2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–18

Pseudo-Bayesian Learning via Direct Loss Minimization with Applications to Sparse Gaussian Process Models

Rishit Sheth RISHET@MICROSOFT.COM
Microsoft Research

Roni Khardon RKHARDON@IU.EDU
Indiana University, Bloomington

Abstract

We propose that approximate Bayesian algorithms should optimize a new criterion, directly derived from the loss, to calculate their approximate posterior which we refer to as pseudo-posterior. Unlike standard variational inference which optimizes a lower bound on the log marginal likelihood, the new algorithms can be analyzed to provide loss guarantees on the predictions with the pseudo-posterior. Our criterion can be used to derive new sparse Gaussian process algorithms that have error guarantees applicable to various likelihoods.

1. Introduction

Results in learning theory show that, under some general conditions, minimizing training set loss, also known as empirical risk minimization (ERM), provides good solutions in the sense that the true loss of such procedures is bounded relative to the best loss possible in hindsight. Alternative algorithms such as structural risk minimization or regularized loss minimization (RLM) have similar guarantees under more general conditions. On the other hand, Bayesian approaches are, in a sense, prescriptive. Given prior and data, we calculate a posterior distribution that compactly captures all our knowledge about the problem. Then, given a prediction task with an associated loss for wrong predictions, we pick the best prediction given our posterior. This is optimal when the model is correct and the exact posterior is tractable. However, the algorithmic choices are less clear with misspecified models or, even if the model is correct, when exact inference is not possible and the learning algorithm can only return an approximation to the posterior. Since the choices are often heuristically motivated we call such approximations *pseudo-posteriors*. The question is how the pseudo-posterior should be calculated. In this paper we propose to use learning theory to guide this process.

To motivate our approach consider the variational approximation which is one of the most effective methods for approximate inference in Bayesian models. In lieu of finding the exact posterior, variational inference maximizes the ELBO, a lower bound on the marginal likelihood. It is well known that this can be seen alternatively as performing regularized loss minimization. For example, in a model with parameters w, prior $p(w)$, and data y where $p(y|w,x) = \prod_i p(y_i|w,x_i)$, we have

$$\log p(y) \geq \text{ELBO} \triangleq \mathop{\mathrm{E}}_{q(w)}[\log p(y|w)] - d_{\mathrm{KL}}(q(w),p(w)) = \sum_i \mathop{\mathrm{E}}_{q(w)}[\log p(y_i|w)] - d_{\mathrm{KL}}(q(w),p(w))$$

© R. Sheth & R. Khardon.

where $q(w)$ is the variational posterior and we have suppressed the dependence on x for visual clarity. Minimizing the negative ELBO, we have a loss term $\sum_i \mathrm{E}_{q(w)}[-\log p(y_i|w, x_i)]$ and a regularization term $d_{\mathrm{KL}}(q(w), p(w))$. The RLM viewpoint is attractive from the perspective of statistical learning theory because such algorithms are known to have good generalization guarantees (under some conditions). However, the ELBO objective is not matched to the intended use of Bayesian predictors: given a posterior $q(w)$ and test example x_*, the Bayesian predictor first calculates the predictive distribution $p(y_*|x_*) = \mathrm{E}_{q(w)}[p(y_*|x_*, w)]$ and then, assuming we are interested in the *log loss*, suffers the loss $-\log p(y_*|x_*)$. In other words, seen from the perspective of learning theory, variational inference optimizes for $L_{\mathrm{G}} = \sum_i \mathrm{E}_{q(w)}[-\log p(y_i|w)]$, sometimes known as the Gibbs loss, instead of $L_{\mathrm{B}} = \sum_i -\log \mathrm{E}_{q(w)}[p(y_i|x_i, w)]$, which is the loss of the Bayesian predictor.

These observations immediately raise several questions: Should we design empirical risk minimization (ERM) algorithms minimizing L_{B} that produce pseudo-posteriors? Should a regularization term, e.g., d_{KL}, be added? Can we use standard analysis, that typically handles frequentist models, to provide guarantees for such algorithms? We emphasize that this differs from standard non-Bayesian algorithms that perform ERM or RLM to find the best parameter w. Here, we propose to perform ERM or RLM to find the best pseudo-posterior $q(w)$ as given by the parameters that define it.

In this paper, we show that such an analysis can indeed be performed, and provide results which are generally applicable to Bayesian predictors optimized using ERM. Then, we focus on sparse Gaussian processes (sGP) for which we develop risk bounds for a smoothed variant of log loss[1] and any observation likelihood (the non-conjugate case). The significance of this is conceptual, in that it points to a different principle for designing approximate inference algorithms where we no longer aim to optimize the marginal likelihood (or ELBO), but instead a criterion that is directly related to the loss — this diverges from current practice in the literature.

The paper highlights sparse GP because it is an important model with significant recent interest and work. But the approach and results are more generally applicable. To illustrate this point the appendix shows how the results can be applied to the Correlated Topic Model (CTM) of Blei and Lafferty (2006).

It is important to distinguish this work from two previous lines of work. Our earlier work (Sheth and Khardon, 2017) made similar observations w.r.t. the mismatch between the optimization criterion and the intended objective. However, the goal there was to analyze existing algorithms where possible. More concretely we showed that optimizing a criterion related to L_{G} does have some risk guarantees, though these are weaker than the ones in this paper. Here, we propose to explore new algorithms based on direct loss minimization with stronger associated guarantees. In Alaoui and Mahoney (2015) and Burt et al. (2019), the goal is to show that the sparse GP approximation can be chosen to be very close to the full GP solution. Conditions on the kernel functions and on the algorithm to select inducing input locations and variational distribution are given for this to be true. This is a very strong result showing that nothing is lost by using the sparse approximation. However, in many cases, the number of inducing inputs required is too large (e.g., for Matern kernels).

1. For technical reasons, our results hold for a smoothed variant of log loss which is a limitation. As discussed below, it may be possible to remove this restriction with an alternative bound on $\Psi()$.

In contrast, our analysis aims at identifying the best sGP posterior in terms of the resulting prediction performance, whether it is close to the full GP posterior or not. In other words, we seek an "agnostic PAC guarantee" for the sparse GP posterior.

2. Technical Results

Due to space constraints, the main paper sketches the technical results with full details given in Appendices A to E. In short, three different approaches to proving agnostic PAC guarantees for learning with a Lipschitz loss under a bounded hypothesis space are provided. The three results use slightly different variants of ERM as the optimization algorithm. All three provide bounds if, in addition, the loss itself is bounded. Approach 1 (Appendix A) uses this directly and proves bounds using a standard discretization argument. Approach 2 (Appendix B) requires a bounded loss but adapts results based on Rademacher complexity (Meir and Zhang, 2003) to provide risk bounds that do not depend on the dimension of the hypothesis space and, in this way, potentially improves on approach 1. Approach 3 (Appendix C), which we present below, is new and has the potential to provide bounds with unbounded losses, although, for the application in this paper, we will be using bounded loss functions. We stress, though, that any of these approaches can be utilized to obtain guarantees under a Lipschitz loss and bounded hypothesis space. Appendices D and E develop the details for sGP and CTM.

2.1. Agnostic Learning with Randomized ERM

In the following, we consider a loss $\ell : \Theta \times (X, Y) \mapsto \mathbb{R}$ over a hypothesis space $\Theta \subset \mathbb{R}^M$ and example/label spaces X and Y. We assume that the hypothesis space is closed and bounded w.r.t. infinity norm with $\sup_{\theta\in\Theta} ||\theta||_\infty \leq B$. We further assume that ℓ is L-Lipschitz in its first argument w.r.t. the same norm, i.e., $\forall \theta, \theta' \in \Theta, |\ell(\theta) - \ell(\theta')| \leq L||\theta - \theta'||$. Let S denote an i.i.d. sample $\{(x_i, y_i)\}_{i=1}^n$ from an unknown distribution $\mathcal{D}$ over $X \times Y$.

The **Randomized ERM Algorithm** is as follows

$$\bar{\theta}_{\text{ERM}} \triangleq \arg\min_{\bar{\theta}\in\bar{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathrm{E}}_{q_{\text{jit}}(\theta|\bar{\theta})} \Big[\ell(\theta, (x_i, y_i))\Big],$$

where $\bar{\Theta} = [-B + \frac{M}{3L\lambda}, B - \frac{M}{3L\lambda}]^M$, $q_{\text{jit}}(\theta|\bar{\theta}) = \prod_{m=1}^{M} \mathcal{U}(\theta_m|\bar{\theta}_m - \frac{M}{3L\lambda}, \bar{\theta}_m + \frac{M}{3L\lambda})$, $\mathcal{U}$ denotes the uniform distribution, and $\lambda > 0$ is a scalar. The algorithm averages[2] the ERM objective of random neighbors of the solution $\bar{\theta}$. We have:

Theorem 1 *Let $p(\theta)$ be some sample-independent distribution over Θ. For all $\theta \in \Theta$,*

$$\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n} \left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\bar{\theta}_{ERM}, (x, y)) \right] \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta, (x, y)) + \frac{1}{\lambda} \left[M + M \log\left(\frac{3BL\lambda}{M}\right) + \Psi(\lambda, n) \right] \tag{1}$$

where

$$\Psi(\lambda, n) \triangleq \log \mathop{\mathrm{E}}_{S\sim\mathcal{D}^n} \left[\mathop{\mathrm{E}}_{p(\theta)} \exp\left(\lambda\Big(\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \Big[\ell(\theta, (x, y))\Big] - \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, (x_i, y_i))\Big)\right)\right].$$

2. Given the other approaches described in the appendix, it is reasonable to consider this an artifact of the proof. In this case, ERM may be used directly.

The proof (Appendix C) uses the compression lemma, $\mathrm{E}_{q(\theta)} f(\theta) \leq \mathrm{KL}(q(\theta), p(\theta)) + \log \mathrm{E}_{p(\theta)} e^{f(\theta)}$, but applied to the variational parameters θ in contrast with Germain et al. (2016) and Sheth and Khardon (2017) that applied it on w. This new approach is the source of jitter in the randomized ERM objective. Specifically we apply the compression lemma with $q(\theta) = q_{\text{jit}}(\theta|\theta_{\text{ERM}})$ and $f(\theta) = \lambda[\mathrm{E}_{(x,y)\sim\mathcal{D}} \ell(\theta, (x, y)) - \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, (x_i, y_i))]$. This bounds the potential overfitting, expressed by $\frac{1}{\lambda} f(\theta)$, by a KL term that we can compute explicitly and $\log \mathrm{E}_{p(\theta)} e^{f(\theta)}$ which results in Ψ.

If the loss is bounded, i.e., $|\ell| \leq c$, then, $\Psi(\lambda, n) \leq \frac{2\lambda^2 c^2}{n}$ (see Germain et al. (2016); Sheth and Khardon (2017)) implying the following corollary showing that the expected risk of Randomized ERM is bounded by the risk of any posterior in Θ plus a term that decays at a rate of $1/\sqrt{n}$. $\Psi(\lambda, n)$ can be bounded under some conditions even if the loss is not bounded[3] but we leave further exploration of this for future work.

Corollary 2 *If the loss is bounded, i.e., $|\ell| \leq c$, then using $\lambda = \sqrt{n}$ we have*

$$\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\bar{\theta}_{ERM}, (x, y))\right] \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta, (x, y)) + \frac{1}{\sqrt{n}}\left[M + M\log\left(\frac{3BL\sqrt{n}}{M}\right) + 2c^2\right]. \quad (2)$$

2.2. Applications to Sparse GP

In the (zero-mean) sparse GP model of Titsias (2009), w represents the latent function at the M inducing inputs $U = (u_1, u_2, \ldots, u_M)$, $f(x)$ is the latent function at x, $K(\cdot,\cdot)$ is the covariance function, and $p(f|w) = \mathcal{N}(f|K_{XU}K_{UU}^{-1}w, K_{XX} - K_{XU}K_{UU}^{-1}K_{UX})$ where $(K_{UU})_{kl} \triangleq K(u_k, u_l)$, $(K_{XU})_{ik} \triangleq K(x_i, u_k)$, and $K_{XU} \triangleq K_{UX}^\top$ for $1 \leq i, j \leq n$ and $1 \leq k, l \leq M$.

Here, the pseudo-posterior is given by $q(w|\theta) = \mathcal{N}(w|m, C^\top C)$ and the parameter space Θ includes both the mean and the Cholesky factor of the covariance of the pseudo-posterior, i.e., $\theta \triangleq \begin{pmatrix} m \\ \mathrm{vec}(C) \end{pmatrix}$. Given $q(w|\theta)$, the induced distribution $q(f|\theta) \triangleq \int_w p(f|w)q(w)\mathrm{d}w$ can be calculated exactly from Gaussian identities. Then, the log loss of the Bayesian prediction is $\ell((m, C), (x_*, y_*)) = -\log \mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)]$ where $q(f_*|\theta) = \mathcal{N}(f_*|a_*^\top m, \text{const} + a_*^\top C^\top C a_*)$, $a_* \triangleq K_{UU}^{-1}K_{U*}$, and const signifies terms that do not depend on m or C.

To apply Corollary 2, we require a bounded loss function which is also Lipschitz w.r.t. θ. To enable this, we define a "smoothed" log loss. Assume for now that $p \leq \xi < \infty$. We use a smoothing parameter $\alpha \in (0, 1)$ and define $\mathrm{nlog}^{(\alpha)}(p) \triangleq -\log((1-\alpha)p + \alpha)$. Then, the loss is bounded as $|\mathrm{nlog}^{(\alpha)}(p)|_\infty \leq \max\{|\log(\alpha)|, |\mathrm{nlog}^{(\alpha)}(\xi)|\}$. We also have that $\mathrm{nlog}^{(\alpha)}(p)$ is $L^{(\alpha)}$-Lipschitz w.r.t. p with $L^{(\alpha)} = \frac{1-\alpha}{\alpha}$.

We show in Appendix D that $\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)]$ is Lipschitz w.r.t. θ and infinity norm yielding that $\mathrm{nlog}^{(\alpha)}(\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)])$ is Lipschitz with constant $L^{(\alpha)}(L_m + L_C)$ where

$$L_m \triangleq \frac{\sqrt{M}}{\lambda_{\min}(K_{UU})}\|K_{U*}\|_2 \max_{f_*}\left|\frac{\mathrm{d}}{\mathrm{d}f_*}p(y_*|f_*)\right|, \quad (3)$$

$$L_C \triangleq \frac{M^3(M+1)B}{2(\lambda_{\min}(K_{UU}))^2}\|K_{U*}\|_2^2 \max_{f_*}\left|\frac{\mathrm{d}^2}{\mathrm{d}f_*^2}p(y_*|f_*)\right|, \quad (4)$$

3. Note, however, that prior results on linear regression in Germain et al. (2016) are not valid.

and $\lambda_{\min}(K_{UU})$ denotes the minimum eigenvalue of K_{UU}.

Therefore, to apply Corollary 2 to any non-conjugate sparse GP model with smoothed log loss, all we need is to (i) verify that $\exists \xi$ s.t. $\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)] < \xi$ and (ii) calculate bounds on $\left|\frac{\mathrm{d}}{\mathrm{d}f_*}p(y_*|f_*)\right|$ and $\left|\frac{\mathrm{d}^2}{\mathrm{d}f_*^2}p(y_*|f_*)\right|$. Condition (i) is easily achieved when Y is discrete, e.g., for binary classification and count regression. For standard regression, we can guarantee this by lower bounding the noise variance σ_Y^2 and upper bounding the range of X, Y. Bounds on the first and second derivatives (condition (ii)) are easily derived for the same likelihoods.

Corollary 3 *Randomized ERM using smoothed log loss with the sparse GP predictive distribution enjoys the bounds of Corollary 2 for regression, binary classification, Poisson regression.*

3. DLM for Other Loss Functions

We have shown that ERM-type algorithms performing direct minimization of log loss have strong performance guarantees for the Bayesian predictor, and we applied these results to the non-conjugate sparse GP model under a smoothed log loss. However, in some scenarios, we may want to minimize a different loss function requiring an explicit prediction. In this case, given a posterior $q(w)$ and example x with label y_{true}, the Bayesian predictor first identifies the optimal prediction $\hat{y} = \hat{y}_{q(w)}(x) = \arg\min_{y \in Y} E_{q(w)p(y'|x,w)}[\ell(y, y')]$ and then suffers the loss $\ell(q(w), (x, y_{\text{true}})) = \ell(\hat{y}_{q(w)}(x), y_{\text{true}})$. Therefore, the natural loss term for optimization is $L_{\text{B}} = \sum_i \ell(\hat{y}_{q(w)}(x_i), y_i)$. We note that L_{G} from the introduction, which implicitly uses the Gibbs log loss, is even less directly related to the learning goal in this case. On the other hand, the results of this paper do potentially apply to this more general setting as long as the conditions for the theorem hold.

Our theory does not directly apply to the square loss $(\hat{y} - y)^2$ because of the need for smoothing. However, it is interesting to consider the use of DLM for square loss and the resulting algorithms. In this case, sGP uses the standard regression model with Gaussian noise for prediction, that is, for calculating $\hat{y}$. It is well known that for the square loss the optimal predictor is the mean of the predictive distribution. As discussed above, for sGP the mean of the predictive distribution on example i is equal to $a_i^\top m$ where $a_i = K_{UU}^{-1} K_{Ui}$. Therefore the ERM algorithm will minimize $\sum_i (a_i^\top m - y_i)^2$ and similarly for the randomized ERM. We therefore see that, if we do not use regularization, the optimization criterion does not depend on the covariance of w and the optimization simplifies into a sparse variant of kernel least squares. The role of the posterior covariance, might become apparent with regularization, which might also be helpful to reduce some of the conditions in our theorem. We leave the derivation of DLM algorithms and performance guarantees for other loss functions to future work.

4. Conclusion

The paper points out the potential of DLM to yield a new type of approximate pseudo-Bayesian algorithm. In this paper we focused on the analysis of ERM and application to sparse GP. There are many important questions for future work including analysis for RLM, analysis for hyperparameter selection, removing the need for bounded or smoothed loss in our theorem, and investigating empirical properties of these algorithmic variants.

Acknowledgments

This work was partly supported by NSF under grants IIS-1906694 and IIS-1714440.

References

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*, pages 775–783. 2015.

Arindam Banerjee. On Bayesian bounds. In *ICML*, pages 81–88, 2006.

David Blei and John Lafferty. Correlated topic models. In *NIPS*, pages 147–154. 2006.

David Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *ICML*, pages 862–871, 2019.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *NIPS*, pages 1876–1884, 2016.

Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *JMLR*, pages 839–860, 2003.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Rishit Sheth and Roni Khardon. Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models. In *NIPS*, pages 5151–5161. 2017.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, pages 567–574, 2009.

Appendix A. Discretization

This straightforward proof shows that having a Lipschitz condition and bounded loss are sufficiently strong to make the problem simple by essentially learning on a grid. We include it here in order to put the other proofs and their potential improvements in context.

Let $\Theta \subset \mathbb{R}^M$ and $||\cdot||$ denote the infinity norm. Recall that we assume a bounded loss for the application of the discretization approach, i.e., $|\ell| \leq c$. Since Θ is assumed bounded, there exists a finite ρ-cover of Θ, $\dot{\Theta}$, i.e., $\forall \theta \in \Theta, \exists \dot{\theta} \in \dot{\Theta}$ s.t. $||\theta - \dot{\theta}|| \leq \rho$. Let

$$\theta_{\text{ERM}} \triangleq \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, (x_i, y_i)).$$

For an arbitrary $\theta \in \Theta$, let $\dot{\theta}$ denote the closest point in $\dot{\Theta}$ to θ. Since the loss is assumed to be L-Lipschitz in the hypothesis parameter, we have that

$$\forall (x,y) \in X \times Y, \quad |\ell(\theta, (x,y)) - \ell(\dot{\theta}, (x,y))| < L\rho \tag{5}$$

$$\forall S \in (X \times Y)^n, \quad \left| \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, (x_i, y_i)) - \frac{1}{n} \sum_{i=1}^{n} \ell(\dot{\theta}, (x_i, y_i)) \right| < L\rho. \tag{6}$$

In addition, by combining the union bound and Hoeffding's bound for bounded loss $|\ell| \leq c$ we have that, with probability $\geq 1 - \delta$ over the choice of sample S, for all $\theta \in \dot{\Theta}$:

$$\left| \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta, (x,y)) - \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, (x_i, y_i)) \right| \leq c \sqrt{\frac{2 \log(2|\dot{\Theta}|/\delta)}{n}}. \tag{7}$$

Let θ be any competitor for the posterior parameters. With probability $\geq 1 - \delta$ we have

$$\begin{aligned}
\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta_{\text{ERM}}, (x,y)) &\overset{(5)}{\leq} \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\dot{\theta}_{\text{ERM}}, (x,y)) + L\rho \\
&\overset{(7)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \ell(\dot{\theta}_{\text{ERM}}, (x_i, y_i)) + c\sqrt{\frac{2 \log(2|\dot{\Theta}|/\delta)}{n}} + L\rho \\
&\overset{(6)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta_{\text{ERM}}, (x_i, y_i)) + c\sqrt{\frac{2 \log(2|\dot{\Theta}|/\delta)}{n}} + 2L\rho \\
&\overset{(e)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \ell(\dot{\theta}, (x_i, y_i)) + c\sqrt{\frac{2 \log(2|\dot{\Theta}|/\delta)}{n}} + 2L\rho \\
&\overset{(7)}{\leq} \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\dot{\theta}, (x,y)) + 2c\sqrt{\frac{2 \log(2|\dot{\Theta}|/\delta)}{n}} + 2L\rho \\
&\overset{(5)}{\leq} \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta, (x,y)) + 2c\sqrt{\frac{2 \log(2|\dot{\Theta}|/\delta)}{n}} + 3L\rho
\end{aligned} \tag{8}$$

where (e) follows because ERM minimizes training set loss. With $|\dot{\Theta}| \leq \left(\frac{2B}{\rho}\right)^M$, the terms on the RHS of (8) depending on ρ are given by

$$2c\sqrt{\frac{2\log(2/\delta) + 2M\log(\frac{2B}{\rho})}{n}} + 3L\rho \leq \frac{2c}{\sqrt{n}}\Big(2\log(2/\delta) + 2M\log\Big(\frac{2B}{\rho}\Big)\Big) + 3L\rho$$
$$\leq \text{const} - \frac{4cM}{\sqrt{n}}\log\rho + 3L\rho.$$

The last expression is optimized when $\rho = \frac{4cM}{3\sqrt{n}L}$. Hence, we have that, with probability $\geq 1-\delta$ over the choice of S, $\forall\theta \in \Theta$,

$$\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta_{\mathrm{ERM}}, (x,y)) \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \ell(\theta, (x,y)) + 2c\sqrt{\frac{2\log(2/\delta) + 2M\log\left(\frac{3BL\sqrt{n}}{2cM}\right)}{n}} + \frac{4cM}{\sqrt{n}}. \tag{9}$$

Appendix B. Rademacher complexity

In this section, we show how the result of Meir and Zhang (2003) can be adapted to handle Bayesian predictors. Meir and Zhang (2003) assume a set of parameterized predictors $h(x;w) : X \mapsto Y$ and, in addition, assume that predictions can be averaged so that $\mathrm{E}_{q(w|\theta)}[h(x;w)]$ is a meaningful prediction. One can then apply the loss $\ell(y, \mathrm{E}_{q(w|\theta)}[h(x;w)])$. For Bayesian predictors, we average the probabilities in $\hat{p} = \mathrm{E}_{q(w|\theta)}[p(y|x,w)]$ but not the predictions themselves. Nonetheless, the same proof technique can be adapted to yield a result for some loss functions, specifically the smoothed log loss discussed in the main paper.

We next develop the details. Note that, although the results of Meir and Zhang (2003) are for unbounded losses, their conditions are complex and it is not clear how to apply these results directly for Bayesian predictors such as the sparse GP discussed in this paper.

Assuming a family of distributions Q over w and an upper bound $p(y|x,w) \leq p_{y|w}$ (where $p_{y|w}$ is a constant), uniform convergence for the averaged predictor $\mathrm{E}_{q(w)}[p(y|x,w)]$ under the smoothed log loss $\mathrm{nlog}^{(\alpha)}()$ will be shown. From Theorem 26.5.1 of Shalev-Shwartz and Ben-David (2014)[4], for all $q \in Q$, the following holds with probability $1-\delta$ over the choice of S:

$$\left|\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\Big[\mathrm{nlog}^{(\alpha)}\big(\mathop{\mathrm{E}}_{q(w)}[p(y|w,x)]\big)\Big] - \frac{1}{n}\sum_{i=1}^{n}\mathrm{nlog}^{(\alpha)}\big(\mathop{\mathrm{E}}_{q(w)}[p(y_i|w,x_i)]\big)\right|$$
$$\leq 2\mathop{\mathrm{E}}_{S'\sim\mathcal{D}^n} R_n(\ell\circ\mathcal{H}\circ S') + c\sqrt{\frac{2\log(2/\delta)}{n}}, \tag{10}$$

where $\mathcal{H} \triangleq \{\mathrm{E}_{q(w)}[p(\cdot;w,\cdot)] : q \in Q\}$, $\circ$ stands for function composition,

$$R_n(\ell\circ\mathcal{H}\circ S) \triangleq \mathop{\mathrm{E}}_{\sigma}\Big[\sup_{q\in Q}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ \mathrm{nlog}^{(\alpha)}\big(\mathop{\mathrm{E}}_{q(w)}[p(y_i|w,x_i)]\big)\Big]$$

4. See also Corollary 4 of Meir and Zhang (2003). These results give one-sided bounds but can be easily adapted to give the two sided bound shown here.

for Rademacher variables σ, and $c = \max\{|\log(\alpha)|, |\mathrm{nlog}^{(\alpha)}(p_{y|w})|\}$. Since $\mathrm{nlog}^{(\alpha)}(\cdot)$ is $L^{(\alpha)}$-Lipschitz, by Theorem 7 of Meir and Zhang (2003), we have

$$R_n(\ell \circ \mathcal{H} \circ S) \leq L^{(\alpha)} \mathop{\mathrm{E}}_{\sigma} \Big[\sup_{q \in Q} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathop{\mathrm{E}}_{q(w)} [p(y_i|w, x_i)] \Big].$$

Next, we slightly adapt the argument outlined in Sections 5 and 6.1 of Meir and Zhang (2003). Fix some constant $\lambda \in (0, \infty)$ and sample-independent distribution $p(w)$ over w. By applying the compression lemma (Banerjee, 2006) to $\mathrm{E}_{q(w)} \left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i p(y_i|w, x_i) \right)$, we have

$$\begin{aligned}
R_n(\ell \circ \mathcal{H} \circ S) &\leq \frac{L^{(\alpha)}}{\lambda} \mathop{\mathrm{E}}_{\sigma} \left[\sup_{q \in Q} \left(\mathrm{KL}(q(w), p(w)) + \log \mathop{\mathrm{E}}_{p(w)} \exp \left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i p(y_i|w, x_i) \right) \right) \right] \\
&= \frac{L^{(\alpha)}}{\lambda} \left(\sup_{q \in Q} \mathrm{KL}(q(w), p(w)) + \mathop{\mathrm{E}}_{\sigma} \left[\log \mathop{\mathrm{E}}_{p(w)} \exp \left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i p(y_i|w, x_i) \right) \right] \right) \\
&\leq \frac{L^{(\alpha)}}{\lambda} \left(\sup_{q \in Q} \mathrm{KL}(q(w), p(w)) + \log \mathop{\mathrm{E}}_{\sigma} \left[\mathop{\mathrm{E}}_{p(w)} \exp \left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i p(y_i|w, x_i) \right) \right] \right) \\
&= \frac{L^{(\alpha)}}{\lambda} \left(\sup_{q \in Q} \mathrm{KL}(q(w), p(w)) + \log \mathop{\mathrm{E}}_{p(w)} \left[\mathop{\mathrm{E}}_{\sigma} \exp \left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i p(y_i|w, x_i) \right) \right] \right) \\
&\leq \frac{L^{(\alpha)}}{\lambda} \left(\sup_{q \in Q} \mathrm{KL}(q(w), p(w)) + \log \mathop{\mathrm{E}}_{p(w)} \left[\exp \left(\frac{\lambda^2}{2n^2} \sum_{i=1}^{n} (p(y_i|w, x_i))^2 \right) \right] \right) \qquad (11) \\
&\leq \frac{L^{(\alpha)}}{\lambda} \left(\sup_{q \in Q} \mathrm{KL}(q(w), p(w)) + \log \mathop{\mathrm{E}}_{p(w)} \left[\exp \left(\frac{\lambda^2}{2n} p_{y|w}^2 \right) \right] \right) \\
&= \frac{L^{(\alpha)}}{\lambda} \left(A + \frac{\lambda^2}{2n} p_{y|w}^2 \right), \qquad (12)
\end{aligned}$$

where (11) follows from the inequality $\mathrm{E}_{\sigma_i} \exp(\sigma_i a_i) \leq \exp(a_i^2/2)$ (Lemma A.6 of Shalev-Shwartz and Ben-David (2014)), and we have defined $A \triangleq \sup_{q \in Q} \mathrm{KL}(q(w), p(w))$. Optimizing (12) w.r.t. λ yields $\lambda^\star = \sqrt{\frac{2An}{p_{y|w}}}$. Substituting this value in (12) results in

$$R_n(\ell \circ \mathcal{H} \circ S) \leq L \sqrt{\frac{A p_{y|w}(1 + p_{y|w})}{2n}}. \qquad (13)$$

Utilizing (13) in (10), we have that with probability $1 - \delta$ over the choice of S, for all $q \in Q$,

$$\begin{aligned}
&\left| \mathop{\mathrm{E}}_{(x,y) \sim \mathcal{D}} \Big[\mathrm{nlog}^{(\alpha)} \big(\mathop{\mathrm{E}}_{q(w)} p(y|w, x) \big) \Big] - \frac{1}{n} \sum_{i=1}^{n} \mathrm{nlog}^{(\alpha)} \big(\mathop{\mathrm{E}}_{q(w)} p(y_i|w, x_i) \big) \right| \\
&\qquad \leq L^{(\alpha)} \sqrt{\frac{A p_{y|w}(1 + p_{y|w})}{2n}} + c \sqrt{\frac{2 \log(2/\delta)}{n}}. \qquad (14)
\end{aligned}$$

Defining $Q_A \triangleq \{q \in Q \text{ s.t. } \mathrm{KL}(q,p) \le A\}$, and the ERM hypothesis as

$$q_{\mathrm{ERM}}(w) \triangleq \arg\min_{q \in Q_A} \frac{1}{n} \sum_{i=1}^{n} \mathrm{nlog}^{(\alpha)}\big(\mathop{\mathrm{E}}_{q(w)} p(y_i|w,x_i)\big),$$

we can use the above with the standard argument for ERM to get that, with probability $1-\delta$ over the choice of S, for all $q \in Q_A$,

$$\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \Big[\mathrm{nlog}^{(\alpha)}\big(\mathop{\mathrm{E}}_{q_{\mathrm{ERM}}(w)} p(y|w,x)\big)\Big] \le \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \Big[\mathrm{nlog}^{(\alpha)}\big(\mathop{\mathrm{E}}_{q(w)} p(y|w,x)\big)\Big] + L^{(\alpha)}\sqrt{\frac{2Ap_{y|w}(1+p_{y|w})}{n}} + c\sqrt{\frac{8\log(2/\delta)}{n}}. \quad (15)$$

Applications of this results to sparse GP are possible as outlined in the main paper. Comparing this result to the discretization proof and randomization proof (below), we see that the requirements for Lipschitz constants are weaker. Here, we only need $L^{(\alpha)}$ whereas other proofs require a Lipschitz condition w.r.t. the parameter θ. This proof can potentially yield bounds that do not depend on the dimension M. Note that, applied to Gaussian distributions, A implicitly depends on M, so a direct application does include such a dependence. But Meir and Zhang (2003) show how to use structural risk minimization to get around this dimension dependence through data-dependent bounds.

Appendix C. Randomized ERM

Let $\bar{\Theta}$ denote some known subset of Θ, i.e., $\bar{\Theta} \subset \Theta$, and let $\{q_{\mathrm{jit}}(\theta|\bar{\theta})\}$ denote a family of distributions over Θ parameterized by members of the subset $\bar{\Theta}$. The members of the family are as yet unspecified, but represent "jitter" distributions which will be defined shortly. Let

$$\bar{\theta}_{\mathrm{ERM}} \triangleq \arg\min_{\bar{\theta}\in\bar{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})} \Big[\ell(\theta,(x_i,y_i))\Big].$$

Note, where we exchange order of expectations in the following development, we assume the conditions of Fubini's theorem are met. The following lemma is standard (see Shalev-Shwartz and Ben-David (2014)):

Lemma 1. For all $\bar{\theta} \in \bar{\Theta}$,

$$\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n} \left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})} \Big[\ell(\theta,(x,y))\Big]\right]\right] \le \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})} \Big[\ell(\theta,(x,y))\Big]\right] + \mathop{\mathrm{E}}_{S\sim\mathcal{D}^n} \left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}} \left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})} \Big[\ell(\theta,(x,y))\Big]\right] - \frac{1}{n}\sum_{i=1}^{n} \left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})} \Big[\ell(\theta,(x_i,y_i))\Big]\right]\right]. \quad (16)$$

Proof It is sufficient to prove that

$$\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\Big[\ell(\theta,(x_i,y_i))\Big]\right]\leq\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[\ell(\theta,(x,y))\Big]\right]$$

holds for all $\bar{\theta}\in\bar{\Theta}$: Since $\bar{\theta}_{\mathrm{ERM}}$ is the ERM hypothesis, it follows that $\forall\bar{\theta}\in\bar{\Theta}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\Big[\ell(\theta,(x_i,y_i))\Big]\leq\frac{1}{n}\sum_{i=1}^{n}\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[\ell(\theta,(x_i,y_i))\Big].$$

Taking expectations of both sides w.r.t. $\mathcal{D}^n$ yields the result. ■

The following lemma uses a technique from Germain et al. (2016) and Sheth and Khardon (2017). The novelty, however, is to apply the compression lemma at a level higher than previous work. Here, we use it at the level of parameters θ defining the posterior distribution, which requires us to introduce the jitter, whereas previous work applied it at the level of base parameter w. This gives a qualitatively different result.

Lemma 2. Let $p(\theta)$ be any sample-independent distribution over Θ and define

$$\Psi(\lambda,n)\triangleq\log\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{p(\theta)}\exp\Big(\lambda\Big(\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\Big[\ell(\theta,(x,y))\Big]-\frac{1}{n}\sum_{i=1}^{n}\ell(\theta,(x_i,y_i))\Big)\Big)\right].$$

Then, $\forall\bar{\theta}\in\bar{\Theta}$,

$$\begin{aligned}\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}&\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\Big[\ell(\theta,(x,y))\Big]\right]\right]\\&\leq\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[\ell(\theta,(x,y))\Big]\right]\\&\quad+\frac{1}{\lambda}\left[\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\Big[\mathrm{KL}(q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}}),p(\theta))\Big]+\Psi(\lambda,n)\right],\end{aligned}\tag{17}$$

Proof First, apply Fubini's theorem to change the order of expectations of

$$\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\Big[\ell(\theta,(x,y))\Big]\right]$$

in (16). Then, apply the compression lemma, $\mathrm{E}_{q(\theta)}f(\theta)\leq\mathrm{KL}(q(\theta),p(\theta))+\log\mathrm{E}_{p(\theta)}e^{f(\theta)}$, with $q(\theta)=q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})$ and $f(\theta)=\lambda\Big(\mathrm{E}_{(x,y)\sim\mathcal{D}}\,\ell(\theta,(x,y))-\frac{1}{n}\sum_{i=1}^{n}\ell(\theta,(x_i,y_i))\Big)$ to the resulting expression within the expectation w.r.t. $S\sim\mathcal{D}^n$. Finally, take the expectation w.r.t. $S\sim\mathcal{D}^n$ and note that $\mathrm{E}_{S\sim\mathcal{D}^n}\log(\cdot)\leq\log\mathrm{E}_{S\sim\mathcal{D}^n}(\cdot)$ by Jensen's inequality to yield the statement of the lemma. ■

Lemma 3. Let some norm $||\cdot||$ over Θ be given. For an L-Lipschitz function $\ell(\theta)$ w.r.t. $||\cdot||$, we have that $\forall\bar{\theta}\in\bar{\Theta}, \forall\theta'\in\Theta$,

$$\ell(\theta',(x,y)) - L \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[||\theta-\theta'||\Big] \leq \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})} \ell(\theta,(x,y)) \leq \ell(\theta',(x,y)) + L \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[||\theta-\theta'||\Big]. \tag{18}$$

Proof If $\ell(\theta)$ is L-Lipschitz w.r.t. $||\cdot||$, then $\forall\theta,\theta'\in\Theta, |\ell(\theta)-\ell(\theta')|\leq L||\theta-\theta'||$, or, $\ell(\theta')-L||\theta-\theta'||\leq\ell(\theta)\leq\ell(\theta')+L||\theta-\theta'||$. Since this holds for all values of θ (given some θ'), it also holds in expectation over any distribution in θ, specifically $q_{\mathrm{jit}}(\theta|\bar{\theta})$. ■

Lemma 4. Let $p(\theta)$ be any sample-independent distribution over Θ and $\ell(\theta,(x,y))$ be L-Lipschitz in its first argument. Then, $\forall\theta'\in\Theta$,

$$\begin{aligned}\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\mathrm{ERM}},(x,y))\right]\right] &\\ \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta',(x,y))\right] &+ 2L\max_{\bar{\theta}\in\bar{\Theta}}\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[||\theta-\bar{\theta}||\Big] + LD\\ &+\frac{1}{\lambda}\left[\max_{\bar{\theta}\in\bar{\Theta}}\left[\mathrm{KL}(q_{\mathrm{jit}}(\theta|\bar{\theta}),p(\theta))\right]+\Psi(\lambda,n)\right],\end{aligned} \tag{19}$$

where $D\triangleq\max_{\theta'\in\Theta\setminus\bar{\Theta}}\min_{\bar{\theta}\in\bar{\Theta}}||\bar{\theta}-\theta'||$ and $\Psi(\lambda,n)$ is defined in Lemma 2.
Proof Following from the left inequality of (18) with $\bar{\theta}=\theta'=\bar{\theta}_{\mathrm{ERM}}$, we have

$$\begin{aligned}\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\mathrm{ERM}},(x,y))\right] - L\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\Big[||\theta-\bar{\theta}_{\mathrm{ERM}}||\Big]\right]&\\ \leq \mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\ell(\theta,(x,y))\right]\right].&\end{aligned} \tag{20}$$

Utilizing (20) in (17) yields that, $\forall\bar{\theta}\in\bar{\Theta}$,

$$\begin{aligned}\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\mathrm{ERM}},(x,y))\right]\right]&\\ \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\Big[\ell(\theta,(x,y))\Big]\right]&\\ + L\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\Big[||\theta-\bar{\theta}_{\mathrm{ERM}}||\Big]\right]&\\ +\frac{1}{\lambda}\left[\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathrm{KL}(q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}}),p(\theta))\right]+\Psi(\lambda,n)\right].&\end{aligned} \tag{21}$$

Following from the right inequality of (18), we have that, $\forall \bar{\theta} \in \bar{\Theta}, \forall \theta' \in \Theta$,

$$\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})} \ell(\theta,(x,y))\right] \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta',(x,y))\right] + L \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\theta'||\right]. \tag{22}$$

Utilizing (22) in (21) yields that, $\forall \bar{\theta} \in \bar{\Theta}, \theta' \in \Theta$,

$$\begin{aligned}
&\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\mathrm{ERM}},(x,y))\right]\right] \\
&\qquad\qquad \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta',(x,y))\right] \\
&\qquad + L \mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}})}\left[||\theta-\bar{\theta}_{\mathrm{ERM}}||\right]\right] + L \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\theta'||\right] \\
&\qquad\qquad + \frac{1}{\lambda}\left[\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathrm{KL}(q_{\mathrm{jit}}(\theta|\bar{\theta}_{\mathrm{ERM}}),p(\theta))\right] + \Psi(\lambda,n)\right].
\end{aligned} \tag{23}$$

Next, we develop a uniform bound over θ' for the term $\mathrm{E}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\theta'||\right]$. Since (23) holds for all $\bar{\theta} \in \bar{\Theta}$, we consider how $\bar{\theta}$ can be selected per $\theta' \in \Theta$. First, note that

$$\begin{aligned}
\mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\theta'||\right] &= \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\bar{\theta}+\bar{\theta}-\theta'||\right] \\
&\leq \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\bar{\theta}||+||\bar{\theta}-\theta'||\right] \\
&= \mathop{\mathrm{E}}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\bar{\theta}||\right] + ||\bar{\theta}-\theta'||.
\end{aligned} \tag{24}$$

Now, when $\theta' \in \bar{\Theta}$, a uniform bound over $\bar{\theta}$ for $\mathrm{E}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\bar{\theta}||\right]$ translates to a uniform bound over θ' for $\mathrm{E}_{q_{\mathrm{jit}}(\theta|\bar{\theta})}\left[||\theta-\theta'||\right]$ since it is possible to select $\bar{\theta} = \theta'$ in (24). When $\theta' \in \Theta\backslash\bar{\Theta}$, the distance to the "closest" point in $\bar{\Theta}$ to θ' is $\min_{\bar{\theta}\in\bar{\Theta}} ||\bar{\theta}-\theta'||$. Then, the second term of (24) is uniformly upper-bounded over θ' by $D \triangleq \max_{\theta'\in\Theta\backslash\bar{\Theta}} \min_{\bar{\theta}\in\bar{\Theta}} ||\bar{\theta}-\theta'||$. Combining the two cases yields the lemma. ■

Jitter distributions. First, we define Θ parametrically as a function of some $\rho > 0$ and relative to $\bar{\Theta}$. Assume $\bar{\Theta}$ is a subset of some space T (equipped with norm $||\cdot||$), and define Θ as the set $\{\theta \in T \text{ s.t. } \min_{\bar{\theta}\in\bar{\Theta}} ||\theta-\bar{\theta}|| \leq \rho\}$. Then, $\bar{\Theta} \subset \Theta$, and $D \leq \rho$. The ρ-ball centered at $\bar{\theta} \in \bar{\Theta}$ is denoted $B_\rho(\bar{\theta}) \triangleq \{\theta \in \Theta \text{ s.t. } ||\theta-\bar{\theta}|| \leq \rho\}$. Let the jitter distribution $q_{\mathrm{jit}}(\theta|\bar{\theta})$ be defined as the following uniform density with support in Θ:

$$q_{\mathrm{jit}}(\theta|\bar{\theta}) = \begin{cases} \frac{1}{\mathrm{vol}(B_\rho(\bar{\theta}))}, & \theta \in B_\rho(\bar{\theta}), \\ 0, & \text{else.} \end{cases}$$

Letting $\text{supp}(q_{\text{jit}}(\theta|\bar{\theta})) \triangleq B_\rho(\bar{\theta})$, we have

$$\begin{aligned}\mathop{\mathrm{E}}_{q_{\text{jit}}(\theta|\bar{\theta})}\Big[||\theta-\bar{\theta}||\Big] &= \int_{\text{supp}(q_{\text{jit}}(\theta|\bar{\theta}))} \frac{1}{\text{vol}(\text{supp}(q_{\text{jit}}(\theta|\bar{\theta})))}||\theta-\bar{\theta}||\mathrm{d}\theta \\ &\leq \int_{\text{supp}(q_{\text{jit}}(\theta|\bar{\theta}))} \frac{1}{\text{vol}(\text{supp}(q_{\text{jit}}(\theta|\bar{\theta})))}\rho\mathrm{d}\theta \\ &= \rho.\end{aligned}$$

For

$$p(\theta) = \begin{cases} \frac{1}{\text{vol}(\Theta)}, \theta \in \Theta, \\ 0, \text{else}, \end{cases}$$

the KL divergence $\text{KL}(q_{\text{jit}}(\theta|\bar{\theta}), p(\theta))$ is given by

$$\begin{aligned}\text{KL}(q_{\text{jit}}(\theta|\bar{\theta}), p(\theta)) &= \int q_{\text{jit}}(\theta|\bar{\theta}) \log \frac{q_{\text{jit}}(\theta|\bar{\theta})}{p(\theta)}\mathrm{d}\theta \\ &= \int_{\text{supp}(q_{\text{jit}}(\theta|\bar{\theta}))} \frac{1}{\text{vol}(\text{supp}(q_{\text{jit}}(\theta|\bar{\theta})))} \log \frac{\text{vol}(\Theta)}{\text{vol}(\text{supp}(q_{\text{jit}}(\theta|\bar{\theta})))}\mathrm{d}\theta \\ &= \log \frac{\text{vol}(\Theta)}{\text{vol}(\text{supp}(q_{\text{jit}}(\theta|\bar{\theta})))}. \end{aligned} \tag{25}$$

For this choice of jitter distribution and prior, (19) becomes

$$\begin{aligned}&\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\text{ERM}}, (x,y))\right]\right] \\ &\leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta', (x,y))\right] + 3L\rho + \frac{1}{\lambda}\left[\max_{\bar{\theta}\in\bar{\Theta}}\Big[-\log \text{vol}(B_\rho(\bar{\theta}))\Big] + \log \text{vol}(\Theta) + \Psi(\lambda, n)\right].\end{aligned} \tag{26}$$

If $\text{vol}(B_\rho(\bar{\theta})) = \kappa\rho^M$ for some constants κ, M, then the RHS of (26) is minimized for $\rho = \frac{M}{3L\lambda}$, and we have that, $\forall\theta' \in \Theta$,

$$\begin{aligned}&\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\text{ERM}}, (x,y))\right]\right] \\ &\leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta', (x,y))\right] + \frac{1}{\lambda}\left[M + M\log\left(\frac{3L\lambda}{M}\right) + \log\frac{1}{\kappa} + \log \text{vol}(\Theta) + \Psi(\lambda, n)\right].\end{aligned} \tag{27}$$

C.1. Applications

M-dimensional parameter. Let $T = \mathbb{R}^M$, $||\cdot|| = ||\cdot||_\infty$, and $\bar{\Theta} = [-B+\rho, B-\rho]^M$. Then, $\Theta = [-B, B]^M$, $\text{vol}(B_\rho(\bar{\theta})) = 2^M\rho^M$, $\text{vol}(\Theta) = 2^M B^M$, and the last term of (27) is equal to

$$\frac{1}{\lambda}\left[M + M\log\left(\frac{3BL\lambda}{M}\right) + \Psi(\lambda, n)\right].$$

This completes the proof of Theorem 1 from the main paper. Next to prove Corollary 2, recall that, for bounded loss $|\ell| \leq c$, $\Psi(\lambda, n) \leq \frac{2\lambda^2 c^2}{n}$. Setting $\lambda = \sqrt{n}$, we have that, $\forall \theta' \in \Theta$,

$$\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\mathrm{ERM}},(x,y))\right]\right] \leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta',(x,y))\right]+\frac{1}{\sqrt{n}}\left[M+M\log\left(\frac{3BL\sqrt{n}}{M}\right)+2c^2\right]. \tag{28}$$

This bound holds when using the randomized ERM learning rule with the class

$$q_{\mathrm{jit}}(\theta|\bar{\theta}) = \prod_{m=1}^{M} \mathcal{U}(\theta_m|\bar{\theta}_m - \frac{M}{3L\sqrt{n}}, \bar{\theta}_m + \frac{M}{3L\sqrt{n}})$$

for $\bar{\theta} \in [-B + \frac{M}{3L\sqrt{n}}, B - \frac{M}{3L\sqrt{n}}]^M$.

Product spaces. Let $T = T_1 \times T_2 = \mathbb{R}^{M_1} \times \mathbb{R}^{M_2}$, $\|\cdot\|_T = \|\cdot\|_{T_1,\infty} + \|\cdot\|_{T_2,\infty}$ and $\bar{\Theta} = \bar{\Theta}_1 \times \bar{\Theta}_2 = [-B_1+\rho_1, B_1-\rho_1]^{M_1} \times [-B_2+\rho_2, B_2-\rho_2]^{M_2}$. Then, $\Theta = \Theta_1 \times \Theta_2 = [-B_1, B_1]^{M_1} \times [-B_2, B_2]^{M_2}$ and $\mathrm{vol}(B_\rho(\bar{\theta})) = 2^{M_1}\rho_1^{M_1}2^{M_2}\rho_2^{M_2}$, $\mathrm{vol}(\Theta) = 2^{M_1}{B_1}^{M_1}2^{M_2}{B_2}^{M_2}$. In this case, the RHS of (26) is optimized for $\rho_1^\star = \frac{M_1}{3L\lambda}$ and $\rho_2^\star = \frac{M_2}{3L\lambda}$. Assuming the same value L of Lipschitz constant for both spaces Θ_1 and Θ_2, we have that $\forall \theta' \in \Theta$,

$$\begin{aligned}\mathop{\mathrm{E}}_{S\sim\mathcal{D}^n}&\left[\mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\bar{\theta}_{\mathrm{ERM}},(x,y))\right]\right] \\ &\leq \mathop{\mathrm{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\theta',(x,y))\right] \\ &+ \frac{1}{\lambda}\left[M_1 + M_2 + M_1\log\left(\frac{3B_1L\lambda}{M_1}\right) + M_2\log\left(\frac{3B_2L\lambda}{M_2}\right) + \Psi(\lambda,n)\right].\end{aligned} \tag{29}$$

This bound holds when using the randomized ERM learning rule with the class

$$q_{\mathrm{jit}}(\theta|\bar{\theta}) = \prod_{m_1=1}^{M_1} \mathcal{U}(\theta_{m_1}|\bar{\theta}_{m_1} - \rho_1^\star, \bar{\theta}_{m_1} + \rho_1^\star) \prod_{m_2=1}^{M_2} \mathcal{U}(\theta_{m_2}|\bar{\theta}_{m_2} - \rho_2^\star, \bar{\theta}_{m_2} + \rho_2^\star)$$

for $\bar{\theta} \in [-B_1+\rho_1^\star, B_1-\rho_1^\star]^{M_1} \times [-B_2+\rho_2^\star, B_2-\rho_2^\star]^{M_2}$.

Appendix D. Sparse GPs

To show that the smoothed log loss of the pseudo-Bayesian prediction $\mathrm{nlog}^{(\alpha)}\,\mathrm{E}_{q(f_*|\theta)}\,p(y_*|f_*)$ is Lipschitz, we use the fact that the composition of Lipschitz functions is Lipschitz, and focus on determining a Lipschitz constant of $\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)]$ w.r.t. $\theta = \begin{pmatrix} m \\ \mathrm{vec}(C)\end{pmatrix}$ and infinity norm. For a differentiable function, a Lipschitz constant is given by the maximum of the dual norm of its gradient[5]. Since we use infinity norm on Θ, the dual norm will

5. This fact follows from the multivariate mean value theorem and Holder's inequality.

be the 1-norm. Following standard Gaussian (see e.g., Rezende et al. (2014)) and matrix derivative identities (see e.g., Petersen et al. (2008)), the gradient w.r.t. m is

$$a_* \int_{f_*} \Big(\frac{\mathrm{d}}{\mathrm{d}f_*} p(y_*|f_*)\Big) q(f_*) \mathrm{d}f_*.$$

The 1-norm of this is quantity is upper-bounded by

$$\begin{aligned}
\|a_*\|_1 \max_{f_*} \left|\frac{\mathrm{d}}{\mathrm{d}f_*} p(y_*|f_*)\right| &\leq \sqrt{M}\|a_*\|_2 \max_{f_*} \left|\frac{\mathrm{d}}{\mathrm{d}f_*} p(y_*|f_*)\right| \\
&\leq \sqrt{M}\lambda_{\max}(K_{UU}^{-1})\|K_{U*}\|_2 \max_{f_*} \left|\frac{\mathrm{d}}{\mathrm{d}f_*} p(y_*|f_*)\right| \\
&= \frac{\sqrt{M}}{\lambda_{\min}(K_{UU})}\|K_{U*}\|_2 \max_{f_*} \left|\frac{\mathrm{d}}{\mathrm{d}f_*} p(y_*|f_*)\right|,
\end{aligned}$$

where the first inequality follows from $\|\cdot\|_1 \leq \sqrt{M}\|\cdot\|_2$, the second inequality follows from $\|Ax\|_2 \leq \|A\|_2\|x\|_2$, and $\lambda_{\min}(K_{UU})$ denotes the minimum eigenvalue of K_{UU}. The Lipschitz constant of $\mathrm{E}_{q(f_*)}[p(y_*|f_*)]$ w.r.t. m and infinity norm is, therefore, bounded by

$$\frac{\sqrt{M}}{\lambda_{\min}(K_{UU})}\|K_{U*}\|_2 \max_{f_*} \left|\frac{\mathrm{d}}{\mathrm{d}f_*} p(y_*|f_*)\right|. \tag{30}$$

Similarly, the gradient of $\mathrm{nlog}^{(\alpha)}(\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)])$ w.r.t. the vectorized Cholesky factor of the variational covariance[6] is given by

$$[(a_* a_*^\top) \otimes I]\mathrm{vec}(C) \int_{f_*} \Big(\frac{\mathrm{d}^2}{\mathrm{d}{f_*}^2} p(y_*|f_*)\Big) q(f_*) \mathrm{d}f_*.$$

The 1-norm of $[(a_* a_*^\top) \otimes I]\mathrm{vec}(C)$ is given by

$$\begin{aligned}
\|[(a_* a_*^\top) \otimes I]\mathrm{vec}(C)\|_1 &\leq \|[(a_* a_*^\top) \otimes I]\|_1 \|\mathrm{vec}(C)\|_1 \\
&= M\|a_* a_*^\top\|_1 \|\mathrm{vec}(C)\|_1 \\
&\leq M\|a_*\|_1^2 \|\mathrm{vec}(C)\|_1 \\
&\leq \frac{1}{2}M^2(M+1)B\|a_*\|_1^2,
\end{aligned}$$

where the last inequality follows from the infinity-norm bound on the hypothesis space, $\|\mathrm{vec}(C)\|_\infty \leq B$, and the fact that only $\frac{1}{2}M(M+1)$ entries of the Cholesky factor are non-zero. Hence, the Lipschitz constant of $\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)]$ w.r.t. $\mathrm{vec}(C)$ and infinity norm is bounded by

$$\frac{M^3(M+1)B}{2(\lambda_{\min}(K_{UU}))^2}\|K_{U*}\|_2^2 \max_{f_*} \left|\frac{\mathrm{d}^2}{\mathrm{d}f_*^2} p(y_*|f_*)\right|. \tag{31}$$

The total Lipschitz constant of $\mathrm{E}_{q(f_*|\theta)}[p(y_*|f_*)]$ w.r.t. $\begin{pmatrix} m \\ \mathrm{vec}(C) \end{pmatrix}$ and infinity norm is bounded by the sum of (30) and (31).

6. Recall $\mathrm{vec}(AXB) = (B^\top \otimes A)\mathrm{vec}(X)$.

Appendix E. Correlated topic model

The correlated topic model (CTM) of Blei and Lafferty (2006) is a generative model for documents. For each document, CTM first draws $w \sim \mathcal{N}(\mu, \Sigma)$, $w \in \mathbb{R}^M$ where $\{\mu, \Sigma\}$ are model parameters, and then maps this vector to the $(M+1)$-simplex with the logistic transformation, $\phi = h(w)$. The function $h(w)$ is given by $h_k(w) = \frac{\exp(w_k)}{1+\sum_{\ell=1}^M \exp(w_\ell)}$ for $k < M+1$ and $h_{M+1}(w) = \frac{1}{1+\sum_{\ell=1}^M \exp(w_\ell)}$. For each position i in the document, the latent topic variable, f_i, is drawn from Discrete(ϕ), and the word y_i is drawn from a Discrete($\beta_{f_i,\cdot}$) where β denotes the $M+1$ topics and is treated as a parameter of the model. In this case $p(f|w)$ can be integrated out analytically and $p(y|w) = \sum_{k=1}^{M+1} \beta_{k,y} h_k(w)$. See further discussion of these details in Sheth and Khardon (2017). Here we assume that the parameter β is constrained to be smoothed, that is, $\forall k, y,\ \beta_{k,y} \geq \gamma > 0$.

First, we derive the Lipschitz constant w.r.t. the mean of the approximate posterior. Following standard Gaussian identites (Rezende et al., 2014), we have

$$\frac{\partial}{\partial m_j} \log\Big(\mathop{\mathrm{E}}_{q(w|(m,\mathrm{vec}(C)))}[p(y|w)]\Big) = \frac{1}{\mathrm{E}_{q(w)}[p(y|w)]} \mathop{\mathrm{E}}_{q(w)}\Big[\frac{\partial}{\partial w_j} p(y|w)\Big].$$

Letting $\delta(\cdot)$ denote the delta function, we have $\frac{\partial h_k(w)}{\partial w_j} = (-h_k(w) + \delta(k-j))h_j(w)$ for $k < M+1$ and $\frac{\partial h_{M+1}(w)}{\partial w_j} = -h_{M+1}(w)h_j(w)$. Since $h_k(w) \leq 1$, $\left|\frac{\partial h_k(w)}{\partial w_j}\right| \leq 1$, and therefore $\left|\frac{\partial}{\partial w_j} p(y|w)\right| = \left|\sum_{k=1}^{M+1} \beta_{k,y} \frac{\partial h_k(w)}{\partial w_j}\right| \leq M+1$. Hence,

$$\left|\frac{\partial}{\partial m_j} \log\Big(\mathop{\mathrm{E}}_{q(w|(m,\mathrm{vec}(C)))}[p(y|w)]\Big)\right| \leq \frac{M+1}{\mathrm{E}_{q(w)}[p(y|w)]},$$

and

$$\left\|\nabla_m \log\Big(\mathop{\mathrm{E}}_{q(w|(m,\mathrm{vec}(C)))} p(y|w)\Big)\right\|_1 \leq \frac{(M+1)M}{\mathrm{E}_{q(w)}[p(y|w)]} \leq \frac{(M+1)M}{\gamma}. \tag{32}$$

To derive the Lipschitz constant w.r.t. Cholesky factor C of the covariance V, we proceed by bounding the entries of the derivative. For a scalar-valued function g of the covariance $V = C^\top C$, we have[7] $\frac{\partial g(C^\top C)}{\partial C_{rs}} = 2\sum_{t \leq r} C_{rt} \frac{\partial g}{\partial V_{ts}}$ where r, s range over the entries of the Cholesky factor. For $g(V) = \log\Big(\mathrm{E}_{q(w|(m,V))}[p(y|w)]\Big)$, from Rezende et al. (2014) we have

$$\frac{\partial}{\partial V_{ts}} \log\Big(\mathop{\mathrm{E}}_{q(w|(m,V))}[p(y|w)]\Big) = \frac{1}{2}\frac{1}{\mathrm{E}_{q(w)}[p(y|w)]} \mathop{\mathrm{E}}_{q(w)}\Big[\frac{\partial^2}{\partial w_t \partial w_s} p(y|w)\Big].$$

The second derivative $\frac{\partial^2}{\partial w_t \partial w_s} h_k(w) = -h_s(w)(-h_k(w) + \delta(k-t))h_t(w) + (-h_k(w) + \delta(k-s))(-h_s(w) + \delta(s-t))h_t(w)$ has entries bounded as $\left|\frac{\partial^2}{\partial w_t \partial w_s} h_k(w)\right| \leq 2$. Therefore, $\left|\frac{\partial^2}{\partial w_t \partial w_s} p(y|w)\right| \leq 2(M+1)$, and

$$\left|\frac{\partial}{\partial V_{ts}} \log\Big(\mathop{\mathrm{E}}_{q(w|(m,V))}[p(y|w)]\Big)\right| \leq \frac{M+1}{\mathrm{E}_{q(w)}[p(y|w)]}.$$

7. The matrix identity is $\frac{\partial g(C^\top C)}{\partial C} = 2\mathrm{triu}\Big(C \frac{\partial g}{\partial V}\Big)$ where triu$(\cdot)$ is the matrix-valued operation that zeros the input matrix above the diagonal.

As a crude bound, we therefore have

$$\left| \frac{\partial}{\partial C_{rs}} \log \Big(\mathop{\mathrm{E}}_{q(w|(m,C^\top C))} [p(y|w)] \Big) \right| \leq \frac{2B(M+1)r}{\mathrm{E}_{q(w)}[p(y|w)]}.$$

Row r of the Cholesky factor has r entries, so

$$\sum_{s=1}^{r} \left| \frac{\partial}{\partial C_{rs}} \log \Big(\mathop{\mathrm{E}}_{q(w|(m,C^\top C))} [p(y|w)] \Big) \right| \leq \frac{2B(M+1)r^2}{\mathrm{E}_{q(w)}[p(y|w)]},$$

and summing over $r \in \{1, \ldots, M+1\}$ yields

$$\left\| \nabla_{\mathrm{vec}(C)} \log \Big(\mathop{\mathrm{E}}_{q(w|(m,\mathrm{vec}(C)))} p(y|w) \Big) \right\|_1 \leq \frac{BM(M+1)^2(2M+1)}{3\,\mathrm{E}_{q(w)}[p(y|w)]} \leq \frac{BM(M+1)^2(2M+1)}{3\gamma}. \tag{33}$$