol{x}\|_{\boldsymbol{M}^{-1}} \le r\right\} \qquad r > 0$$

and for any Positive Semi-Definite (PSD) $d \times d$ matrix $\boldsymbol{A}$ we have that

$$\mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \,\middle|\, \boldsymbol{x} \in \mathcal{E}(r)\right] = \frac{F_{d+2}(r^2)}{F_d(r^2)} \operatorname{tr}(\boldsymbol{A}\boldsymbol{M}) \le \operatorname{tr}(\boldsymbol{A}\boldsymbol{M})\,.$$

where F_k is the CDF of a $\mathcal{X}^2$-distribution with k degrees of freedom.
Moreover the above implies that

$$\lim_{r \to 0} \mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \,\middle|\, \boldsymbol{x} \in \mathcal{E}(r)\right] = 0\,. \tag{40}$$

Proof Observe that

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \,\middle|\, \boldsymbol{x} \in \mathcal{E}(r)\right] &= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\right) \,\middle|\, \boldsymbol{x} \in \mathcal{E}(r)\right] && (41)\\
&= \operatorname{tr}\left(\boldsymbol{A}\,\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top \,\middle|\, \boldsymbol{x} \in \mathcal{E}(r)\right]\right) && \text{(by linearity of trace)}\\
&= \operatorname{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{M}}\right) && (42)
\end{aligned}$$

where $\tilde{\boldsymbol{M}}$ is the covariance matrix of the Gaussian density $\mathcal{N}(\mathbf{0}, \boldsymbol{M})$ conditioned on $\mathcal{E}(r)$. Next, we apply a result about moments of multivariate Gaussian densities under elliptical truncation (Tallis, 1963, p. 941) to get that

$$\tilde{\boldsymbol{M}} = \frac{F_{d+2}(r^2) - F_{d+2}(0)}{F_d(r^2) - F_d(0)}\, \boldsymbol{M} = \frac{F_{d+2}(r^2)}{F_d(r^2)}\, \boldsymbol{M}\,. \tag{43}$$

This proves the first identity.
The inequality is proven by expanding $\tilde{\boldsymbol{M}}$ further in terms of the Gamma function $\Gamma(\cdot)$ and the incomplete Gamma function $\gamma(\cdot,\cdot)$:

$$\tilde{\boldsymbol{M}} = \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(1+\frac{d}{2}\right)} \frac{\gamma\left(1+\frac{d}{2}, \frac{r^2}{2}\right)}{\gamma\left(\frac{d}{2}, \frac{r^2}{2}\right)} \boldsymbol{M} = \frac{2}{d} \frac{\gamma\left(1+\frac{d}{2}, \frac{r^2}{2}\right)}{\gamma\left(\frac{d}{2}, \frac{r^2}{2}\right)} \boldsymbol{M} \preceq \boldsymbol{M}\,.$$

Finally, we look at the limit of the ratio in the right-hand side of (43) as $r \to 0$. By L'Hôpital's rule,

$$\begin{aligned}
\lim_{r\to 0} \frac{F_{d+2}(r^2)}{F_d(r^2)} &= \lim_{r\to 0} \frac{\mathcal{X}^2_{d+2}(r^2)}{\mathcal{X}^2_d(r^2)} \\
&= \lim_{r\to 0} \frac{r^d e^{-\frac{r^2}{2}}}{2^{1+\frac{d}{2}}\Gamma\left(1+\frac{d}{2}\right)} \frac{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}{r^{d-2}e^{-\frac{r^2}{2}}} \\
&= \lim_{r\to 0} \frac{r^d e^{-\frac{r^2}{2}}}{2^{1+\frac{d}{2}}\Gamma\left(1+\frac{d}{2}\right)} \frac{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}{r^{d-2}e^{-\frac{r^2}{2}}} \\
&= \lim_{r\to 0} \frac{r^2}{d} = 0
\end{aligned}$$

concluding the proof. ■

Recall that $\mathcal{E}^\star(r) \equiv \mathcal{E}(\boldsymbol{w}^\star_\lambda, \boldsymbol{H}^\star_\lambda, r)$ is the ellipsoid of radius r centered at $\boldsymbol{w}^\star_\lambda$.

Lemma 3 (restated) For any minimizer $\boldsymbol{w}^\star_\lambda$ of the regularized risk we have

$$\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[\widehat{R}_S(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right]\right] - R(\boldsymbol{w}^\star_\lambda) \le \frac{1}{\gamma}\operatorname{tr}\left(\boldsymbol{H}^\star \boldsymbol{H}^{\star -1}_\lambda\right) + \frac{\varepsilon(r)}{6} + \frac{M}{2}\sqrt{\frac{\gamma\varepsilon(r)}{3} + \frac{M^2\gamma^2}{2m}}\,.$$

Proof We abbreviate the regularized empirical risk by $\widehat{R}_{S,\lambda}(\boldsymbol{w}) = \widehat{R}_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2$ and recall that the regularized risk is denoted by $R_\lambda(\boldsymbol{w}) = R(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2$. Throughout the proof, we drop γ from the notation for the Gibbs densities $\widehat{p}_{S,\gamma}$. Let $\widehat{p}_{S|\mathcal{E}^\star}$ be the Gibbs density (2) conditioned on the ellipsoid $\mathcal{E}^\star(r)$. Similarly, let $q_{|\mathcal{E}^\star}$ be the the Gaussian density

$$q(\boldsymbol{w}) = \frac{1}{Z_q} e^{-\frac{\gamma}{2}\|\boldsymbol{w}-\widehat{\boldsymbol{w}}_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}} \qquad \boldsymbol{w} \in \mathbb{R}^d$$

conditioned on $\mathcal{E}^\star(r)$.

We begin by observing that $\widehat{R}_S$ is trivially $M^2/8$-sub-Gaussian since the loss function is bounded by M. Hence, by the transportation Lemma 8,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_S}\left[\widehat{R}_S(\boldsymbol{w})\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]-\mathop{\mathbb{E}}_{q}\left[\widehat{R}_S(\boldsymbol{w})\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]\right]\\
&=\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\widehat{p}_{S|\mathcal{E}^\star}}\left[\widehat{R}_S(\boldsymbol{w})\right]-\mathop{\mathbb{E}}_{q_{|\mathcal{E}^\star}}\left[\widehat{R}_S(\boldsymbol{w})\right]\right] &(44)\\
&\le\frac{M}{2}\mathop{\mathbb{E}}_{S}\left[\sqrt{\mathrm{KL}\left(\widehat{p}_{S|\mathcal{E}^\star}\,||\,q_{|\mathcal{E}^\star}\right)}\right]\le\frac{M}{2}\sqrt{\mathop{\mathbb{E}}_{S}\left[\mathrm{KL}\left(\widehat{p}_{S|\mathcal{E}^\star}\,||\,q_{|\mathcal{E}^\star}\right)\right]} &(45)
\end{aligned}$$

where the last inequality is obtained by Jensen's inequality. The KL term can be written as follows

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}\left[\mathrm{KL}\left(\widehat{p}_{S|\mathcal{E}^\star}\,||\,q_{|\mathcal{E}^\star}\right)\right]=&\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_{S|\mathcal{E}^\star}}\left[\ln\left(\frac{\widehat{p}_{S|\mathcal{E}^\star}(\boldsymbol{w})}{q_{|\mathcal{E}^\star}(\boldsymbol{w})}\right)\right]\\
=&\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[\ln\left(\widehat{p}_{S|\mathcal{E}^\star}(\boldsymbol{w})\right)\right]-\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[\ln\left(q_{|\mathcal{E}^\star}(\boldsymbol{w})\right)\right]\\
=&-\gamma\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[\widehat{R}_{S,\lambda}(\boldsymbol{w})\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]-\mathop{\mathbb{E}}_{S}\left[\ln(\mathbb{P}_{\widehat{p}_S}(\mathcal{E}^\star(r))\,Z_{\widehat{p}_S})\right]\\
&+\frac{\gamma}{2}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]+\mathop{\mathbb{E}}_{S}\left[\ln(\mathbb{P}_q(\mathcal{E}^\star(r))\,Z_q)\right]\ . \quad (46)
\end{aligned}$$

Now we relate the regularized empirical risk (46) to the regularized risk. By applying Theorem 1 with $A\equiv\mathcal{E}^\star(r)$ we get

$$\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[R_\lambda(\boldsymbol{w})-\widehat{R}_{S,\lambda}(\boldsymbol{w})\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]=\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[R(\boldsymbol{w})-\widehat{R}_S(\boldsymbol{w})\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]\le\frac{M^2\gamma}{2m}\ .$$

Using this result we can write

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}\left[\mathrm{KL}\left(\widehat{p}_{S|\mathcal{E}^\star}\,||\,q_{|\mathcal{E}^\star}\right)\right]\le&-\gamma\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[R_\lambda(\boldsymbol{w})\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]+\frac{M^2\gamma^2}{2m}\\
&+\frac{\gamma}{2}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\widehat{p}_S}\left[\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}\,\middle|\,\boldsymbol{w}\in\mathcal{E}^\star(r)\right]\\
&-\mathop{\mathbb{E}}_{S}\left[\ln\left(\frac{\mathbb{P}_{\widehat{p}_S}(\mathcal{E}^\star(r))\,Z_{\widehat{p}_S}}{\mathbb{P}_q(\mathcal{E}^\star(r))\,Z_q}\right)\right]\\
\le&-\gamma R^\star_\lambda+\frac{\gamma\varepsilon(r)}{6}+\frac{M^2\gamma^2}{2m} &(47)\\
&-\mathop{\mathbb{E}}_{S}\left[\ln\left(\frac{\mathbb{P}_{\widehat{p}_S}(\mathcal{E}^\star(r))\,Z_{\widehat{p}_S}}{\mathbb{P}_q(\mathcal{E}^\star(r))\,Z_q}\right)\right] &(48)
\end{aligned}$$

where (47) is obtained by applying the lower Taylor expansion (26) to $\boldsymbol{w} \in \mathcal{E}^\star(r)$. Now we bound the expected log-ratio term in (48) as

$$
\begin{aligned}
-\mathop{\mathbb{E}}_{S}\left[\ln\left(\frac{\mathbb{P}_{\widehat{p}_S}(\mathcal{E}^\star(r))Z_{\widehat{p}_S}}{\mathbb{P}_q(\mathcal{E}^\star(r))Z_q}\right)\right] &= -\mathop{\mathbb{E}}_{S}\left[\ln\left(\frac{\int_{\mathcal{E}^\star(r)} e^{-\gamma \widehat{R}_{S,\lambda}(\boldsymbol{w})}}{\int_{\mathcal{E}^\star(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}}}\right)\right] \qquad (49)\\
&\le \gamma \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{q}\left[\widehat{R}_{S,\lambda}(\boldsymbol{w}) - \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda} \mid \boldsymbol{w}\in\mathcal{E}^\star(r)\right] \\
&\qquad\qquad\qquad\qquad\text{(by Lemma 9)}\\
&= \gamma \mathop{\mathbb{E}}_{q}\left[R_\lambda(\boldsymbol{w}) - \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda} \mid \boldsymbol{w}\in\mathcal{E}^\star(r)\right] \\
&\qquad\qquad\qquad\qquad\text{(Since } \mathbb{E}_S[\widehat{R}_{S,\lambda}(\boldsymbol{w})] = R_\lambda(\boldsymbol{w}))\\
&\le \gamma R^\star_\lambda + \frac{\gamma\varepsilon(r)}{6}\,. \qquad (50)
\end{aligned}
$$

where the last inequality is derived from the upper Taylor expansion (27). Substituting the above into (48) gives

$$
\mathop{\mathbb{E}}_{S}\left[\mathrm{KL}\left(\widehat{p}_{S|\mathcal{E}^\star} \,||\, q_{|\mathcal{E}^\star}\right)\right] \le \frac{\gamma\varepsilon(r)}{3} + \frac{M^2\gamma^2}{2m}\,. \qquad (51)
$$

Now we go back to (44) and, using the upper Taylor expansion (27), we get

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{q_{|\mathcal{E}^\star}}\left[\widehat{R}_S(\boldsymbol{w})\right] &= \mathop{\mathbb{E}}_{q_{|\mathcal{E}^\star}}[R(\boldsymbol{w})] \qquad \text{(since } \mathbb{E}_S[\widehat{R}_S(\boldsymbol{w})] = R(\boldsymbol{w}))\\
&\le R(\boldsymbol{w}^\star_\lambda)\\
&+ \nabla R(\boldsymbol{w}^\star_\lambda)^\top\left(\mathop{\mathbb{E}}_{q}\left[\boldsymbol{w}-\boldsymbol{w}^\star_\lambda \,\middle|\, \boldsymbol{w}\in\mathcal{E}^\star(r)\right]\right) \qquad (52)\\
&+ \mathop{\mathbb{E}}_{q}\left[\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star} \,\middle|\, \boldsymbol{w}\in\mathcal{E}^\star(r)\right] + \frac{1}{6}L^\star(r)\mathop{\mathbb{E}}_{q}\left[\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^3 \,\middle|\, \boldsymbol{w}\in\mathcal{E}^\star(r)\right]\\
&\le R(\boldsymbol{w}^\star_\lambda)\\
&+ \mathop{\mathbb{E}}_{q}\left[\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star} \,\middle|\, \boldsymbol{w}\in\mathcal{E}^\star(r)\right] + \frac{\varepsilon(r)}{6}\,. \qquad (53)
\end{aligned}
$$

where (52) vanishes since the first moment of elliptically-truncated Gaussian is zero (Tallis, 1963). Finally, we bound the first term in (53) by invoking Lemma 10. By taking $\boldsymbol{M} = \gamma\boldsymbol{H}^\star_\lambda$, $\boldsymbol{A} = \boldsymbol{H}^\star$, and $\boldsymbol{x} = \boldsymbol{w} - \boldsymbol{w}^\star_\lambda$, and using Lemma 10 we get

$$
\mathcal{E}^\star(r) \equiv \left\{\boldsymbol{w}\in\mathbb{R}^d \;:\; \sqrt{\gamma(\boldsymbol{w}-\boldsymbol{w}^\star_\lambda)^\top\boldsymbol{H}^\star_\lambda(\boldsymbol{w}-\boldsymbol{w}^\star_\lambda)} \le r\right\}
$$

and

$$
\mathop{\mathbb{E}}_{q}\left[\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star} \,\middle|\, \boldsymbol{w}\in\mathcal{E}^\star(r)\right] \le \frac{1}{\gamma}\operatorname{tr}\left(\boldsymbol{H}^\star\boldsymbol{H}^{\star-1}_\lambda\right)\,.
$$

Now, combining these results with the bound on KL-divergence (51), and substituting into (45), gives the stated result. ■

Theorem 4 (restated) Assume the same as in Lemma 3. Then,

$$\Delta(\boldsymbol{w}_\lambda^\star) \le \frac{1}{\gamma} \operatorname{tr}\left(\boldsymbol{H}^\star \boldsymbol{H}_\lambda^{\star-1}\right) + \frac{\varepsilon(r)}{6} + \frac{M}{2}\sqrt{\frac{\gamma\varepsilon(r)}{3} + \frac{M^2\gamma^2}{2m}} + \frac{M^2\gamma}{2m} \,.$$

Proof From the definition of local generalization error,

$$\begin{aligned}
\Delta(\boldsymbol{w}_\lambda^\star) &= \mathbb{E}_S\left[\mathbb{E}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[R(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right]\right] - R(\boldsymbol{w}_\lambda^\star) \\
&= \mathbb{E}_S\left[\mathbb{E}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[R(\boldsymbol{w}) - \widehat{R}_S(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right]\right] \\
&+ \mathbb{E}_S\left[\mathbb{E}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[\widehat{R}_S(\boldsymbol{w}) \mid \boldsymbol{w} \in \mathcal{E}^\star(r)\right]\right] - R(\boldsymbol{w}_\lambda^\star) \\
&\le \frac{M^2\gamma}{2m} + \frac{1}{\gamma} \operatorname{tr}\left(\boldsymbol{H}^\star \boldsymbol{H}_\lambda^{\star-1}\right) + \frac{\varepsilon(r)}{6} + \frac{M}{2}\sqrt{\frac{\gamma\varepsilon(r)}{3} + \frac{M^2\gamma^2}{2m}}
\end{aligned}$$

where the last inequality is derived from Theorem 1 and Lemma 3. ■

A.4. Statements about Probability Mass of Ellipsoids

Before we prove our bound on the global excess risk, we introduce some necessary technical notions about the *regularized gamma function*, which can be interpreted as the probability of an Euclidean ball of radius z under a Gaussian density with covariance matrix $\boldsymbol{I}$.

Theorem 11 ((NIST, 2018, Regularized Gamma Function)) *Denote the regularized gamma function by*

$$P(a,z) = \frac{\Gamma(a) - \Gamma(a,z)}{\Gamma(a)}$$

where $\Gamma(a,z)$ is the upper incomplete Gamma function given by

$$\Gamma(a,z) = \int_z^\infty t^{a-1} e^{-t}\, \mathrm{d}t \,.$$

Then, for all $z \ge 0$ and $a > 0$,

$$\left(1 - e^{-\alpha_a z}\right)^a \le P(a,z) \tag{54}$$

where

$$\alpha_a = \begin{cases} 1 & 0 < a < 1 \\ \frac{1}{\Gamma(1+a)^{\frac{1}{a}}} & a > 1 \,. \end{cases} \tag{55}$$

with equality in (54) only when $a = 1$.

Proposition 1 (Truncated Gaussian Integrals) *For any $\gamma, r > 0$,*

$$\int_{\mathcal{B}(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|^2}\, \mathrm{d}\boldsymbol{u} = \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right)$$

where $\mathcal{B}(r)$ is the d-dimensional Euclidean ball. In addition, for any $d \times d$ semi-definite matrix $\boldsymbol{A}$,

$$\int_{\mathcal{E}(\mathbf{0},\boldsymbol{A},r)} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|_{\boldsymbol{A}}^2} \,\mathrm{d}\boldsymbol{u} = \frac{1}{\sqrt{\det(\boldsymbol{A})}} \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) .$$

Proof By the integration of radial functions

$$\begin{aligned}
\int_{\mathcal{B}(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|^2} \,\mathrm{d}\boldsymbol{u} &= 2\,\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^r e^{-\frac{\gamma}{2}x^2} x^{d-1} \,\mathrm{d}x \\
&= \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} \frac{\Gamma\left(\frac{d}{2}\right) - \Gamma\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \\
&= \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) .
\end{aligned}$$

In addition we have

$$\begin{aligned}
\int_{\mathcal{E}(\mathbf{0},\boldsymbol{A},r)} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|_{\boldsymbol{A}}^2} \,\mathrm{d}\boldsymbol{u} &= \int_{\|\boldsymbol{u}\|_{\boldsymbol{A}} \le r} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|_{\boldsymbol{A}}^2} \,\mathrm{d}\boldsymbol{u} \\
&= \frac{1}{\sqrt{\det(\boldsymbol{A})}} \int_{\|\boldsymbol{v}\| \le r} e^{-\frac{\gamma}{2}\|\boldsymbol{v}\|^2} \,\mathrm{d}\boldsymbol{v} \\
&= \frac{1}{\sqrt{\det(\boldsymbol{A})}} \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) .
\end{aligned}$$

where the third step is obtained through the change of variables $\mathrm{d}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{u} = \mathrm{d}\boldsymbol{v}$. ■

Recall that $\mathcal{E}^\star(r) \equiv \mathcal{E}(\boldsymbol{w}_\lambda^\star, \boldsymbol{H}_\lambda^\star, r)$, where $\boldsymbol{H}_\lambda^\star = \nabla^2 R_\lambda(\boldsymbol{w}_\lambda^\star)$.

Lemma 12 (Bounds on the Ellipsoid probability mass.) *Let $\boldsymbol{w}_\lambda^\star$ be any minimizer of R_λ. Then the following results hold for probabilities of ellipsoids under the density $e^{-\gamma R_\lambda(\boldsymbol{w})}/Z$,*

$$\begin{aligned}
\mathbb{P}\big(\mathcal{E}^\star(r)\big) &\le \frac{1}{Z}\, e^{-\gamma R_\lambda(\boldsymbol{w}_\lambda^\star) + \frac{\gamma}{6}\varepsilon(r)}\, \frac{1}{\sqrt{\det(\boldsymbol{H}_\lambda^\star)}} \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) \\
\mathbb{P}\big(\mathcal{E}^\star(r)\big) &\ge \frac{1}{Z}\, e^{-\gamma R_\lambda(\boldsymbol{w}_\lambda^\star) - \frac{\gamma}{6}\varepsilon(r)}\, \frac{1}{\sqrt{\det(\boldsymbol{H}_\lambda^\star)}} \left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) \\
\mathbb{P}\big(\mathcal{E}^\star(r)\big) &\ge e^{-\frac{\gamma}{3}\varepsilon(r)} P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) .
\end{aligned}$$

Proof By applying the lower Taylor expansion (26) in the exponent of the Gibbs density we get

$$\begin{aligned}
\mathbb{P}\big(\mathcal{E}^\star(r)\big) &= \frac{1}{Z}\int_{\mathcal{E}^\star(r)} e^{-\gamma R_\lambda(\boldsymbol{w})}\,\mathrm{d}\boldsymbol{w} \\
&\le \frac{1}{Z}\,e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)+\frac{\gamma}{6}\varepsilon(r)}\int_{\mathcal{E}^\star(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}}\,\mathrm{d}\boldsymbol{w} \\
&= \frac{1}{Z}\,e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)+\frac{\gamma}{6}\varepsilon(r)}\,\frac{1}{\sqrt{\det(\boldsymbol{H}^\star_\lambda)}}\int_{\mathcal{B}(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|^2}\,\mathrm{d}\boldsymbol{u} \qquad (56)\\
&= \frac{1}{Z}\,e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)+\frac{\gamma}{6}\varepsilon(r)}\,\frac{1}{\sqrt{\det(\boldsymbol{H}^\star_\lambda)}}\left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2},\frac{r^2\gamma}{2}\right) \qquad (57)
\end{aligned}$$

where (56) is obtained via the change of variables $\boldsymbol{u} = \boldsymbol{H}^{\star\frac{1}{2}}_\lambda(\boldsymbol{w}-\boldsymbol{w}^\star_\lambda)$ and (57) via Proposition 1. This shows the first result. The second result follows in a similar way exploiting the upper Taylor expansion (27),

$$\begin{aligned}
\mathbb{P}\big(\mathcal{E}^\star(r)\big) &\ge \frac{1}{Z}\,e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)-\frac{\gamma}{6}\varepsilon(r)}\int_{\mathcal{E}^\star(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}}\,\mathrm{d}\boldsymbol{w} \\
&= \frac{1}{Z}\,e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)-\frac{\gamma}{6}\varepsilon(r)}\,\frac{1}{\sqrt{\det(\boldsymbol{H}^\star_\lambda)}}\int_{\mathcal{B}(r)} e^{-\frac{\gamma}{2}\|\boldsymbol{u}\|^2}\,\mathrm{d}\boldsymbol{u} \\
&= \frac{1}{Z}\,e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)-\frac{\gamma}{6}\varepsilon(r)}\,\frac{1}{\sqrt{\det(\boldsymbol{H}^\star_\lambda)}}\left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}} P\left(\frac{d}{2},\frac{r^2\gamma}{2}\right)\,. \qquad (58)
\end{aligned}$$

Finally, we give a lower bound on the probability of $\mathcal{E}^\star(r)$. We start by upper bounding the normalization constant using the lower Taylor expansion (26),

$$\begin{aligned}
Z &= \int_{\mathbb{R}^d} e^{-\gamma R_\lambda(\boldsymbol{w})}\,\mathrm{d}\boldsymbol{w} \\
&\le e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)+\frac{\gamma}{6}\varepsilon(r)}\int_{\mathbb{R}^d} e^{-\frac{\gamma}{2}\|\boldsymbol{w}-\boldsymbol{w}^\star_\lambda\|^2_{\boldsymbol{H}^\star_\lambda}}\,\mathrm{d}\boldsymbol{w} \\
&\le e^{-\gamma R_\lambda(\boldsymbol{w}^\star_\lambda)+\frac{\gamma}{6}\varepsilon(r)}\,\frac{1}{\sqrt{\det(\boldsymbol{H}^\star_\lambda)}}\left(\frac{2\pi}{\gamma}\right)^{\frac{d}{2}}\,.
\end{aligned}$$

Combining the above with (58) gives $\mathbb{P}\big(\mathcal{E}^\star(r)\big) \ge e^{-\frac{\gamma}{3}\varepsilon(r)} P\left(\frac{d}{2},\frac{r^2\gamma}{2}\right)$ thus completing the proof. ∎

Lemma 5 (restated) For all $r > 0$,

$$\pi_{\gamma,r}(i) \le \frac{e^{\frac{\gamma}{3}\max_{k\in\mathcal{I}}\varepsilon_k(r)}}{\sum_{j\in\mathcal{I}} e^{\gamma\left(R_\lambda(\boldsymbol{w}^\star_{\lambda,i})-R_\lambda(\boldsymbol{w}^\star_{\lambda,j})\right)}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} \qquad i\in\mathcal{I}\,.$$

Moreover, assuming without loss of generality that $R_\lambda(\boldsymbol{w}^\star_{\lambda,i}) = 0$ for all $i \in \mathcal{I}^{\text{GLOB}}$, and setting $r = \gamma^{-p}$ for $p > 0$, we have

$$\lim_{\gamma\to\infty}\pi_{\gamma,r}(i) = \begin{cases} \dfrac{1}{\sum_{j\in\mathcal{I}^{\text{GLOB}}}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} & i\in\mathcal{I}^{\text{GLOB}} \\ 0 & i\in\mathcal{I}^{\text{SUBOPT}}\,. \end{cases}$$

Proof Throughout this proof we consider probabilities of ellipsoids under the density $e^{-\gamma R_\lambda(\boldsymbol{w})}/Z$, and we abbreviate $R^\star_{\lambda,i} = R_\lambda(\boldsymbol{w}^\star_{\lambda,i})$. Applying Lemma 12 with $\boldsymbol{w}^\star_\lambda = \boldsymbol{w}^\star_{\lambda,i}$ readily gives

$$
\begin{aligned}
\pi_{\gamma,r}(i) &= \frac{\mathbb{P}(\boldsymbol{w} \in \mathcal{E}^\star_i(r))}{\sum_{j\in\mathcal{I}} \mathbb{P}(\boldsymbol{w} \in \mathcal{E}^\star_j(r))} \\
&\leq \frac{e^{\frac{\gamma}{6}\varepsilon_i(r)}}{\sum_{j\in\mathcal{I}} e^{\gamma\left(R^\star_{\lambda,i}-R^\star_{\lambda,j}\right)-\frac{\gamma}{6}\varepsilon_j(r)}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} \\
&\leq \frac{e^{\frac{\gamma}{6}\max_{k\in\mathcal{I}}\varepsilon_k(r)}}{\sum_{j\in\mathcal{I}} e^{\gamma\left(R^\star_{\lambda,i}-R^\star_{\lambda,j}\right)-\frac{\gamma}{6}\max_{k\in\mathcal{I}}\varepsilon_k(r)}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} \\
&= \frac{e^{\frac{\gamma}{3}\max_{k\in\mathcal{I}}\varepsilon_k(r)}}{\sum_{j\in\mathcal{I}} e^{\gamma\left(R^\star_{\lambda,i}-R^\star_{\lambda,j}\right)}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} \,.
\end{aligned} \tag{59}
$$

This proves the first statement.

Now we look at the asymptotics of $\pi_{\gamma,r}(i)$ as $\gamma \to \infty$ assuming that $r = \gamma^{-p}$ for $p > 0$. First, observe that for any $i \in \mathcal{I}$

$$
\lim_{\gamma\to\infty} \varepsilon_i(\gamma^{-p}) = \lim_{\gamma\to\infty} L^\star_i(\gamma^{-p})\left(\frac{1}{\gamma^p\sqrt{\lambda_{\min,i}+\lambda}}\right)^3 = 0
$$

because $\lim_{\gamma\to\infty} L^\star_i(\gamma^{-p}) = \mathcal{O}(1)$ and $\lambda_{\min,i} + \lambda > 0$. Thus, the numerator of (59) approaches 1. Now, we turn our attention to the denominator. First, we consider global minimizers recalling our assumption that $R^\star_{\lambda,i} = 0$. Denoting $\delta_{i,j}(\gamma) = e^{\gamma\left(R^\star_{\lambda,i}-R^\star_{\lambda,j}\right)}$, we observe that for all $\gamma \geq 0$ and $i \in \mathcal{I}^{\text{GLOB}}$,

$$
\lim_{\gamma\to\infty} \delta_{i,j}(\gamma) = \begin{cases} 1 & j \in \mathcal{I}^{\text{GLOB}} \\ 0 & j \in \mathcal{I}^{\text{SUBOPT}} \end{cases}
$$

where the second case holds because the exponent in $\delta_{i,j}(\gamma)$ is negative. This implies

$$
\lim_{\gamma\to\infty} \pi_{\gamma,r}(i) \leq \frac{1}{\sum_{j\in\mathcal{I}^{\text{GLOB}}}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}} \qquad i \in \mathcal{I}^{\text{GLOB}}.
$$

Next, we consider the local minima, and observe that for all $\gamma \geq 0$ and $i \in \mathcal{I}^{\text{SUBOPT}}$,

$$
\lim_{\gamma\to\infty} \delta_{i,j}(\gamma) = \begin{cases} 0 & \text{if } R^\star_{\lambda,i} \leq R^\star_{\lambda,j} \\ \infty & \text{otherwise.} \end{cases}
$$

Therefore, for all $i \in \mathcal{I}^{\text{SUBOPT}}$, $\lim_{\gamma\to\infty} \pi_{\gamma,r}(i) = 0$. This proves the second statement and completes the proof. ■

A.5. Global Excess Risk Bounds

We first show a nonasymptotic (i.e., finite γ) global excess risk bound.

Theorem 6 (restated) Assume the same as in Lemma 3. Then for any $r \in [0, r_0]$ the global excess risk satisfies

$$\Delta(\pi_{\gamma,r}) \lesssim \frac{1}{\gamma}\,\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{H}_I^\star \boldsymbol{H}_{\lambda,I}^{\star -1}\right)\right] + \frac{\gamma}{\sqrt{m}} + \mathbb{E}[\varepsilon_I(r)] + \sqrt{\gamma\,\mathbb{E}[\varepsilon_I(r)]} + \mathbb{P}_\gamma(\mathcal{C}^\star(r))$$

where the expectation is taken with respect to $I \sim \pi_{\gamma,r}$ and the probability of the complement of the minima is bounded as

$$\mathbb{P}_\gamma(\mathcal{C}^\star(r)) \le 1 - \left(1 - d e^{-r^2\gamma\alpha_{d/2}}\right)\sum_{i\in\mathcal{I}} e^{-\frac{1}{3}\gamma\varepsilon_i(r)} \tag{60}$$

with $\alpha_{d/2}$ defined in (55).

Proof Denote the sample-dependent global excess risk by

$$\Delta_S(\pi_{\gamma,r}) = \mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}[R(\boldsymbol{w})] - \mathop{\mathbb{E}}_{I\sim\pi_{\gamma,r}}\left[R(\boldsymbol{w}_{\lambda,I}^\star)\right]$$

and let the probabilities $\mathbb{P}\big(\mathcal{E}_i^\star(r)\big)$ and $\mathbb{P}\big(\mathcal{C}^\star(r)\big)$ be defined with respect to the population Gibbs distribution $p_\gamma(\boldsymbol{w}) \propto e^{-\gamma R_\lambda(\boldsymbol{w})}$ with $\gamma > 0$.

We first focus on the first term on the right-hand side of $\Delta_S(\pi_{\gamma,r})$. By the law of total expectation, for any $r \in [0, r_0]$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}[R(\boldsymbol{w})] &= \sum_{i\in\mathcal{I}} \mathbb{P}\big(\mathcal{E}_i^\star(r)\big)\,\mathbb{E}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{E}_i^\star(r)\right] + \mathbb{P}\big(\mathcal{C}^\star(r)\big)\,\mathbb{E}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{C}^\star(r)\right] \\
&\le \sum_{i\in\mathcal{I}} \frac{\mathbb{P}\big(\mathcal{E}_i^\star(r)\big)}{\sum_{j\in\mathcal{I}}\mathbb{P}\big(\mathcal{E}_j^\star(r)\big)}\,\mathbb{E}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{E}_i^\star(r)\right] \\
&\qquad\qquad\qquad \text{(ellipsoids are disjoint by Assumption 1)} \\
&\quad + \mathbb{P}\big(\mathcal{C}^\star(r)\big) M \qquad\qquad \text{(risk is bounded)} \\
&= \sum_{i\in\mathcal{I}} \pi_{\gamma,r}(I=i)\,\mathbb{E}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{E}_i^\star(r)\right] \qquad \text{(by definition of } \pi_{\gamma,r}\text{)} \\
&\quad + \mathbb{P}\big(\mathcal{C}^\star(r)\big) M\,.
\end{aligned} \tag{61}$$

An upper bound on $\mathbb{E}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{E}_i^\star(r)\right]$ is given by Theorem 4, thus all that is left to show is that the probability of the complement is small. Since the ellipsoids are disjoint

$$\mathbb{P}\big(\mathcal{C}^\star(r)\big) = 1 - \sum_{i\in\mathcal{I}} \mathbb{P}\big(\mathcal{E}_i^\star(r)\big)\,.$$

To upper bound $\mathbb{P}\big(\mathcal{C}^\star(r)\big)$ we need a lower bound on $\mathbb{P}\big(\mathcal{E}^\star(r)\big)$. This is provided by the last inequality in Lemma 12, that is,

$$\begin{aligned}\sum_{i\in\mathcal{I}} \mathbb{P}\big(\mathcal{E}_i^\star(r)\big) &\geq P\left(\frac{d}{2}, \frac{r^2\gamma}{2}\right) \sum_{i\in\mathcal{I}} e^{-\frac{\gamma\varepsilon_i(r)}{3}} \\ &\geq \left(1 - e^{-r^2\gamma\alpha_{d/2}}\right)^{\frac{d}{2}} \sum_{i\in\mathcal{I}} e^{-\frac{\gamma\varepsilon_i(r)}{3}} \qquad (62)\\ &\geq \left(1 - de^{-r^2\gamma\alpha_{d/2}}\right) \sum_{i\in\mathcal{I}} e^{-\frac{\gamma\varepsilon_i(r)}{3}}\end{aligned}$$

where (62) is derived from the lower bound on the regularized Gamma function (54), and the last inequality is obtained from the Bernoulli inequality

$$(1+x)^{\frac{d}{2}} \geq (1+x)^d \geq 1 + dx \qquad d \in \mathbb{N},\ x \geq -1\ .$$

Thus,

$$\mathbb{P}\big(\mathcal{C}^\star(r)\big) \leq 1 - \left(1 - de^{-r^2\gamma\alpha_{d/2}}\right) \sum_{i\in\mathcal{I}} e^{-\frac{1}{3}\gamma\varepsilon_i(r)}\ . \qquad (63)$$

Taking expectation with respect to S, and combining Theorem 4 with (61) and Jensen's inequality, we obtain

$$\begin{aligned}\Delta(\pi_{\gamma,r}) &\leq \mathop{\mathbb{E}}_{I\sim\pi_{\gamma,r}}\left[\mathop{\mathbb{E}}_{S}\left[\mathbb{E}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{E}_i^\star(r)\right]\right] - R(\boldsymbol{w}^\star_{\lambda,I})\right] + M\,\mathbb{P}\big(\mathcal{C}^\star(r)\big) \\ &\leq \frac{1}{\gamma}\mathop{\mathbb{E}}_{I\sim\pi_{\gamma,r}}\left[\operatorname{tr}\left(\boldsymbol{H}_I^\star \boldsymbol{H}_{\lambda,I}^{\star-1}\right)\right] + \frac{1}{6}\mathop{\mathbb{E}}_{I\sim\pi_{\gamma,r}}[\varepsilon_I(r)] + \frac{M}{2}\sqrt{\frac{\gamma}{3}\mathop{\mathbb{E}}_{I\sim\pi_{\gamma,r}}[\varepsilon_I(r)] + \frac{M^2\gamma^2}{2m}} + \frac{M^2\gamma}{2m} \\ &\quad + M\,\mathbb{P}\big(\mathcal{C}^\star(r)\big)\ .\end{aligned}$$

The proof is concluded by stating the above with respect to radius $r \in [0, r_0]$ —recall that the radius cannot exceed r_0, the largest radius ensuring that ellipsoids remain disjoint. ■

Corollary 1 (restated) Assume the same as in Lemma 3. Then, for any $r > 0$, the global asymptotic pseudo-excess risk satisfies

$$\Delta_r^\infty \lesssim \frac{1}{\gamma}\,\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{H}_I^\star \boldsymbol{H}_{\lambda,I}^{\star-1}\right)\right] + \mathbb{E}\left[\varepsilon_I(r)\right] + \sqrt{\gamma\,\mathbb{E}\left[\varepsilon_I(r)\right] + \frac{\gamma^2}{m}} + \frac{\gamma}{m}$$

where I is distributed according to

$$\pi_\infty(i) = \frac{1}{\sum_{j\in\mathcal{I}^{\text{GLOB}}}\sqrt{\frac{\det(\boldsymbol{H}^\star_{\lambda,i})}{\det(\boldsymbol{H}^\star_{\lambda,j})}}}\ .$$

Proof Recall that the global asymptotic pseudo-excess risk is defined as

$$\Delta_r^\infty = \mathop{\mathbb{E}}_{I\sim\pi_\infty}\left[\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\widehat{p}_{S,\gamma}}\left[R(\boldsymbol{w}) \,\middle|\, \boldsymbol{w}\in\mathcal{E}_I^\star(r)\right]\right] - R(\boldsymbol{w}^\star_{\lambda,I})\right]\ .$$

Distribution π_∞ is given by Lemma 5, while the local excess risk centered at $\boldsymbol{w}^\star_{\lambda,I}$ is bounded by Theorem 4. This immediately yields the statement. ■