# Distribution-Dependent PAC-Bayes Priors

Guy Lever[1], François Laviolette[2], and John Shawe-Taylor[1]

[1] University College London
[2] Université Laval

**Abstract.** We develop the idea that the PAC-Bayes prior can be informed by the data-generating distribution. We prove sharp bounds for an existing framework, and develop insights into function class complexity in this model and suggest means of controlling it with new algorithms. In particular we consider controlling capacity with respect to the *unknown* geometry of the data-generating distribution. We finally extend this localization to more practical learning methods.

## 1 Introduction

The paper takes its inspiration from Catoni (2007), who developed localised PAC-Bayes analysis by using a prior defined in terms of the data generating distribution. At first sight this might appear to be 'cheating', since we must define the prior before seeing the data. However, by defining in terms of the distribution we avoid this difficulty since the distribution itself can be considered as fixed before the sample is generated. PAC-Bayes bounds are one of the sharper analyses of the learning process. A weakness in the standard PAC-Bayes approach is that analysis is constrained by the choice of prior distribution, since the divergence between prior and posterior forms part of the bound. This choice of prior tends to be rather generic; typically not tailored to the particular problem, so that, in particular, good classifiers do not generally receive large prior weight. Thus the divergence term in the PAC-Bayes analysis can typically be large. By tuning the prior to the distribution Catoni is able to remove the Kullback-Leibler (KL) term from the bound hence significantly reducing the complexity penalty.

The paper begins by using Catoni's definition of the prior involving a Boltzmann distribution, but proves a new sharp bound (Theorem 3) using a new lemma (Lemma 1) and the re-use of the PAC-Bayes bound to remove the KL term (Lemma 2). The resulting bound suggests a new complexity parameter $\gamma$ that enters as a $\gamma/m^{3/2}$ term (where $m$ is the sample size). This opens a potential new direction in generalization analysis of learning machines.

In our context this suggests the need to regularize in this learning method. The flexibility of the framework we develop is that it allows us to encode our prior meta-assumptions about how we anticipate a good classifier will interact with the data; we can control capacity, for example, with respect to the smoothness of a classifier over the *unknown* data generating distribution thus giving high weight to classifiers that are smooth over the manifold defined by the support of the data distribution. The analysis is achieved with a novel PAC-Bayes bound on U-statistics estimation.

The final section of the paper covers the third main theme which is the extension of the data distribution dependent priors to the Gaussian prior and posterior PAC-Bayes

M. Hutter et al. (Eds.): ALT 2010, LNAI 6331, pp. 119–133, 2010.

bounds for, for example, SVMs (Langford & Shawe-taylor, 2002). Here we are able to remove the KL term again leaving a term that only involves a similar complexity parameter $\gamma$ appearing as $O(\gamma/(\eta^2 m^2))$, where $\eta$ is the regularization parameter, in contrast to the usual $O(\|\mathbf{w}\|^2/m)$. This again suggests a new measure of complexity for SVM classifiers with the possibility of using the bound to optimise the regularization parameter.

We now review the relation of our approach to earlier work. The luckiness framework explored the possibility that we could learn the hierarchy of classes of hypotheses from the data as part of the learning process giving rise to so-called data-dependent structural risk minimization (Shawe-Taylor et al., 1998). The most successful example of this approach was large margin classification such as support vector machines. However, although we cannot measure a margin without seeing the data, by moving to real-valued functions, we can equate large margin with small norm when we constrain $y_i f(\mathbf{x}_i) \geq 1$ on the training data, $i = 1, \ldots, m$, resulting in a fixed prior. Nonetheless this is equivalent to placing a prior over the classifiers in terms of the data generating distribution, that is we favour hyperplanes that have low input density in the slab defined by shifting the decision boundary parallel to itself by $\pm\gamma$.

Further research in this direction has been developed by Balcan and Blum (Balcan & Blum, 2010) with their notion of compatibility, which is used to restrict the hypotheses considered in the learning process to those satisfying a given level of compatibility *estimated from the training data*, hence reducing the effective complexity of the class. Perhaps less well-known is work by Catoni (Catoni, 2007) where he introduces 'localised' PAC-Bayes analysis effectively defining the prior in terms of the data-generating distribution in a PAC-Bayes bound on generalization.

We should finally distinguish between distribution defined priors and using part of the data to learn a prior and the rest to learn the function (Ambroladze et al., 2006).

## 2 Preliminaries

We consider the general setting in which we are given a sample of labelled and unlabelled[1] points $\mathcal{S} = \{(X_1, Y_1), ...(X_m, Y_m)\} \cup \{X_{m+1}, ...X_n\}$ drawn i.i.d. from distribution $D$ (with density $d$) over $\mathcal{X} \times \mathcal{Y}$, the product space of labelled inputs (or its marginalization to $\mathcal{X}$). Our analysis therefore includes the settings of supervised and semi-supervised learning and some transductive settings. We are interested in the case where $\mathcal{Y} = \{-1, +1\}$, and study binary classification.

We are interested in the notion of *risk* of a hypothesis $h \in \mathcal{H}$,

$$\mathrm{risk}^{\ell}(h) := \mathbb{E}_{(X,Y)\sim D}\ell(h(X), Y),$$

and its empirical counterpart on a labelled sample $\mathcal{S}$,

$$\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h) := \frac{1}{|\mathcal{S}|}\sum_{(X,Y)\in\mathcal{S}} \ell(h(X), Y),$$

[1] Throughout, the unlabelled set may be empty.

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some loss function. When $\ell(\cdot,\cdot)$ is the $0-1$ loss of binary classification, $\ell_{0-1}(y,y') := \begin{cases} 0 \text{ if } y = y' \\ 1 \text{ if } y \neq y' \end{cases}$, then for simplicity we denote the corresponding binary classification risk and its empirical counterpart by $\operatorname{risk}(\cdot)$ and $\widehat{\operatorname{risk}}_{\mathcal{S}}(\cdot)$ respectively. Our objective is to obtain a probabilistic guarantee on the true binary classification risk of a classifier by relating it to its empirical counterpart.

The PAC-Bayes bounds apply to a stochastic Gibbs classifier $G_Q$ drawn from a distribution $Q$ over a hypothesis class $\mathcal{H}$. We denote $\operatorname{risk}(G_Q) := \mathbb{E}_{h\sim Q}\operatorname{risk}(h)$. The following quantities feature in these bounds: the Kullback-Leibler divergence between distributions $Q$ and $P$, and its specialization to Bernouilli distributions,

$$\operatorname{KL}(Q||P) := \mathbb{E}_{h\sim Q} \ln \frac{dQ}{dP}(h), \quad \operatorname{kl}(q,p) := q \ln \frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \quad q,p \in (0,1),$$

and we define

$$\xi(m) := \sum_{k=0}^{m} \binom{m}{k} \left(\frac{k}{m}\right)^{k} \left(1-\frac{k}{m}\right)^{m-k} \in [\sqrt{m}, 2\sqrt{m}].$$

The following is a generalization of (Germain et al., 2009, Th 2.1) and is proved using the same sequence of arguments.

**Theorem 1.** *For any functions $A(h)$, $B(h)$ over $\mathcal{H}$, either of which may be a statistic of the sample $\mathcal{S}$, any distributions $P$ over $\mathcal{H}$, any $\delta \in (0,1]$, any $t > 0$, and a convex function $\mathcal{D} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we have with probability at least $1-\delta$ over the draw of $\mathcal{S}$,*

$$\forall\, Q \,\text{on}\, \mathcal{H} : \mathcal{D}(\mathbb{E}_{h\sim Q}A(h), \mathbb{E}_{h\sim Q}B(h)) \leq \frac{1}{t}\left(\operatorname{KL}(Q||P) + \ln\left[\frac{\mathcal{L}_P}{\delta}\right]\right),$$

*where $\mathcal{L}_P := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{h\sim P}\left[e^{t\mathcal{D}(A(h),B(h))}\right]$.*

Theorem 1 is a recipe for generating a variety of PAC-Bayes bounds, by specializing to a choice for $\mathcal{D}(\cdot,\cdot)$, $t$, $A(\cdot)$ and $B(\cdot)$, and choosing $P$ to be a "prior" (i.e. not sample-dependent) so that the order of expectation in the r.h.s. can be exchanged and evaluated. For example, one can derive Seeger's bound and a slightly relaxed version of Catoni's bound, which will be needed later:

**Theorem 2.** *(Seeger, 2002; Langford, 2005; Catoni, 2007) For any (unknown) distribution $D$, any set $\mathcal{H}$ of classifiers, any distribution $P$ of support $\mathcal{H}$, any $\delta \in (0,1]$, and any positive constant $C$, we have, where $C^\star := \frac{C}{1-e^{-C}}$,*

$$\mathbb{P}_{\mathcal{S}}\left(\forall\, Q \,\text{on}\, \mathcal{H} : \operatorname{kl}(\widehat{\operatorname{risk}}_{\mathcal{S}}(G_Q), \operatorname{risk}(G_Q)) \leq \tfrac{1}{m}\left(\operatorname{KL}(Q||P) + \ln\tfrac{\xi(m)}{\delta}\right)\right) \geq 1-\delta$$
$$\mathbb{P}_{\mathcal{S}}\left(\forall\, Q \,\text{on}\, \mathcal{H} : \operatorname{risk}(G_Q) \leq C^\star\left(\widehat{\operatorname{risk}}_{\mathcal{S}}(G_Q) + \tfrac{1}{C\cdot m}\left(\operatorname{KL}(Q||P) + \ln\tfrac{1}{\delta}\right)\right)\right) \geq 1-\delta$$

Note that the PAC-Bayes bound proposed by McAllester in his pioneer work on the subject (McAllester, 1999) can be retrieved from Seeger's bound using the inequality

$$2(q-p)^2 \leq \operatorname{kl}(q,p). \tag{1}$$

Note also that the $\mathrm{KL}(Q||P)$ term can significantly deteriorate these bound if classifiers with small empirical risk receive low probability from the prior $P$, i.e. if the prior has been "badly" chosen. The data distribution-defined priors we investigate are specifically constructed to give large weight to classifiers with low true risk, and the KL-divergence between $Q$ and $P$ decays with the sample size.

### 2.1 Choosing a Distribution-Dependent Prior

Suppose an algorithm takes as input a training sample $\mathcal{S}$ from the distribution $D^m$ over $\mathcal{Z}^m$ and outputs a posterior distribution $Q$ over $\mathcal{H}$. Let $\mathcal{P}_{\mathcal{H}}$ be the space of probability distributions over $\mathcal{H}$, and in the interest of obtaining a good PAC-Bayes bound for $Q$, consider the minimizer of $\mathrm{KL}(Q||P)$ *in expectation*: $\mathrm{argmin}_{P\in\mathcal{P}_{\mathcal{H}}} \mathbb{E}_{\mathcal{S}}[\mathrm{KL}(Q||P)] = \mathbb{E}_{\mathcal{S}}[Q]$. The implication of this result is noted in this context in (Catoni, 2007) as is the immediate fact that the resulting divergence is equal to the mutual information, $I(h;\mathcal{S})$, between sample and classifier (considered as random variables drawn from the distribution $Q\times D^m$), $\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(Q||\mathbb{E}_{\mathcal{S}}[Q])] = I(h;\mathcal{S})$. We note that PAC-Bayes analysis using this prior appears to be quite difficult since the prior can be difficult to manipulate. As suggested by Catoni we study other more flexible choices of prior which enable us to obtain very sharp PAC-Bayes bounds. We assume that there exists a reference measure $\mu$ on $\mathcal{H}$ (when $\mathcal{H}$ is of finite dimensionality this would typically be a uniform measure such as Lebesgue measure). We consider the case when the posterior and prior are from the exponential family, defined by their densities w.r.t. $\mu$,

$$q(h) := \frac{dQ}{d\mu}(h) := \frac{1}{Z}e^{-F_Q(h)} \qquad\qquad p(h) := \frac{dP}{d\mu}(h) := \frac{1}{Z'}e^{-F_P(h)},$$

where $F_Q$, $F_P$ are "energy functions", to be chosen, and $Z = \int_{\mathcal{H}} e^{-F_Q(h)}\mathrm{d}\mu$, $Z' = \int_{\mathcal{H}} e^{-F_P(h)}\mathrm{d}\mu$. We note the following upper bound on the KL divergence, which reduces obtaining a bound on the KL divergence to establishing a PAC-Bayesian concentration result for the energy functions.

**Lemma 1**

$$\mathrm{KL}(Q||P) \leq \left(\mathbb{E}_{h\sim Q} - \mathbb{E}_{h\sim P}\right)\left[F_P(h) - F_Q(h)\right]. \tag{2}$$

*Proof*

$$\begin{aligned}\mathrm{KL}(Q||P) &= \mathbb{E}_{h\sim Q}\ln\frac{Z'e^{-F_Q(h)}}{Ze^{-F_P(h)}}\\ &= \mathbb{E}_{h\sim Q}\left[F_P(h) - F_Q(h)\right] - \ln\frac{\int e^{-F_Q(h)}\mathrm{d}\mu}{Z'}\\ &= \mathbb{E}_{h\sim Q}\left[F_P(h) - F_Q(h)\right] - \ln\int p(h)e^{F_P(h)-F_Q(h)}\mathrm{d}\mu\\ &\leq \left(\mathbb{E}_{h\sim Q} - \mathbb{E}_{h\sim P}\right)\left[F_P(h) - F_Q(h)\right], \end{aligned} \tag{3}$$

where the final line follows from the convexity of $-\ln(\cdot)$. □

Note that the r.h.s. of (3) is precisely the type of quantity that PAC-Bayes theory provides a bound for. In particular, the choice $F_P = \mathbb{E}_{\mathcal{S}}[F_Q]$ seems natural and we remark that (3) is then reduced to obtaining a concentration inequality for $F_Q$ to its expectation.

## 3 Stochastic Empirical Risk Minimization-Type Prediction

We first consider posterior and prior densities, w.r.t. $\mu$, over $\mathcal{H}$ of the following forms:

$$q(h) = \frac{1}{Z} e^{-\left(\gamma\widehat{\text{risk}}_{\mathcal{S}}(h)+\eta F_Q(h)\right)} \tag{4}$$

$$p(h) = \frac{1}{Z'} e^{-(\gamma\text{risk}(h)+\eta F_P(h))}. \tag{5}$$

where the $F : \mathcal{H} \to \mathbb{R}$ are regularization functions, and $Z$ a normalization constant. The unregularized case corresponds to "Gibbs algorithms", e.g. (Catoni, 2007). $F_Q(\cdot)$ and $F_P(\cdot)$ may be different and in particular we will consider the special case where $F_Q(\cdot)$ is a sample statistic, allowing us to perform data-dependent regularization.

We note that Lemma 1 implies the following upper bound on the KL divergence

$$\text{KL}(Q||P) \leq (\mathbb{E}_{h\sim Q} - \mathbb{E}_{h\sim P}) \left[\gamma\text{risk}(h) - \gamma\widehat{\text{risk}}_{\mathcal{S}}(h) + \eta F_P(h) - \eta F_Q(h)\right]. \tag{6}$$

As we will see later, for suitable choices of parameters $\gamma$ and $\eta$, the divergence decays with the sample. We now consider several choices of $F_Q(\cdot)$ and $F_P(\cdot)$ and give PAC-Bayes bounds for the resulting Gibbs classifiers.

### 3.1 The Non-regularized Case: $\eta = 0$

We recall that the distribution $D$ over $\mathcal{X} \times \mathcal{Y}$ is unknown, hence so is the prior distribution given by (5). To obtain a bound, we need to bound the KL divergence $\text{KL}(Q||P)$. With reference to (6), in the situation where $\eta = 0$ such an upper bound can be obtained given an upper bound for $\text{risk}(G_Q) - \widehat{\text{risk}}_{\mathcal{S}}(G_Q)$ and a lower bound for $\text{risk}(G_P) - \widehat{\text{risk}}_{\mathcal{S}}(G_P)$, and such bounds can obtained via Theorem 2.

**Lemma 2.** *Let $P$ and $Q$ be defined as in (4) and (5) with $\eta = 0$ then with probability at least $1 - \delta$, the following holds,*

$$\text{KL}(Q||P) \leq \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{\xi(m)}{\delta}} + \frac{\gamma^2}{4m}.$$

*Proof.* From (1) and the Seeger bound of Theorem 2 (applied for the choices $Q = Q$ and $Q = P$) we obtain that, simultaneously,

$$\begin{aligned} \text{risk}(G_Q) - \widehat{\text{risk}}_{\mathcal{S}}(G_Q) &\leq \frac{1}{2\sqrt{m}}\sqrt{\text{KL}(Q||P) + \ln\frac{\xi(m)}{\delta}}, \\ -\left(\text{risk}(G_P) - \widehat{\text{risk}}_{\mathcal{S}}(G_P)\right) &\leq \frac{1}{2\sqrt{m}}\sqrt{\ln\frac{\xi(m)}{\delta}}. \end{aligned}$$

Together with (6) the last inequalities give

$$\begin{aligned}\mathrm{KL}(Q||P) &\le \gamma\left(\mathrm{risk}(G_Q)-\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q)\right) - \gamma\left(\mathrm{risk}(G_P)-\widehat{\mathrm{risk}}_{\mathcal{S}}(G_P)\right)\\ &\le \frac{\gamma}{2\sqrt{m}}\sqrt{\mathrm{KL}(Q||P)+\ln\frac{\xi(m)}{\delta}} + \frac{\gamma}{2\sqrt{m}}\sqrt{\ln\frac{\xi(m)}{\delta}}\,.\end{aligned}$$

If $\mathrm{KL}(Q||P) \le \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{\xi(m)}{\delta}}$, we are done. Otherwise, by straightforward algebraic manipulations we then obtain the following inequality, which, together with the fact that $\mathrm{KL}(Q||P) \ge 0$, directly implies the result.

$$\begin{aligned}(\mathrm{KL}(Q||P))^2 - \frac{2\gamma}{2\sqrt{m}}\sqrt{\ln\frac{\xi(m)}{\delta}}\,\mathrm{KL}(Q||P) + \frac{\gamma^2}{4m}\ln\frac{\xi(m)}{\delta}&\\ \le \frac{\gamma^2}{4m}\mathrm{KL}(Q||P) + \frac{\gamma^2}{4m}\ln\frac{\xi(m)}{\delta}&.\end{aligned}$$

□

Thus, Theorem 2 can be specialized to the following bound. (Note, for the first result no union bound is required since we need to apply Theorem 2 once only.)

**Theorem 3.** *Let $P$ and $Q$ be defined as in (4) and (5) with $\eta = 0$, then*

$$\mathbb{P}_{\mathcal{S}}\left(\mathrm{kl}(\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q),\mathrm{risk}(G_Q)) \le \frac{1}{m}\left(\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{\xi(m)}{\delta}}+\frac{\gamma^2}{4m}+\ln\frac{\xi(m)}{\delta}\right)\right) \ge 1-\delta,$$

$$\mathbb{P}_{\mathcal{S}}\left(\mathrm{risk}(G_Q) \le C^\star\,\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q) + \frac{C^\star}{C\cdot m}\left(\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta}}+\frac{\gamma^2}{4m}+\ln\frac{2}{\delta}\right)\right) \ge 1-\delta$$

Observe that for a large value of $\gamma$, the posterior Gibbs classifier $G_Q$ will be concentrated on the classifiers of $\mathcal{H}$ with smallest empirical risk. Hence the two bounds of Theorem 3 are risk bounds for a type of stochastic empirical risk minimization algorithm. Since the KL-divergence term has been evaluated and is small, it appears that there is no component of the bound that depends on the complexity of the learning problem or the class of classifiers. In fact the parameter that controls the effective complexity is the "inverse temperature", $\gamma$ (or $\gamma^2$ if we view it in the role of a VC dimension). If the problem is 'easy' in the sense that the measure of the set of classifiers with low empirical risk is not too small then a low value of $\gamma$ will deliver low empirical risk for the Gibbs classifier. If, however, the measure of the classifiers that have low empirical risk is very small (as would be likely if the function class itself is large) then we require a larger value of $\gamma$ before the Gibbs risk is controlled. The complexity that $\gamma$ measures is related to the fit between input distribution and function class in that it will depend on the measure of the distribution $Q$ on the low empirical risk functions.

In practice $\gamma$ would need to be chosen from a grid $\Gamma$ of values in response to the particular training problem. Hence, in order to apply the bound we would need to use the union bound over the $|\Gamma|$ applications of the bound resulting in an extra $\log(|\Gamma|)$ term in the right hand side brackets. Another possibility would be to make use the generalized union bound known as Occam's hammer (Blanchard & Fleuret, 2007).

### 3.2 Regularization with $F_Q(\cdot) = F_P(\cdot)$

Given the above argument it appears necessary to control function class capacity in this model in order to deliver low empirical Gibbs risk. We therefore consider the presence of regularization terms in (4), (5) which encode a preference for classifiers which satisfy some notion of simplicity. The flexibility of this model is such that, with reference to (6), when $F_Q(\cdot) = F_P(\cdot)$, the bounds of Theorem 3 hold for this case. We can therefore apply arbitrary (non data-dependent) regularization and attain the same bound of Theorem 3, and there are many natural possibilities. For example, if $\mathcal{H}$ is equipped with a norm $||\cdot||_{\mathcal{H}}$ we can choose $F_Q(\cdot) = F_P(\cdot) = ||\cdot||_{\mathcal{H}}$. This should permit learning with smaller $\gamma$.

### 3.3 Regularization in the Intrinsic Data Geometry

The flexibility of this model further permits, in a straightforward way, regularization w.r.t. the geometry defined by the *unknown* data-generating distribution, and we detail one way of achieving this. The regularization methods considered in Section 3.2 utilise a geometry which is extrinsic to the data, that is, determined by the ambient representation space rather than the intrinsic geometry of data. For example, if the data generating distribution has support on some submanifold of the ambient space, then encouraging smoothness on the manifold ought to be more suitable for learning (since if the structure of data is a key factor in the learnability of a task, it is the intrinsic geometry which will capture this relevant structure most accurately). In general, when using a regularizer informed by the intrinsic geometry of the data-generating distribution the prior and posterior regularizers must be different since the posterior regularizer will be an empirical quantity (here, chosen to be an estimate, based on the sample, of the prior regularizer).

Given a sample $\mathcal{S} = \{(X_1, Y_1), ...(X_m, Y_m)\} \cup \{X_{m+1}, ...X_n\}$, we consider regularizing via the following *smoothness functional* (typical in semi-supervised learning e.g. (Belkin et al., 2004; Zhu et al., 2003)) over functions from some function class $\mathcal{H}$:

$$\widehat{U}_{\mathcal{S}}(h) := \frac{1}{n(n-1)} \sum_{ij} (h(X_i) - h(X_j))^2 W(X_i, X_j) \tag{7}$$

where the symmetric $W : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ captures similarity or "weight" between data points, for example $W(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} 1 \text{ if } \boldsymbol{x}, \boldsymbol{x}' \text{ are a pair of } k\text{−nearest neighbours} \\ 0 \text{ otherwise} \end{cases}$ or $W(\boldsymbol{x}, \boldsymbol{x}') = e^{-||\boldsymbol{x}-\boldsymbol{x}'||^2}$ for some norm $||\cdot||$ over $\mathcal{X}$. Note that $\widehat{U}_{\mathcal{S}}(h) = \frac{2}{n(n-1)} \boldsymbol{h}^\top \boldsymbol{L} \boldsymbol{h}$ where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the graph Laplacian of a graph $\mathcal{G}$ whose vertices are the sample instances and whose *edge weights* are controlled by $W$, and $D_{ij} = \delta_{ij} \sum_k W_{ik}$ and where $\boldsymbol{h} \in \mathbb{R}^n$ is the "point evaluation" of $h$ on the sample, $h_i := h(\boldsymbol{x}_i)$. Minimizing (7) encourages functions to be smooth over the sample $\mathcal{S}$. Note that $\widehat{U}_{\mathcal{S}}(h)$ is a $U$-statistic of order 2 with *kernel* $f_h(X_i, X_j) := (h(X_i) - h(X_j))^2 W(X_i, X_j)$ indexed by $\mathcal{H}$. A family of $U$-statistics indexed by a function space is often called a $U$-process. We suppose that the weights are bounded, $|W(\boldsymbol{x}, \boldsymbol{x}')| \leq w$, for example if $W(\boldsymbol{x}, \boldsymbol{x}') = e^{-||\boldsymbol{x}-\boldsymbol{x}'||^2}$ we have $w = 1$, and that $\sup_{h \in \mathcal{H}, \boldsymbol{x} \in \mathcal{X}} |h(\boldsymbol{x})| = b$.

A series of results (Hein et al., 2006) demonstrate that under certain conditions on the distribution of instances, certain constructions of graph Laplacian converge to a generalized Laplace operator on the support of the data generating distribution and the smoothness functional converges to a natural distribution-dependent measure of smoothness over functions defined over the data.

We choose $F_Q(\cdot) = \widehat{U}_{\mathcal{S}}(\cdot)$ so that,

$$q(h) = \frac{1}{Z} e^{-\left(\gamma \widehat{\mathrm{risk}}_{\mathcal{S}}(h) + \eta \widehat{U}_{\mathcal{S}}(h)\right)}. \tag{8}$$

The exponent simply minimizes empirical risk plus the smoothness on the graph formed on the sample, as is a typical methodology in semi-supervised learning (Belkin et al., 2006; Belkin et al., 2004).

We further choose $F_P(h) = U(h) := \mathbb{E}_{X,X'}[(h(X) - h(X'))^2 W(X,X')] = \mathbb{E}_{\mathcal{S}}[\widehat{U}_{\mathcal{S}}(h)]$, giving the prior $p(h) = \frac{1}{Z'} e^{-(\gamma \mathrm{risk}(h) + \eta U(h))}$.

**Convergence of the smoothness functional.** We consider PAC-Bayes convergence of the U-process (see (Ralaivola et al., 2009) for an alternative PAC-Bayes analysis of $U$-statistics). Let $\mathcal{S} = \{X_1, ... X_n\}$ be an i.i.d. sample. For any second-order $U$-statistic $\widehat{U}_{\mathcal{S}}(h) = \frac{1}{n(n-1)} \sum_{i \neq j} f_h(X_i, X_j)$ with expectation $U(h)$, and with kernel $f_h(x, x')$ indexed by $\mathcal{H}$ and bounded, $a \leq f_h(x, x') \leq b$, we have the following.

**Theorem 4.** *For all t, any prior P and simultaneously for all posteriors Q over $\mathcal{H}$,*

$$\mathbb{P}_{\mathcal{S}}\left(\mathbb{E}_{h\sim Q}[\widehat{U}_{\mathcal{S}}(h) - U(h)] \leq \frac{1}{t}\left(\mathrm{KL}(Q||P) + \frac{t^2(b-a)^2}{2n} + \ln\left(\frac{1}{\delta}\right)\right)\right) \geq 1 - \delta \tag{9}$$

$$\mathbb{P}_{\mathcal{S}}\left(\mathbb{E}_{h\sim Q}[U(h) - \widehat{U}_{\mathcal{S}}(h)] \leq \frac{1}{t}\left(\mathrm{KL}(Q||P) + \frac{t^2(b-a)^2}{2n} + \ln\left(\frac{1}{\delta}\right)\right)\right) \geq 1 - \delta. \tag{10}$$

*In particular, choosing $t = \sqrt{n}$ gives $\mathcal{O}(\frac{1}{\sqrt{n}})$ convergence.*

*Proof.* We note that Theorem 1 implies that with probability at least $1 - \delta$, $\forall Q$ on $\mathcal{H}$:

$$\mathbb{E}_{h\sim Q}[\widehat{U}_{\mathcal{S}}(h) - U(h)] \leq \frac{1}{t}\left(\mathrm{KL}(Q||P) + \ln\left(\frac{1}{\delta}\mathbb{E}_{h\sim P}\mathbb{E}_{\mathcal{S}}\left[e^{t(\widehat{U}_{\mathcal{S}}(h) - U(h))}\right]\right)\right),$$

so we simply need to bound $\mathbb{E}_{\mathcal{S}}\left[e^{t(\widehat{U}_{\mathcal{S}}(h) - U(h))}\right]$. Employing Hoeffding's canonical decomposition of $U$-statistics into forward martingales (e.g. (Serfling, 1980)), let,

$$V_k := \sum_{i=1}^{k} \left(\mathbb{E}[f_h(X_i, X) \mid X_i] - U(h)\right)$$

$$W_k := \sum_{j=1}^{k} \sum_{i=1}^{j-1} \left(f_h(X_i, X_j) + U(h) - \mathbb{E}[f_h(X_i, X) \mid X_i] - \mathbb{E}[f_h(X, X_j) \mid X_j]\right),$$

so that, $\widehat{U}_{\mathcal{S}}(h) - U(h) = \frac{2}{n} V_n + \frac{2}{n(n-1)} W_n$. We then have that

$$\begin{aligned} V_k - V_{k-1} &= \mathbb{E}[f_h(X_k, X \mid X_k)] - U(h) \\ W_k - W_{k-1} &= \sum_{i=1}^{k-1} f_h(X_i, X_k) + U(h) - \mathbb{E}[f_h(X_i, X) \mid X_i] - \mathbb{E}[f_h(X_k, X) \mid X_k], \end{aligned}$$

and note the martingale structure $\mathbb{E}_{X_k}[V_k - V_{k-1}] = \mathbb{E}_{X_k}[W_k - W_{k-1}] = 0$. Note further that,

$$\begin{aligned} V_k - V_{k-1} + \frac{1}{n-1}(W_k - W_{k-1}) = &\frac{n-k}{n-1}(\mathbb{E}[f_h(X_k, X) \mid X_k] - U(h)) \\ &+ \frac{1}{n-1} \sum_{i=1}^{k-1} f_h(X_i, X_k) - \mathbb{E}[f_h(X_i, X) \mid X_i], \end{aligned}$$

so that,

$$|V_k - V_{k-1} + \frac{1}{n-1}(W_k - W_{k-1})| \leq (b-a)\frac{n-k}{n-1} + (b-a)\frac{k-1}{n-1} = b-a. \quad (11)$$

Now,

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}}\left[e^{t(\widehat{U}_{\mathcal{S}}(h) - U(h))}\right] = \mathbb{E}_{\mathcal{S}}\left[e^{\frac{2t}{n}\sum_{i=1}^n V_i - V_{i-1} + \frac{1}{n-1}(W_i - W_{i-1})}\right] \\ =&\mathbb{E}_{X_1,\ldots X_{n-1}}\left[\mathbb{E}_{X_n}\left[e^{\frac{2t}{n}\sum_{i=1}^n V_i - V_{i-1} + \frac{1}{n-1}(W_i - W_{i-1})} \mid X_1, \ldots X_{n-1}\right]\right] \\ =&\mathbb{E}_{X_1,\ldots X_{n-1}}\left[e^{\frac{2t}{n}\sum_{i=1}^{n-1} V_i - V_{i-1} + \frac{1}{n-1}(W_i - W_{i-1})}\mathbb{E}_{X_n}\left[e^{\frac{2t}{n}\left(V_n - V_{n-1} + \frac{1}{n-1}(W_n - W_{n-1})\right)}\right]\right] \\ \leq&\mathbb{E}_{X_1,\ldots X_{n-1}}\left[e^{\frac{2t}{n}\sum_{i=1}^{n-1} V_i - V_{i-1} + \frac{1}{n-1}(W_i - W_{i-1})}\right] e^{\frac{t^2(b-a)^2}{2n^2}} \\ &\vdots \\ \leq& \prod_{i=1}^n e^{\frac{t^2(b-a)^2}{2n^2}} \quad = \quad e^{\frac{t^2(b-a)^2}{2n}}, \end{aligned}$$

where in the final lines we used Hoeffding's lemma, Lemma 6, combined with (11) recursively. This proves (9), and (10) follows by a symmetrical argument. □

We can now give the following bound for the classification risk of the Gibbs classifier $G_Q$ drawn from the distribution (8) over $\mathcal{H}$:

**Theorem 5.** *For $\eta < \sqrt{n}$,*

$$\mathbb{P}_{\mathcal{S}}\left(\mathrm{kl}(\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(A^2 + B + A\sqrt{2B + A^2} + \ln\frac{\xi(m)}{\delta}\right)\right) \geq 1 - \delta,$$

*where*

$$\begin{aligned} A &:= \frac{\gamma\sqrt{n}}{2\sqrt{m}(\sqrt{n} - \eta)} = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right) \\ B &:= \frac{\sqrt{n}}{\sqrt{n} - \eta}\left(\gamma\sqrt{\frac{2}{m}\ln\frac{4\xi(m)}{\delta}} + \frac{2\eta}{\sqrt{n}}\left(32b^4w^2 + \ln\frac{4}{\delta}\right)\right) = \mathcal{O}\left(\sqrt{\frac{\ln m}{m}}\right). \end{aligned}$$

*Proof.* From (6) we have

$$\begin{aligned}\mathrm{KL}(Q||P) \leq \gamma(\mathrm{risk}(G_Q) - \widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q)) + \gamma(\widehat{\mathrm{risk}}_{\mathcal{S}}(G_P) - \mathrm{risk}(G_P)) \\ + \eta \mathbb{E}_{h\sim Q}\left[U(h) - \widehat{U}_{\mathcal{S}}(h)\right] + \eta \mathbb{E}_{h\sim P}\left[\widehat{U}_{\mathcal{S}}(h) - U(h)\right]. \qquad (12)\end{aligned}$$

And now, recalling (1), and noting that, because $|h(\boldsymbol{x})| \leq b$, $W(\boldsymbol{x}, \boldsymbol{x}') \leq w$, the kernel satisfies $|f_h(\boldsymbol{x}, \boldsymbol{x}')| \leq 4b^2 w$, as in Theorem 3 we apply Seeger's bound of Theorem 2 and Theorem 4 to the relevant terms in (12), and apply the union bound, so that with probability at least $1 - \delta$ over the draw of $\mathcal{S}$,

$$\begin{aligned}\mathrm{KL}(Q||P) &\leq \gamma\sqrt{\frac{1}{2m}\left(\mathrm{KL}(Q||P) + \ln\frac{4\xi(m)}{\delta}\right)} + \gamma\sqrt{\frac{1}{2m}\ln\frac{4\xi(m)}{\delta}} \\ &\quad + \frac{\eta}{\sqrt{n}}\left(\mathrm{KL}(Q||P) + 32b^4w^2 + \ln\frac{4}{\delta}\right) + \frac{\eta}{\sqrt{n}}\left(32b^4w^2 + \ln\frac{4}{\delta}\right) \\ &\leq \gamma\sqrt{\frac{1}{2m}\mathrm{KL}(Q||P)} + \frac{\eta}{\sqrt{n}}\mathrm{KL}(Q||P) + \gamma\sqrt{\frac{2}{m}\ln\frac{4\xi(m)}{\delta}} + \frac{2\eta}{\sqrt{n}}\left(32b^4w^2 + \ln\frac{4}{\delta}\right) \\ &\left(\sqrt{\mathrm{KL}(Q||P)} - \frac{1}{\sqrt{2}}A\right)^2 \leq B + \frac{A^2}{2} \\ &\mathrm{KL}(Q||P) \leq A^2 + B + A\sqrt{2B + A^2},\end{aligned}$$

which we plug into the bound of Theorem 2. □

We remark that the ease with which we can obtain this bound for regularization w.r.t. the geometry defined by the unknown data-generating distribution, with apparently very little deterioration in the bound, is unusual and that in classical frameworks this type of structuring of a function class usually results in significant deterioration in the bound.

## 4 Prediction by RKHS Regularization

We now extend the localization framework to the more practical setting of predicting with a Gaussian process whose mean is the solution to a loss minimization with RKHS regularization, such as an SVM solution. We consider a separable[2] RKHS $\mathcal{H} = \mathrm{span}\{K(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$, for some kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, of functions which are identified as binary classifiers via $h_{\mathrm{class}}(\boldsymbol{x}) = \mathrm{sgn}(h(\boldsymbol{x})) = \mathrm{sgn}(\langle h, K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}})$. For any chosen loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, we are interested in the classifiers,

$$h^*_{\mathcal{S}} := \underset{h\in\mathcal{H}}{\mathrm{argmin}}\{\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h) + \eta||h||^2_{\mathcal{H}}\} \quad \text{and} \quad h^* := \mathbb{E}_{\mathcal{S}}[h^*_{\mathcal{S}}],$$

where $\eta$ is a regularization parameter and expectation is taken with respect to samples $\mathcal{S}$ with $m$ labelled points. For our intended applications, typically $\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(\cdot)$ will be convex so that $h^*_{\mathcal{S}}$ is unique and $h^*$ well-defined. Our posterior $Q$ and prior $P$ will be Gaussian processes on $\mathcal{X}$ with mean $h^*_{\mathcal{S}}$ and $h^*$ respectively. To define this, denote

[2] This is a mild condition, an RKHS $\mathcal{H}$ is separable if $\mathcal{X}$ is and if the kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is continuous.

by $\mathcal{L}^2(\mathcal{X})$ the Hilbert space of square integrable real-valued functions on $\mathcal{X}$ with inner product $\langle h, g\rangle_{\mathcal{L}^2} = \int_{\mathcal{X}} h(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and consider the countable orthonormal basis $\{\phi_i\}$ for $\mathcal{L}^2(\mathcal{X})$ provided by the eigenfunctions of the integral operator $A_K$ defined by $(A_K h)(\boldsymbol{x}) := \int K(\boldsymbol{x}, \boldsymbol{x}')h(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'$, i.e. such that $A_K(\phi_i) = \lambda_i\phi_i$, for eigenvalues $\{\lambda_i\}$ and $\langle\phi_i, \phi_j\rangle_{\mathcal{L}^2} = \delta_{ij}$. Denote $h_i := \langle h, \phi_i\rangle_{\mathcal{L}^2}$ and consider the isomorphism $I : \mathcal{L}^2(\mathcal{X}) \to \ell^2$ given by $I(h) = (h_i)$ identifying $\mathcal{L}^2(\mathcal{X})$ with the space of square summable real-valued sequences. Denote by $N_{h_i,\frac{1}{\gamma}\lambda_i}$ the one-dimensional Gaussian measure on (the Borel $\sigma$-algebra on) $\mathbb{R}$ with mean $h_i$ and variance $\frac{1}{\gamma}\lambda_i$. We then define the product measures[3],

$$Q := \prod_{i=1}^{\infty} N_{(h^*_{\mathcal{S}})_i,\frac{1}{\gamma}\lambda_i} \quad \text{and } P := \prod_{i=1}^{\infty} N_{h^*_i,\frac{1}{\gamma}\lambda_i}. \tag{13}$$

That $Q$ and $P$ define probability measures on $\mathcal{L}^2(\mathcal{X})$ is the subject of (Da Prato, 2006, Chapter 1). A full treatment of this subject requires more space and will be presented in a longer version of this paper (or see (Lever et al., 2010)). To build intuition, when the distributions are of finite dimensionality they have density (w.r.t. Lebesgue measure under the above isomorphism),

$$q(h) = \frac{1}{Z}e^{-\frac{\gamma}{2}||h-h^*_{\mathcal{S}}||^2_{\mathcal{H}}} \text{ and } \qquad p(h) = \frac{1}{Z'}e^{-\frac{\gamma}{2}||h-h^*||^2_{\mathcal{H}}} \tag{14}$$

where, $Z$, $Z'$ enforce normalization. When the dimension of $\mathcal{H}$ is infinite any marginalization to a finite-dimensional linear subspace of $\mathcal{L}^2(\mathcal{X})$ has a similar density. Note that $\gamma$ is a parameter of the algorithm which controls the variance of the Gaussian distribution.

We note, when predicting on a finite set of points, the equivalence between the Gibbs classifier drawn from the posterior (13) and a Gaussian process $\{G_{\boldsymbol{x}}\}_{\boldsymbol{x}\in\mathcal{X}}$ on $\mathcal{X}$ with mean $\mathbb{E}[G_{\boldsymbol{x}}] = h^*_{\mathcal{S}}(\boldsymbol{x})$ and covariance $\mathbb{E}[(G_{\boldsymbol{x}}-\mathbb{E}[G_{\boldsymbol{x}}])(G_{\boldsymbol{x}'}-\mathbb{E}[G_{\boldsymbol{x}'}])] = \frac{1}{\gamma}K(\boldsymbol{x}, \boldsymbol{x}')$.

To obtain a PAC-Bayes bound for the Gibbs classifier drawn from $Q$, we proceed to establish a bound for the relative entropy between $Q$ and $P$. For any Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we denote

$$\kappa(\boldsymbol{x}) := \sup_{h\in\mathcal{H}} \frac{|h(\boldsymbol{x})|}{||h||_{\mathcal{H}}} = \sqrt{K(\boldsymbol{x}, \boldsymbol{x})} \quad \text{and} \quad \kappa := \sup_{\boldsymbol{x}\in\mathcal{X}} \kappa(\boldsymbol{x}),$$

and define the distance $d_K(\boldsymbol{x}, \boldsymbol{x}') := ||K(\boldsymbol{x}, \cdot) - K(\boldsymbol{x}', \cdot)||_{\mathcal{H}}$. Note that $d_K(\boldsymbol{x}, \boldsymbol{x}') \leq 2\kappa$. Our analyses will make use of the following property of a loss function:

**Definition 1.** *(Bousquet & Elisseeff, 2002, Definition 19) $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is $\alpha$-admissible with respect to $\mathcal{H}$ if it is convex in its first argument and for all $y \in \mathcal{Y}$,*

$$|\ell(y_1, y) - \ell(y_2, y)| \leq \alpha|y_1 - y_2|,$$

*for all $y_1$, $y_2$ in the domain of the functions from $\mathcal{H}$.*

[3] These measures are defined on $\mathbb{R}^\infty$ but their support is precisely $\ell^2$, i.e. $\mathcal{L}^2(\mathcal{X})$.

We recall the following definition of Bregman divergence on a Hilbert space $\mathcal{V}$: for differentiable convex $\Phi : \mathcal{V} \rightarrow \mathbb{R}$,

$$d_\Phi(u, v) := \Phi(u) - \Phi(v) - \langle \nabla\Phi(v), u - v\rangle_{\mathcal{V}}. \tag{15}$$

**Lemma 3.** $\mathrm{KL}(Q||P) = \frac{\gamma}{2}||h^*_{\mathcal{S}} - h^*||^2_{\mathcal{H}}$.

*Proof.* When $Q$ and $P$ are finite-dimensional distributions this is the well-known formula for the KL divergence between Gaussians. When the dimensionality is infinite some subtleties are required and due to lack of space we refer the reader to a longer version of this paper (Lever et al., 2010). □

We now proceed to upper bound this divergence via a method of bounded differences: consider a sample $\mathcal{S}$ and its "perturbation" $\mathcal{S}^{(i)}$,

$$\begin{aligned}\mathcal{S} &:= \{(X_1, Y_1), ...(X_m, Y_m)\}\\ \mathcal{S}^{(i)} &:= \{(X_1, Y_1), ...(X_{i-1}, Y_{i-1}), (X'_i, Y'_i), (X_{i+1}, Y_{i+1}), ...(X_m, Y_m)\}.\end{aligned}$$

**Lemma 4.** *If $\ell(\cdot, \cdot)$ is $\alpha$-admissible and differentiable*[4] *then*

$$||h^*_{\mathcal{S}^{(i)}} - h^*_{\mathcal{S}}||_{\mathcal{H}} \leq \frac{\alpha}{2\eta m}(\kappa(X_i) + \kappa(X'_i)). \tag{16}$$

*Proof.* The method of proof is a stability argument which follows (Bousquet & Elisseeff, 2002, Theorem 22). Denote the "objectives"

$$\begin{aligned}\Omega(h) &:= \widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h) + \eta||h||^2_{\mathcal{H}},\\ \Omega^{(i)}(h) &:= \widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}^{(i)}}(h) + \eta||h||^2_{\mathcal{H}}.\end{aligned}$$

Since $\nabla\Omega(h^*_{\mathcal{S}}) = \nabla\Omega^{(i)}(h^*_{\mathcal{S}^{(i)}}) = 0$, we have,

$$\begin{aligned}d_\Omega(h^*_{\mathcal{S}^{(i)}}, h^*_{\mathcal{S}}) + d_{\Omega^{(i)}}(h^*_{\mathcal{S}}, h^*_{\mathcal{S}^{(i)}}) &= \Omega(h^*_{\mathcal{S}^{(i)}}) - \Omega(h^*_{\mathcal{S}}) + \Omega^{(i)}(h^*_{\mathcal{S}}) - \Omega^{(i)}(h^*_{\mathcal{S}^{(i)}})\\ &= \frac{1}{m}(\ell(h^*_{\mathcal{S}^{(i)}}(X_i), Y_i) - \ell(h^*_{\mathcal{S}^{(i)}}(X'_i), Y'_i)\\ &\quad +\ell(h^*_{\mathcal{S}}(X'_i), Y'_i) - \ell(h^*_{\mathcal{S}}(X_i), Y_i)).\end{aligned}$$

Noting the additivity, $d_{\Phi+\Psi} = d_\Phi + d_\Psi$, and non-negativity of Bregman divergences and that $d_{\eta||\cdot||^2_{\mathcal{H}}}(h, g) = \eta||h - g||^2_{\mathcal{H}}$ we have,

$$\begin{aligned}&2\eta||h^*_{\mathcal{S}} - h^*_{\mathcal{S}^{(i)}}||^2_{\mathcal{H}}\\ &\quad\leq \frac{1}{m}\big(\ell(h^*_{\mathcal{S}^{(i)}}(X_i), Y_i) - \ell(h^*_{\mathcal{S}^{(i)}}(X'_i), Y'_i) + \ell(h^*_{\mathcal{S}}(X'_i), Y'_i) - \ell(h^*_{\mathcal{S}}(X_i), Y_i)\big)\\ &\quad\leq \frac{\alpha}{m}(|h^*_{\mathcal{S}}(X_i) - h^*_{\mathcal{S}^{(i)}}(X_i)| + |h^*_{\mathcal{S}}(X'_i) - h^*_{\mathcal{S}^{(i)}}(X'_i)|)\\ &\quad\leq \frac{\alpha}{m}(||h^*_{\mathcal{S}} - h^*_{\mathcal{S}^{(i)}}||_{\mathcal{H}}(\kappa(X_i) + \kappa(X'_i))).\end{aligned}$$

□

[4] We note that for the case of the hinge loss or absolute loss this condition can be relaxed – we can define the derivative to be zero at the point at which they are non-differentiable. For general subdifferentiable convex loss functions we recover the results if we define the gradient to be zero at the minimum.

**Lemma 5.** *If $\ell(\cdot,\cdot)$ is $\alpha$-admissible, differentiable*[4] *and $\mathcal{H}$ is separable then*

$$\mathbb{P}_{\mathcal{S}}\left(||h^*_{\mathcal{S}} - h^*||_{\mathcal{H}} \leq \frac{2\alpha\kappa}{\eta}\sqrt{\frac{1}{m}\ln\frac{4}{\delta}}\right) \geq 1-\delta. \tag{17}$$

*Proof.* Define the Doob martingale,

$$V_i = \mathbb{E}[h^*_{\mathcal{S}} - h^* \mid (X_1, Y_1), ...(X_i, Y_i)],$$

and note that $V_0 = 0$, $V_m = h^*_{\mathcal{S}} - h^*$, and that

$$\begin{aligned}\mathbb{E}[V_i \mid (X_1, Y_1), ...(X_{i-1}, Y_{i-1})] &= \mathbb{E}[h^*_{\mathcal{S}} - h^* \mid (X_1, Y_1), ...(X_{i-1}, Y_{i-1})] \\ &= V_{i-1}.\end{aligned}$$

Thus $\{V_i\}_{i=1}^m$ is a martingale and we have further, by Lemma 4, that

$$\begin{aligned}||V_i - V_{i-1}||_{\mathcal{H}} &= ||\mathbb{E}[h^*_{\mathcal{S}} \mid (X_1, Y_1), ...(X_i, Y_i)] - \mathbb{E}[h^*_{\mathcal{S}} \mid (X_1, Y_1), ...(X_{i-1}, Y_{i-1})]||_{\mathcal{H}} \\ &\leq \frac{\kappa\alpha}{\eta m}.\end{aligned}$$

Since $\mathcal{H}$ is separable it has a countable basis and so is isomorphic to either $\ell^2(\mathbb{R})$ or $\mathbb{R}^d$ and the result follows from the result of (Kallenberg & Sztencel, 1991, Theorem 3.1) (which gives a version of Azuma's inequality for $\ell^2$-valued martingales, see the details in Theorem 7 and Corollary 1 of the Appendix). □

We can now give the PAC-Bayes bound for the classification risk of the Gibbs classifier, $G_Q$, drawn from $\mathcal{H}$ according to the distribution $Q$ defined by (13).

**Theorem 6.** *If $\ell(\cdot,\cdot)$ is $\alpha$-admissible, differentiable*[4] *and $\mathcal{H}$ is separable then*

$$\mathbb{P}_{\mathcal{S}}\left(\mathrm{kl}(\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(\frac{2\gamma\alpha^2\kappa^2}{\eta^2 m}\ln\frac{8}{\delta} + \ln\frac{2\xi(m)}{\delta}\right)\right) \geq 1-\delta.$$

*Proof.* Lemma 3 and Lemma 5 immediately imply that,

$$\mathbb{P}_{\mathcal{S}}\left(KL(Q||P) \leq \frac{2\gamma\alpha^2\kappa^2}{\eta^2 m}\ln\frac{8}{\delta}\right) \geq 1-\frac{\delta}{2},$$

which we combine with Theorem 2 using the union bound. □

Note that the PAC-Bayes bounds for Gibbs classifiers presented here will provide sharp bounds on the mean classifier (which, with suitable choices for parameters, could be various types of SVM), with an additional factor of $1+\epsilon$, under a margin assumption, by standard techniques (Langford & Shawe-taylor, 2002). We also remark that it is straightforward to extend the analysis presented here to provide bounds for classifiers obtained by regularizing in the geometry defined by the data, in the manner of Section 3.3, such as LapSVM, and refer the reader to an extended version of this paper (Lever et al., 2010).

## References

Ambroladze, A., Parrado-Hernández, E., Shawe-Taylor, J.: Tighter pac-bayes bounds. In: NIPS, pp. 9–16. MIT Press, Cambridge (2006)

Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Mathematical Journal 68, 357–367 (1967)

Balcan, M., Blum, A.: A discriminative model for semi-supervised learning. JACM, 57 (2010)

Belkin, M., Matveeva, I., Niyogi, P.: Regularization and semi-supervised learning on large graphs. In: Shawe-Taylor, J., Singer, Y. (eds.) COLT 2004. LNCS (LNAI), vol. 3120, pp. 624–638. Springer, Heidelberg (2004)

Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434 (2006)

Blanchard, G., Fleuret, F.: Occam's hammer. In: Bshouty, N.H., Gentile, C. (eds.) COLT. LNCS (LNAI), vol. 4539, pp. 112–126. Springer, Heidelberg (2007)

Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Learn. Res. 2, 499–526 (2002)

Catoni, O.: PAC-Bayesian surpevised classification: the thermodynamics of statistical learning. Monograph Series of the Institute of Mathematical Statistics (2007)

Da Prato, G.: An introduction to infinite-dimensional analysis. Springer, Heidelberg (2006)

Germain, P., Lacasse, A., Laviolette, F., Marchand, M.: Pac-bayesian learning of linear classifiers. In: ICML, p. 45. ACM, New York (2009)

Hein, M., Audibert, J.-Y., von Luxburg, U.: Graph laplacians and their convergence on random neighborhood graphs. CoRR (2006)

Kallenberg, O., Sztencel, R.: Some dimension-free features of vector-valued martingales. Probability Theory and Related Fields 88, 215–247 (1991)

Langford, J.: Tutorial on practical prediction theory for classification. Journal of Machine Learning Research 6, 273–306 (2005)

Langford, J., Shawe-taylor, J.: Pac-bayes and margins. In: Advances in Neural Information Processing Systems, vol. 15, pp. 439–446. MIT Press, Cambridge (2002)

Lever, G., Laviolette, F., Shawe-Taylor, J.: Distribution dependent pac-bayes priors. UCL technical report (2010), http://www.cs.ucl.ac.uk/staff/G.Lever/pubs/DDPB.pdf

McAllester, D.A.: Pac-bayesian model averaging. In: COLT, pp. 164–170 (1999)

Ralaivola, L., Szafranski, M., Stempfel, G.: Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary $\beta$-mixing processes. CoRR, abs/0909.1993 (2009)

Seeger, M.: Pac-bayesian generalisation error bounds for gaussian process classification. Journal of Machine Learning Research 3, 233–269 (2002)

Serfling, R.: Approximation theorems of mathematical statistics. Wiley, Chichester (1980)

Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory 44, 1926–1940 (1998)

Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML 2003, pp. 912–919 (2003)

# A Technical Lemmas

**Lemma 6.** *(Hoeffding's lemma) Let $X$ be a random variable with $\mathbb{E}[X] = 0$ and $a < X < b$ then for $t > 0$,*

$$\mathbb{E}[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}.$$

The following theorem demonstrates that many key properties of martingales are independent of their dimension. The authors note that it is true for any Hilbert space-valued martingale but the proof is just for martingales in $\ell^2$.

**Theorem 7.** *(Kallenberg & Sztencel, 1991, Theorem 3.1) Let $\{V_t\}$ be a martingale in $\mathbb{R}^d$ or $\ell^2$. Then there exists a martingale $\{U_t\}$ in $\mathbb{R}^2$ such that $||V_t|| = ||U_t||$ a.s. and $||V_t - V_{t-1}|| = ||U_t - U_{t-1}||$ a.s..*

Given the above result all that we must do to obtain a large deviation inequality for $\ell^2$-valued martingales is to demonstrate a variation of Azuma-Hoeffding inequality for a martingale in $\mathbb{R}^2$, which is elementary if we are not concerned with obtaining the best constants.

**Corollary 1.** *For a martingale $\{V_i\}_{i=1}^m$ in $\mathbb{R}^d$ or $\ell^2$, such that, for all $i$,*

$$||V_i - V_{i-1}|| \leq c_i,$$

*we have for all $\delta > 0$,*

$$\mathbb{P}\left( ||V_m - V_0|| \leq 2\sqrt{\sum_{i=1}^m c_i^2 \ln \frac{4}{\delta}} \right) \geq 1 - \delta.$$

*Proof.* Consider a martingale $\{U_i\}_{i=1}^m$ in $\mathbb{R}^2$ such that,

$$||U_i - U_{i-1}|| \leq c_i. \tag{18}$$

Let $U_i = (U_i^{(1)}, U_i^{(2)})$, so that we have that $\{U_i^{(1)}\}_{i=1}^n$ and $\{U_i^{(2)}\}_{i=1}^n$ are clearly martingales and that,

$$|U_i^{(1)} - U_{i-1}^{(1)}| \leq c_i$$
$$|U_i^{(2)} - U_{i-1}^{(2)}| \leq c_i.$$

Now,

$$\begin{aligned}
\mathbb{P}\left(||U_m - U_0|| \geq \epsilon\right) &= \mathbb{P}\left((U_m^{(1)} - U_0^{(1)})^2 + (U_m^{(2)} - U_0^{(2)})^2 \geq \epsilon^2\right) \\
&\leq \mathbb{P}\left(|U_m^{(1)} - U_0^{(1)}| \geq \frac{\epsilon}{\sqrt{2}}\right) + \mathbb{P}\left(|U_m^{(2)} - U_0^{(2)}| \geq \frac{\epsilon}{\sqrt{2}}\right) \\
&\leq 4\exp\left(-\frac{\epsilon^2}{4\sum_{i=1}^m c_i^2}\right),
\end{aligned}$$

where the last line follows by the Azuma-Hoeffding inequality (Azuma, 1967). The result then follows by Theorem 7. □