Controlling Confusion via Generalisation Bounds

Reuben Adams, John Shawe-Taylor, Benjamin Guedj

Preprint 1–31

Controlling Confusion via Generalisation Bounds

Reuben Adams REUBEN.ADAMS.20@UCL.AC.UK
John Shawe-Taylor JOHN.SHAWE-TAYLOR@UCL.AC.UK
Benjamin Guedj B.GUEDJ@UCL.AC.UK
Centre for Artificial Intelligence, Department of Computer Science, University College London, UK

Abstract

We establish new generalisation bounds for multiclass classification by abstracting to a more general setting of discretised error types. Extending the PAC-Bayes theory, we are hence able to provide fine-grained bounds on performance for multiclass classification, as well as applications to other learning problems including discretisation of regression losses. Tractable training objectives are derived from the bounds. The bounds are uniform over all weightings of the discretised error types and thus can be used to bound weightings not foreseen at training, including the full confusion matrix in the multiclass classification case.

Keywords: Statistical Learning Theory, PAC-Bayes theory, Classification, Generalisation Bounds.

1. Introduction

Generalisation bounds are a core component of the theoretical understanding of machine learning algorithms. For over two decades now, the PAC-Bayesian theory has been at the core of studies on generalisation abilities of machine learning algorithms. PAC-Bayes originates in the seminal work of McAllester (1998, 1999) and was further developed by Catoni (2003, 2004, 2007), among other authors—we refer to the recent surveys Guedj (2019) and Alquier (2021) for an introduction to the field. The outstanding empirical successes of deep neural networks in the past decade call for better theoretical understanding of deep learning, and PAC-Bayes emerged as one of the few frameworks allowing the derivation of meaningful (and non-vacuous) generalisation bounds for neural networks: the pioneering work of Dziugaite and Roy (2017) has been followed by a number of contributions, including Neyshabur et al. (2018), Zhou et al. (2019), Letarte et al. (2019), Perez-Ortiz et al. (2021a,b) and Biggs and Guedj (2021, 2022a,b), to name but a few.

Much of the PAC-Bayes literature focuses on the case of binary classification, or of multiclass classification where one only distinguishes whether each classification is correct or incorrect. This is in stark contrast to the complexity of contemporary real-world learning problems. This work aims to bridge this gap via generalisation bounds that provide information rich measures of performance at test time by controlling the probabilities of errors of any finite number of types, bounding combinations of these probabilities uniformly over all weightings.

Previous results. We believe our framework of discretised error types to be novel. In the particular case of multiclass classification, little is known from a theoretical perspective and, to the best of our knowledge, only a handful of relevant strategies or generalisation bounds can be compared to the present paper. The closest is the work of Morvant et al. (2012) on a PAC-Bayes generalisation bound on the operator norm of the confusion matrix, to train a Gibbs classifier. We focus on a different performance metric, in the broader setting of discretised error types. Koço and Capponi (2013) suggest to minimise the confusion matrix norm with a focus on the imbalance between classes;

their treatment is not done through PAC-Bayes. Laviolette et al. (2017) extend the celebrated $\mathcal{C}$-bound in PAC-Bayes to weighted majority votes of classifiers, to perform multiclass classification. Benabbou and Lang (2017) present a streamlined version of some of the results from Morvant et al. (2012) in the case where some examples are voluntarily not classified (*e.g.*, in the case of too large uncertainty). More recently, Feofanov et al. (2019) derive bounds for a majority vote classifier where the confusion matrix serves as an error indicator: they conduct a study of the Bayes classifier.

From binary to multiclass classification. A number of PAC-Bayesian bounds have been unified by a single general bound, found in Bégin et al. (2016). Stated as Theorem 1 below, it applies to binary classification. We use it as a basis to prove our Theorem 3, a more general bound that can be applied to, amongst other things, multiclass classification and discretised regression. While the proof of Theorem 3 follows similar lines to that given in Bégin et al. (2016), our generalisation to 'soft' hypotheses incurring any finite number of error types requires a non-trivial extension of a result found in Maurer (2004). This extension (Lemma 5), along with its corollary (Corollary 6) may be of independent interest. The generalisation bound in Maurer (2004), stated below as Corollary 2, is shown in Bégin et al. (2016) to be a corollary of their bound. In a similar manner, we derive Corollary 7 from Theorem 3. Obtaining this corollary is significantly more involved than the analogous derivation in Bégin et al. (2016) or the original proof in Maurer (2004), requiring a number of technical results found in Appendix B.

Briefly, the results in Bégin et al. (2016) and Maurer (2004) consider an arbitrary input set $\mathcal{X}$, output set $\mathcal{Y} = \{-1, 1\}$, hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and i.i.d. sample $S \in (\mathcal{X} \times \mathcal{Y})^m$. They then establish high probability bounds on the discrepancy between the risk (probability of error an a new datapoint) of any stochastic classifier Q (namely, a distribution on $\mathcal{H}$) and its empirical counterpart (the fraction of the sample Q misclassifies). The bounds hold uniformly over all Q and contain a complexity term involving the Kullback-Leibler (KL) divergence between Q and a reference distribution P on $\mathcal{H}$ (often referred to as a prior by analogy with Bayesian inference—see the discussion in Guedj, 2019).

There are two ways in which the results in Bégin et al. (2016) and Maurer (2004) can be described as binary. First, as $\mathcal{Y}$ contains two elements, this is obviously an instance of binary classification. But a more interesting and subtle way to look at this is that only two cases are distinguished—correct classification and incorrect classification. Specifically, since the two different directions in which misclassification can be made are counted together, the bound gives no information on which direction is more likely.

More generally, the aforementioned bounds can be applied in the context of multiclass classification provided one maintains the second binary characteristic by only distinguishing correct and incorrect classifications rather than considering the entire confusion matrix. However, note that these bounds will not give information on the relative likelihood of the different errors. In contrast, our new results can consider the entire confusion matrix, bounding how far the true (read "expected over the data-generating distribution") confusion matrix differs from the empirical one, according to some metric. In fact, our results extend to the case of arbitrary label set $\mathcal{Y}$, provided the number of different errors one distinguishes is finite.

Formally, we let $\bigcup_{j=1}^{M} E_j$ be a user-specified disjoint partition of $\mathcal{Y}^2$ into a finite number of M *error types*[1], where we say that a hypothesis $h \in \mathcal{H}$ makes an error of type j on datapoint (x, y) if $(h(x), y) \in E_j$. It should be stressed that some E_j need not correspond to mislabellings—indeed,

1. By convention, every pair $(\hat{y}, y) \in \mathcal{Y}^2$ is interpreted as a predicted value $\hat{y}$ followed by a true value y, in that order.

some of the E_j may distinguish different correct labellings. We then count up the number of errors of each type that a hypothesis makes on a sample, and bound how far this empirical distribution of errors is from the expected distribution under the data-generating distribution (Theorem 3). Thus, in our generalisation, the (scalar) risk and empirical risk ($R_D(Q)$ and $R_S(Q)$, defined in the next section) are replaced by M-dimensional vectors ($\boldsymbol{R}_D(Q)$ and $\boldsymbol{R}_S(Q)$), and our discrepancy measure d is a divergence between discrete distributions on M elements. Our generalisation therefore allows us to bound how far the true distribution of errors can be from the observed distribution of errors. If we then associate a loss value $\ell_j \in [0, \infty)$ to each E_j we can derive a bound on the *total risk*, defined as the sum of the true error probabilities weighted by the loss values. In fact, the total risk is bounded with high probability uniformly over all such weightings. The loss values need not be distinct; we may wish to understand the distribution of error types even across error types that incur the same loss.

For example, in the case of binary classification with $\mathcal{Y} = \{-1, 1\}$, we can take the usual partition into $E_1 = \{(-1,-1),(1,1)\}$ and $E_2 = \{(-1,1),(1,-1)\}$ and loss values $\ell_1 = 0, \ell_2 = 1$, or the fine-grained partition $\mathcal{Y}^2 = \{(0,0)\} \cup \{(1,1)\} \cup \{(0,1)\} \cup \{(1,0)\}$ and the loss values $\ell_1 = \ell_2 = 0, \ell_3 = 1, \ell_4 = 2$. More generally, for multiclass classification with N classes and $\mathcal{Y} = [N]$, one may take the usual coarse partition into $E_1 = \{(\hat{y}, y) \in \mathcal{Y}^2 : \hat{y} = y\}$ and $E_2 = \{(\hat{y}, y) \in \mathcal{Y}^2 : \hat{y} \neq y\}$ (with $\ell_1 = 0$ and $\ell_2 = 1$), or the fully refined partition into $E_{i,j} = \{(i,j)\}$ for $i, j \in [N]$ (with correspondingly greater choice of the associated loss values), or something in-between. Note that we still refer to E_j as an "error type" even if it contains elements that correspond to correct classification, namely if there exists $y \in \mathcal{Y}$ such that $(y, y) \in E_j$. As we will see later, a more fine-grained partition will allow more error types to be distinguished and bounded, at the expense of a looser bound. As a final example, for regression with $\mathcal{Y} = \mathbb{R}$, we may fix M strictly increasing thresholds $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_M$ and partition $\mathcal{Y}^2$ into $E_j = \{(y_1, y_2) \in \mathcal{Y}^2 : \lambda_j \leq |y_1 - y_2| < \lambda_{j+1}\}$ for $j \in [M-1]$, and $E_M = \{(y_1, y_2) \in \mathcal{Y}^2 : |y_1 - y_2| \geq \lambda_M\}$.

Outline. We set our notation in Section 2. In Section 3 we state and prove generalisation bounds in the setting of discretised error types: this significantly expands the previously known results from Bégin et al. (2016) by allowing for generic output sets $\mathcal{Y}$. Our main results are Theorem 3 and Corollary 7. To make our findings profitable to the broader machine learning community we then discuss how these new bounds can be turned into tractable training objectives in Section 4 (with a general recipe described in greater detail in Appendix A). The paper closes with perspectives for follow-up work in Section 5 and we defer to Appendix B the proofs of technical results.

2. Notation

For any set A, let $\mathcal{M}(A)$ be the set of probability measures on A. For any $M \in \mathbb{Z}_{>0}$, define $[M] := \{1, 2, \ldots, M\}$, the M-dimensional simplex $\triangle_M := \{\boldsymbol{u} \in [0,1]^M : u_1 + \cdots + u_M = 1\}$ and its interior $\triangle_M^{>0} := \triangle_M \cap (0,1)^M$. For $m, M \in \mathbb{Z}_{>0}$, define the integer counterparts $S_{m,M} := \left\{(k_1, \ldots, k_M) \in \mathbb{Z}_{\geq 0}^M : k_1 + \cdots + k_M = m\right\}$ and $S_{m,M}^{>0} := S_{m,M} \cap \mathbb{Z}_{>0}^M$. The set $S_{m,M}$ is the domain of the multinomial distribution with parameters m, M and some $\boldsymbol{r} \in \triangle_M$, which is denoted $\mathrm{Mult}(m, M, \boldsymbol{r})$ and has probability mass function for $\boldsymbol{k} \in S_{m,M}$ given by

$$\mathrm{Mult}(\boldsymbol{k}; m, M, \boldsymbol{r}) := \binom{m}{k_1 \;\; k_2 \;\; \cdots \;\; k_M} \prod_{j=1}^{M} r_j^{k_j}, \quad \text{where} \quad \binom{m}{k_1 \;\; k_2 \;\; \cdots \;\; k_M} := \frac{m!}{\prod_{j=1}^{M} k_j!}.$$

For $\boldsymbol{q}, \boldsymbol{p} \in \triangle_M$, let $\mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p})$ denote the KL-divergence of $\mathrm{Mult}(1, M, \boldsymbol{q})$ from $\mathrm{Mult}(1, M, \boldsymbol{p})$, namely $\mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) := \sum_{j=1}^{M} q_j \ln \frac{q_j}{p_j}$, with the convention that $0 \log \frac{0}{x} = 0$ for $x \geq 0$ and $x \log \frac{x}{0} = \infty$ for $x > 0$. For $M = 2$ we abuse notation and abbreviate $\mathrm{kl}((q, 1-q)\|(p, 1-p))$ to $\mathrm{kl}(q\|p)$, which is then the conventional definition of $\mathrm{kl}(\cdot\|\cdot) : [0,1]^2 \to [0, \infty]$ found in the PAC-Bayes literature (as in Seeger, 2002, for example).

Let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary input (*e.g.*, feature) and output (*e.g.*, label) sets respectively. Let $\bigcup_{j=1}^{M} E_j$ be a partition of $\mathcal{Y}^2$ into a finite sequence of M *error types*, and to each E_j associate a loss value $\ell_j \in [0, \infty)$. The only restriction we place on the loss values ℓ_j is that they are not all equal. This is not a strong assumption, since if they were all equal then all hypotheses would incur equal loss and there would be no learning problem: we are effectively ruling out trivial cases.

Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote a hypothesis class, $D \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ a data-generating distribution and $S \sim D^m$ an i.i.d. sample of size m drawn from D. For $h \in \mathcal{H}$ and $j \in [M]$ we define the *empirical j-risk* and *true j-risk* of h to be $R_S^j(h) := \frac{1}{m} \sum_{(x,y) \in S} \mathbb{1}[(h(x), y) \in E_j]$ and $R_D^j(h) := \mathbb{E}_{(x,y)\sim D}[\mathbb{1}[(h(x), y) \in E_j]]$, respectively, namely, the proportion of the sample S on which h makes an error of type E_j and the probability that h makes an error of type E_j on a new $(x, y) \sim D$.

More generally, suppose $\mathcal{H} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ is a class of *soft* hypotheses of the form $H : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$, where, for any $A \subseteq \mathcal{Y}$, $H(x)[A]$ is interpreted as the probability according to H that the label of x is in A. It is worth stressing that a soft hypothesis is still deterministic since a prediction is not drawn from the distribution it returns. We then define the *empirical j-risk* of H to be $R_S^j(H) := \frac{1}{m} \sum_{(x,y)\in S} H(x)\big[\{\hat{y} \in \mathcal{Y} : (\hat{y}, y) \in E_j\}\big]$, namely the mean—over the elements (x, y) of S—probability mass H assigns to predictions $\hat{y} \in \mathcal{Y}$ incurring an error of type E_j when labelling each x. Further, we define the *true j-risk* of H to be $R_D^j(H) := \mathbb{E}_{(x,y)\sim D}\big[H(x)\big[\{\hat{y} \in \mathcal{Y} : (\hat{y}, y) \in E_j\}\big]\big]$, namely the mean—over $(x, y) \sim D$—probability mass H assigns to predictions $\hat{y} \in \mathcal{Y}$ incurring an error of type E_j when labelling each x. We will see in Section 4 that the more general hypothesis class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ is necessary for constructing a differentiable training objective.

To each ordinary hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ there corresponds a soft hypothesis $H \in \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ that, for each $x \in \mathcal{X}$, returns a point mass on $h(x)$. In this case, it is straightforward to show that $R_S^j(h) = R_S^j(H)$ and $R_D^j(h) = R_D^j(H)$ for all $j \in [M]$, where we have used the corresponding definitions above for ordinary and soft hypotheses. Since, in addition, our results hold identically for both ordinary and soft hypotheses, we henceforth use the same notation h for both ordinary and soft hypotheses and their associated values $R_S^j(h)$ and $R_D^j(h)$. It will always be clear from the context whether we are dealing with ordinary or soft hypotheses and thus which of the above definitions of the empirical and true j-risks is being used.

We define the *empirical risk* and *true risk* of a (ordinary or soft) hypothesis h to be $\boldsymbol{R}_S(h) := (R_S^1(h), \ldots, R_S^M(h))$ and $\boldsymbol{R}_D(h) := (R_D^1(h), \ldots, R_D^M(h))$, respectively. It is straightforward to show that $\boldsymbol{R}_S(h)$ and $\boldsymbol{R}_D(h)$ are elements of $\triangle_M$. Since S is drawn i.i.d. from D, the expectation of the empirical risk is equal to the true risk, namely $\mathbb{E}_S[R_S^j(h)] = R_D^j(h)$ for all j and thus $\mathbb{E}_S[\boldsymbol{R}_S(h)] = \boldsymbol{R}_D(h)$. Finally, we generalise to stochastic hypotheses $Q \in \mathcal{M}(\mathcal{H})$, which predict by first drawing a deterministic hypothesis $h \sim Q$ and then predicting according to h, where a new h is drawn for each prediction. Thus, we define the *empirical j-risk* and *true j-risk* of Q to be the scalars $R_S^j(Q) := \mathbb{E}_{h\sim Q}[R_S^j(h)]$ and $R_D^j(Q) := \mathbb{E}_{h\sim Q}[R_D^j(h)]$, for $j \in [M]$, and simply the *empirical risk* and *true risk* of Q to be the elements of $\triangle_M$ defined by $\boldsymbol{R}_S(Q) := \mathbb{E}_{h\sim Q}[\boldsymbol{R}_S(h)]$ and $\boldsymbol{R}_D(Q) := \mathbb{E}_{h\sim Q}[\boldsymbol{R}_D(h)]$. As before, since S is i.i.d., we have (using Fubini this time) that $\mathbb{E}_S[\boldsymbol{R}_S(Q)] = \boldsymbol{R}_D(Q)$. Finally, given a loss vector $\boldsymbol{\ell} \in [0, \infty)^M$, we define the *total risk* of Q by the scalar $R_D^T(Q) := \sum_{j=1}^{M} \ell_j R_D^j(Q)$. As is conventional in the PAC-Bayes literature, we refer

to sample independent and dependent distributions on $\mathcal{M}(\mathcal{H})$ (*i.e.* stochastic hypotheses) as *priors* (denoted P) and *posteriors* (denoted Q) respectively, even if they are not related by Bayes' theorem.

3. Inspiration and Main Results

We first state the existing results in Bégin et al. (2016) and Maurer (2004) that we will generalise from just two error types (correct and incorrect) to any finite number of error types. These results are stated in terms of the scalars $R_S(Q) := \frac{1}{m}\sum_{(x,y)\in S} \mathbb{1}[h(x) \neq y]$ and $R_D(Q) := \mathbb{E}_{(x,y)\sim D}\mathbb{1}[h(x) \neq y]$ and, as we demonstrate, correspond to the case $M = 2$ of our generalisations.

Theorem 1 *(Bégin et al., 2016, Theorem 4) Let $\mathcal{X}$ be an arbitrary set and $\mathcal{Y} = \{-1, 1\}$. Let $D \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ be a data-generating distribution and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. For any prior $P \in \mathcal{M}(\mathcal{H})$, $\delta \in (0, 1]$, convex function $d : [0, 1]^2 \to \mathbb{R}$, sample size m and $\beta \in (0, \infty)$, with probability at least $1 - \delta$ over the random draw $S \sim D^m$, we have that simultaneously for all posteriors $Q \in \mathcal{M}(\mathcal{H})$*

$$d\big(R_S(Q), R_D(Q)\big) \leq \frac{1}{\beta}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_d(m,\beta)}{\delta}\right],$$

with $\mathcal{I}_d(m,\beta) := \sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\mathrm{Bin}(k;m,r)\exp\left(\beta d\left(\frac{k}{m}, r\right)\right)\right]$, where $\mathrm{Bin}(k;m,r)$ is the binomial probability mass function $\mathrm{Bin}(k;m,r) := \binom{m}{k}r^k(1-r)^{m-k}$.

Note the original statement in Bégin et al. (2016) is for a positive integer m', but the proof trivially generalises to any $\beta \in (0, \infty)$. One of the bounds that Theorem 1 unifies—which we also generalise—is that of Seeger (2002), later tightened in Maurer (2004), which we now state. It can be recovered from Theorem 1 by setting $\beta = m$ and $d(q, p) = \mathrm{kl}(q\|p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$.

Corollary 2 *(Maurer, 2004, Theorem 5) Let $\mathcal{X}$ be an arbitrary set and $\mathcal{Y} = \{-1, 1\}$. Let $D \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ be a data-generating distribution and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. For any prior $P \in \mathcal{M}(\mathcal{H})$, $\delta \in (0, 1]$ and sample size m, with probability at least $1 - \delta$ over the random draw $S \sim D^m$, we have that simultaneously for all posteriors $Q \in \mathcal{M}(\mathcal{H})$*

$$\mathrm{kl}\big(R_S(Q), R_D(Q)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right].$$

We wish to bound the deviation of the empirical vector $\boldsymbol{R}_S(Q)$ from the unknown vector $\boldsymbol{R}_D(Q)$. Since in general the stochastic hypothesis Q we learn will depend on the sample S, it is useful to obtain bounds on the deviation of $\boldsymbol{R}_S(Q)$ from $\boldsymbol{R}_D(Q)$ that are uniform over Q, just as in Theorem 1 and Corollary 2. In Theorem 1, the deviation $d(R_S(Q), R_D(Q))$ between the scalars $R_S(Q), R_D(Q) \in [0, 1]$ is measured by some convex function $d : [0, 1]^2 \to \mathbb{R}$. In our case, the deviation $d(\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q))$ between the vectors $\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q) \in \triangle_M$ is measured by some convex function $d : \triangle_M^2 \to \mathbb{R}$. In Section 3.2 we will derive Corollary 7 from Theorem 3 by selecting $\beta = m$ and $d(\boldsymbol{q}, \boldsymbol{p}) := \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p})$, analogous to how Corollary 2 is obtained from Theorem 1.

3.1. Statement and proof of the generalised bound

We now state and prove our generalisation of Theorem 1. The proof follows identical lines to that of Theorem 1 given in Bégin et al. (2016), but with additional non-trivial steps to account for the greater number of error types and the possibility of soft hypotheses.

Theorem 3 *Let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary sets and $\bigcup_{j=1}^{M} E_j$ be a disjoint partition of $\mathcal{Y}^2$. Let $D \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ be a data-generating distribution and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class. For any prior $P \in \mathcal{M}(\mathcal{H})$, $\delta \in (0, 1]$, jointly convex function $d : \triangle_M^2 \to \mathbb{R}$, sample size m and $\beta \in (0, \infty)$, with probability at least $1 - \delta$ over the random draw $S \sim D^m$, we have that simultaneously for all posteriors $Q \in \mathcal{M}(\mathcal{H})$*

$$d\big(\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q)\big) \leq \frac{1}{\beta}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_d(m,\beta)}{\delta}\right], \tag{1}$$

where $\mathcal{I}_d(m,\beta) := \sup_{\boldsymbol{r}\in\triangle_M}\left[\sum_{\boldsymbol{k}\in S_{m,M}} \mathrm{Mult}(\boldsymbol{k}; m, M, \boldsymbol{r}) \exp\left(\beta d\left(\frac{\boldsymbol{k}}{m}, \boldsymbol{r}\right)\right)\right]$. Further, the bounds are unchanged if one restricts to an ordinary hypothesis class, namely if $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$.

One can derive multiple bounds from this theorem, all of which then hold simultaneously with probability at least $1 - \delta$. For example, one can derive bounds on the individual error probabilities $R_D^j(Q)$ or combinations thereof. It is this flexibility that allows Theorem 3 to provide far richer information on the performance of the posterior Q on unseen data. For a more in depth discussion of how such bounds can be derived, including a recipe for transforming the bound into a differentiable training objective, see Section 4 and Appendix A.

To see that Theorem 3 is a generalisation of Theorem 1, note that we can recover it by setting $\mathcal{Y} = \{-1, 1\}$, $M = 2$, $E_1 = \{(-y, y) : y \in \mathcal{Y}\}$ and $E_2 = \{(y, y) : y \in \mathcal{Y}\}$. Then, for any convex function $d : [0, 1]^2 \to \mathbb{R}$, apply Theorem 3 with the convex function $d' : \triangle_M^2 \to \mathbb{R}$ defined by $d'((u_1, u_2), (v_1, v_2)) := d(u_1, v_1)$ so that Theorem 3 bounds $d'\big(\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q)\big) = d\big(R_S^1(Q), R_D^1(Q)\big)$ which equals $d(R_S(Q), R_D(Q))$ in the notation of Theorem 1. Further,

$$\sum_{\boldsymbol{k}\in S_{m,2}} \mathrm{Mult}(\boldsymbol{k}; m, 2, \boldsymbol{r}) \exp\left(\beta d'\left(\frac{\boldsymbol{k}}{m}, \boldsymbol{r}\right)\right) = \sum_{k=0}^{m} \mathrm{Bin}(k; m, r_1) \exp\left(\beta d\left(\frac{k}{m}, r_1\right)\right),$$

so that the supremum over $r_1 \in [0, 1]$ of the right hand side equals the supremum over $\boldsymbol{r} \in \triangle_2$ of the left hand side, which, when substituted into (1), yields the bound given in Theorem 1.

Our proof of Theorem 3 follows the lines of the proof of Theorem 1 in Bégin et al. (2016), making use of the change of measure inequality Lemma 4. However, a complication arises from the use of soft classifiers $h \in \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$. A similar problem is dealt with in Maurer (2004) when proving Corollary 2 by means of a Lemma permitting the replacement of $[0, 1]$-valued random variables by corresponding $\{0, 1\}$-valued random variables with the same mean. We use a generalisation of this, stated as Lemma 5 (Lemma 3 in Maurer, 2004 corresponds to the case $M = 2$), the proof of which is not insightful for our purposes and thus deferred to Appendix B.1. An immediate consequence of Lemma 5 is Corollary 6, which is a generalisation of the first half of Theorem 1 in Maurer (2004). While we only use it implicitly in the remainder of the paper, we state it as it may be of broader interest.

The consequence of Lemma 5 is that the worst case (in terms of bounding $d(\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q))$) occurs when the $\boldsymbol{R}_{\{(x,y)\}}(h)$ are one-hot vectors for all $(x, y) \in S$ and $h \in \mathcal{H}$, namely when $\mathcal{H} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ only contains hypotheses that, when labelling S, put all their mass on elements $\hat{y} \in \mathcal{Y}$ that incur the same error type[2]. In particular, this is the case for hypotheses that put all their mass on a single element of $\mathcal{Y}$, equivalent to the simpler case $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ as discussed in Section 2. Thus, Lemma 5 shows that the bound given in Theorem 3 cannot be made tighter only by restricting to such hypotheses.

Lemma 4 *(Change of measure, Csiszár, 1975, Donsker and Varadhan, 1975) For any set $\mathcal{H}$, any $P, Q \in \mathcal{M}(\mathcal{H})$ and any measurable function $\phi : \mathcal{H} \to \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{h\sim Q}\phi(h) \leq \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbb{E}}_{h\sim P}\exp(\phi(h))\right).$$

Lemma 5 *(Generalisation of Lemma 3 in Maurer, 2004) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ be i.i.d $\triangle_M$-valued random vectors with mean $\boldsymbol{\mu}$ and suppose that $f : \triangle_M^m \to \mathbb{R}$ is convex. If $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_m$ are i.i.d.* $\mathrm{Mult}(1, M, \boldsymbol{\mu})$ *random vectors, then*

$$\mathbb{E}[f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)] \leq \mathbb{E}[f(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_m)].$$

Corollary 6 *(Generalisation of Theorem 1 in Maurer, 2004) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ be i.i.d $\triangle_M$-valued random vectors with mean $\boldsymbol{\mu}$ and $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_m$ be i.i.d.* $\mathrm{Mult}(1, M, \boldsymbol{\mu})$*. Then*

$$\mathbb{E}\Big[\exp\left(m\mathrm{kl}(\bar{\boldsymbol{X}}\|\mu)\right)\Big] \leq \mathbb{E}\Big[\exp\left(m\mathrm{kl}(\bar{\boldsymbol{X}}'\|\mu)\right)\Big],$$

where $\bar{\boldsymbol{X}} := \frac{1}{m}\sum_{i=1}^m \boldsymbol{X}_i$ and $\bar{\boldsymbol{X}}' := \frac{1}{m}\sum_{i=1}^m \boldsymbol{X}'_i$.

Proof (of Corollary 6) This is immediate from Lemma 5 since the average is linear, the kl-divergence is convex and the exponential is non-decreasing and convex. ■

Proof (of Theorem 3) The case $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ follows directly from the more general case by taking $\mathcal{H}' := \{h' \in \mathcal{M}(\mathcal{Y})^{\mathcal{X}} : \exists h \in \mathcal{H} \text{ such that } \forall x \in \mathcal{X} \;\; h'(x) = \delta_{h(x)}\}$, where $\delta_{h(x)} \in \mathcal{M}(\mathcal{Y})$ denotes a point mass on $h(x)$. For the general case $\mathcal{H} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$, using Jensen's inequality with the convex function $d(\cdot,\cdot)$ and Lemma 4 with $\phi(h) = \beta d(\boldsymbol{R}_S(h), \boldsymbol{R}_D(h))$, we see that for all $Q \in \mathcal{M}(\mathcal{H})$

$$\begin{aligned}
\beta d\big(\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q)\big) &= \beta d\left(\mathop{\mathbb{E}}_{h\sim Q}\boldsymbol{R}_S(h), \mathop{\mathbb{E}}_{h\sim Q}\boldsymbol{R}_D(h)\right) \\
&\leq \mathop{\mathbb{E}}_{h\sim Q}\beta d\big(\boldsymbol{R}_S(h), \boldsymbol{R}_D(h)\big) \\
&\leq \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbb{E}}_{h\sim P}\exp\left(\beta d\big(\boldsymbol{R}_S(h), \boldsymbol{R}_D(h)\big)\right)\right) \\
&= \mathrm{KL}(Q\|P) + \ln(Z_P(S)),
\end{aligned}$$

where $Z_P(S) := \mathbb{E}_{h\sim P}\exp\left(\beta d(\boldsymbol{R}_S(h), \boldsymbol{R}_D(h))\right)$. Note that $Z_P(S)$ is a non-negative random variable, so that by Markov's inequality $\mathop{\mathrm{P}}_{S\sim D^m}\left(Z_P(S) \leq \frac{\mathbb{E}_{S'\sim D^m} Z_P(S')}{\delta}\right) \geq 1 - \delta$. Thus, since

2. More precisely, when $\forall h \in \mathcal{H} \;\; \forall (x, y) \in S \;\; \exists j \in [M]$ such that $h(x)[\{\hat{y} \in \mathcal{Y} : (\hat{y}, y) \in E_j\}] = 1$.

$\ln(\cdot)$ is strictly increasing, with probability at least $1-\delta$ over $S \sim D^m$, we have that simultaneously for all $Q \in \mathcal{M}(\mathcal{H})$

$$\beta d\big(\boldsymbol{R}_S(Q), \boldsymbol{R}_D(Q)\big) \leq \mathrm{KL}(Q\|P) + \ln \frac{\mathop{\mathbb{E}}_{S' \sim D^m} Z_P(S')}{\delta}. \tag{2}$$

To bound $\mathbb{E}_{S' \sim D^m} Z_P(S')$, let $\boldsymbol{X}_i := \boldsymbol{R}_{\{(x_i, y_i)'\}}(h) \in \triangle_M$ for $i \in [m]$, where $(x_i, y_i)'$ is the i'th element of the dummy sample S'. Noting that each $\boldsymbol{X}_i$ has mean $\boldsymbol{R}_D(h)$, define the random vectors $\boldsymbol{X}'_i \sim \mathrm{Mult}(1, M, \boldsymbol{R}_D(h))$ and $\boldsymbol{Y} := \sum_{i=1}^m \boldsymbol{X}'_i \sim \mathrm{Mult}(m, M, \boldsymbol{R}_D(h))$. Finally let $f : \triangle_M^m \to \mathbb{R}$ be defined by $f(x_1, \ldots, x_m) := \exp\left(\beta d\left(\frac{1}{m}\sum_{i=1}^m x_i, \boldsymbol{R}_D(h)\right)\right)$, which is convex since the average is linear, d is convex and the exponential is non-decreasing and convex. Then, by swapping expectations (which is permitted by Fubini's theorem since the argument is non-negative) and applying Lemma 5, we have that $\mathbb{E}_{S' \sim D^m} Z_P(S')$ can be written as

$$\begin{aligned}
\mathbb{E}_{S' \sim D^m} Z_P(S') &= \mathop{\mathbb{E}}_{S' \sim D^m} \mathop{\mathbb{E}}_{h \sim P} \exp\Big(\beta d\big(\boldsymbol{R}_{S'}(h), \boldsymbol{R}_D(h)\big)\Big) \\
&= \mathop{\mathbb{E}}_{h \sim P} \mathop{\mathbb{E}}_{S' \sim D^m} \exp\Big(\beta d\big(\boldsymbol{R}_{S'}(h), \boldsymbol{R}_D(h)\big)\Big) \\
&= \mathop{\mathbb{E}}_{h \sim P} \mathop{\mathbb{E}}_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m} \exp\left(\beta d\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{X}_i, \boldsymbol{R}_D(h)\right)\right) \\
&\leq \mathop{\mathbb{E}}_{h \sim P} \mathop{\mathbb{E}}_{\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_m} \exp\left(\beta d\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{X}'_i, \boldsymbol{R}_D(h)\right)\right) \\
&= \mathop{\mathbb{E}}_{h \sim P} \mathop{\mathbb{E}}_{\boldsymbol{Y}} \exp\left(\beta d\left(\frac{1}{m}\boldsymbol{Y}, \boldsymbol{R}_D(h)\right)\right) \\
&= \mathop{\mathbb{E}}_{h \sim P} \sum_{\boldsymbol{k} \in S_{m,M}} \mathrm{Mult}\big(\boldsymbol{k}; m, M, \boldsymbol{R}_D(h)\big) \exp\Big(\beta d\big(\tfrac{\boldsymbol{k}}{m}, \boldsymbol{R}_D(h)\big)\Big) \\
&\leq \sup_{\boldsymbol{r} \in \triangle_M} \left[\sum_{\boldsymbol{k} \in S_{m,M}} \mathrm{Mult}\big(\boldsymbol{k}; m, M, \boldsymbol{r}\big) \exp\Big(\beta d\big(\tfrac{\boldsymbol{k}}{m}, \boldsymbol{r}\big)\Big)\right].
\end{aligned}$$

Which is the definition of $\mathcal{I}_d(m, \beta)$. Inequality (1) then follows by substituting this bound on $\mathbb{E}_{S' \sim D^m} Z_P(S')$ into (2) and dividing by β. ■

3.2. Statement and proof of the generalised corollary

We now apply our generalised theorem with $\beta = m$ and $d(\boldsymbol{q}, \boldsymbol{p}) = \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p})$. This results in the following corollary, analogous to Corollary 2 (although the multi-dimensionality makes the proof much more involved, requiring multiple lemmas and extra arguments to make the main idea go through). We give two forms of the bound since, while the second is looser, the first is not practical to calculate except when m is very small.

Corollary 7 *Let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary sets and $\bigcup_{j=1}^M E_j$ be a disjoint partition of $\mathcal{Y}^2$. Let $D \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ be a data-generating distribution and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class. For any*

prior $P \in \mathcal{M}(\mathcal{H})$, $\delta \in (0,1]$ and sample size m, with probability at least $1-\delta$ over the random draw $S \sim D^m$, we have that simultaneously for all posteriors $Q \in \mathcal{M}(\mathcal{H})$

$$\mathrm{kl}\big(\boldsymbol{R}_S(Q)\|\boldsymbol{R}_D(Q)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{m!}{\delta m^m}\sum_{\boldsymbol{k}\in S_{m,M}}\prod_{j=1}^{M}\frac{k_j^{k_j}}{k_j!}\right)\right] \tag{3}$$

$$\leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{1}{\delta}\sqrt{\pi}e^{1/12m}\left(\frac{m}{2}\right)^{\frac{M-1}{2}}\sum_{z=0}^{M-1}\binom{M}{z}\frac{1}{(\pi m)^{z/2}\,\Gamma\left(\frac{M-z}{2}\right)}\right)\right], \tag{4}$$

where the second inequality holds provided $m \geq M$. Further, the bounds are unchanged if one restricts to an ordinary hypothesis class, namely if $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$.

While analogous corollaries can be obtained from Theorem 3 by other choices of convex function d, the kl-divergence leads to convenient cancellations that remove the dependence of $\mathcal{I}_{\mathrm{kl}}(m,\beta,\boldsymbol{r})$ on $\boldsymbol{r}$, making $\mathcal{I}_{\mathrm{kl}}(m,\beta) := \sup_{\boldsymbol{r}\in\triangle_M}\mathcal{I}_{\mathrm{kl}}(m,\beta,\boldsymbol{r})$ simple to evaluate. Nevertheless, if one desires a bound on, say, the total variation $\mathrm{TV}(\boldsymbol{R}_S(Q),\boldsymbol{R}_D(Q)) := \frac{1}{2}\|\boldsymbol{R}_S(Q)-\boldsymbol{R}_D(Q)\|_1$, one may apply to Corollary 7 Pinsker's inequality $\mathrm{TV}(\boldsymbol{R}_S(Q),\boldsymbol{R}_D(Q)) \leq \sqrt{\frac{1}{2}\mathrm{kl}(\boldsymbol{R}_S(Q)\|\boldsymbol{R}_D(Q))}$, or $\mathrm{TV}(\boldsymbol{R}_S(Q),\boldsymbol{R}_D(Q)) \leq \sqrt{1-e^{-\mathrm{kl}(\boldsymbol{R}_S(Q)\|\boldsymbol{R}_D(Q))}}$ (due to Bretagnolle and Huber, 1978, Lemma 2.1; see also Tsybakov, 2009, Eq. 2.25). Further, if one desires a bound on the Hellinger distance $\mathrm{H}(\boldsymbol{R}_S(Q),\boldsymbol{R}_D(Q)) := \frac{1}{\sqrt{2}}\|\sqrt{\boldsymbol{R}_S(Q)}-\sqrt{\boldsymbol{R}_D(Q)}\|_2$, one can combine one of the previous inequalities with $\mathrm{H}(\boldsymbol{R}_S(Q),\boldsymbol{R}_D(Q)) \leq \sqrt{\mathrm{TV}(\boldsymbol{R}_S(Q),\boldsymbol{R}_D(Q))}$ (due to Kraft, 1955; see also Steerneman, 1983 and van Erven and Harremoës, 2014).

To prove Corollary 7 we require Lemma 8, the proof of which is deferred to Appendix B.2.

Lemma 8 *For integers $M \geq 1$ and $m \geq M$,*

$$\sum_{\boldsymbol{k}\in S^{>0}_{m,M}}\frac{1}{\prod_{j=1}^{M}\sqrt{k_j}} \leq \frac{\pi^{\frac{M}{2}}m^{\frac{M-2}{2}}}{\Gamma(\frac{M}{2})}.$$

Proof (of Corollary 7) Applying Theorem 3 with $d(\boldsymbol{q},\boldsymbol{p}) = \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p})$ (defined in Section 2) and $\beta = m$ gives that with probability at least $1-\delta$ over $S \sim D^m$, simultaneously for all posteriors $Q \in \mathcal{M}(\mathcal{H})$,

$$\mathrm{kl}\big(\boldsymbol{R}_S(Q)\|\boldsymbol{R}_D(Q)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_{\mathrm{kl}}(m,m)}{\delta}\right],$$

where $\mathcal{I}_{\mathrm{kl}}(m,m) := \sup_{\boldsymbol{r}\in\triangle_M}\left[\sum_{\boldsymbol{k}\in S_{m,M}}\mathrm{Mult}(\boldsymbol{k};m,M,\boldsymbol{r})\exp\left(m\mathrm{kl}\left(\frac{\boldsymbol{k}}{m},\boldsymbol{r}\right)\right)\right]$. Thus, to establish the first inequality of the corollary, it suffices to show that

$$\mathcal{I}_{\mathrm{kl}}(m,m) \leq \frac{m!}{m^m}\sum_{\boldsymbol{k}\in S_{m,M}}\prod_{j=1}^{M}\frac{k_j^{k_j}}{k_j!}. \tag{5}$$

To see this, for each fixed $\boldsymbol{r} = (r_1,\ldots,r_M) \in \triangle_M$ let $J_{\boldsymbol{r}} = \{j \in [M] : r_j = 0\}$. Then $\mathrm{Mult}(\boldsymbol{k};m,M,\boldsymbol{r}) = 0$ for any $\boldsymbol{k} \in S_{m,M}$ such that $k_j \neq 0$ for some $j \in J_{\boldsymbol{r}}$. For the other

$\boldsymbol{k} \in S_{m,M}$, namely those such that $k_j = 0$ for all $j \in J_{\boldsymbol{r}}$, the probability term can be written as

$$\text{Mult}(\boldsymbol{k}; m, M, \boldsymbol{r}) = \frac{m!}{\prod_{j=1}^{M} k_j!} \prod_{j=1}^{M} r_j^{k_j} = \frac{m!}{\prod_{j \notin J_{\boldsymbol{r}}} k_j!} \prod_{j \notin J_{\boldsymbol{r}}} r_j^{k_j},$$

and (recalling the convention that $0 \log \frac{0}{0} = 0$) the term $\exp\left(m\text{kl}\left(\frac{\boldsymbol{k}}{m}, \boldsymbol{r}\right)\right)$ can be written as

$$\exp\left(m \sum_{j=1}^{M} \frac{k_j}{m} \ln \frac{\frac{k_j}{m}}{r_j}\right) = \exp\left(\sum_{j \notin J_{\boldsymbol{r}}} k_j \ln \frac{k_j}{m r_j}\right) = \prod_{j \notin J_{\boldsymbol{r}}} \left(\frac{k_j}{m r_j}\right)^{k_j} = \frac{1}{m^m} \prod_{j \notin J_{\boldsymbol{r}}} \left(\frac{k_j}{r_j}\right)^{k_j},$$

where the last equality is obtained by recalling that the k_j sum to m. Substituting these two expressions into the definition of $\mathcal{I}_{\text{kl}}(m, m)$ and only summing over those $\boldsymbol{k} \in S_{m,M}$ with non-zero probability, we obtain

$$\begin{aligned}
\sum_{\boldsymbol{k} \in S_{m,M}} \text{Mult}(\boldsymbol{k}; m, M, \boldsymbol{r}) \exp\left(m\text{kl}\left(\tfrac{\boldsymbol{k}}{m}, \boldsymbol{r}\right)\right) &= \sum_{\substack{\boldsymbol{k} \in S_{m,M}: \\ \forall j \in J_{\boldsymbol{r}}\ k_j = 0}} \text{Mult}(\boldsymbol{k}; m, M, \boldsymbol{r}) \exp\left(m\text{kl}\left(\tfrac{\boldsymbol{k}}{m}, \boldsymbol{r}\right)\right) \\
&= \sum_{\substack{\boldsymbol{k} \in S_{m,M}: \\ \forall j \in J_{\boldsymbol{r}}\ k_j = 0}} \frac{m!}{\prod_{j \notin J_{\boldsymbol{r}}} k_j!} \prod_{j \notin J_{\boldsymbol{r}}} r_j^{k_j} \frac{1}{m^m} \prod_{j \notin J_{\boldsymbol{r}}} \left(\frac{k_j}{r_j}\right)^{k_j} \\
&= \frac{m!}{m^m} \sum_{\substack{\boldsymbol{k} \in S_{m,M}: \\ \forall j \in J_{\boldsymbol{r}}\ k_j = 0}} \prod_{j \notin J_{\boldsymbol{r}}} \frac{k_j^{k_j}}{k_j!} \\
&= \frac{m!}{m^m} \sum_{\substack{\boldsymbol{k} \in S_{m,M}: \\ \forall j \in J_{\boldsymbol{r}}\ k_j = 0}} \prod_{j=1}^{M} \frac{k_j^{k_j}}{k_j!} \qquad \text{(because } \tfrac{0^0}{0!} = 1\text{)} \\
&\leq \frac{m!}{m^m} \sum_{\boldsymbol{k} \in S_{m,M}} \prod_{j=1}^{M} \frac{k_j^{k_j}}{k_j!}.
\end{aligned}$$

Since this is independent of $\boldsymbol{r}$, it also holds after taking the supremum over $\boldsymbol{r} \in \triangle_M$ of the left hand side. We have thus established (5) and hence (3).

Defining $f : \bigcup_{M=2}^{\infty} S_{m,M} \to \mathbb{R}$ by $f(\boldsymbol{k}) = \prod_{j=1}^{|\boldsymbol{k}|} \frac{k_j^{k_j}}{k_j!}$, we see that to establish inequality (4) it suffices to show that

$$\frac{m!}{m^m} \sum_{\boldsymbol{k} \in S_{m,M}} f(\boldsymbol{k}) \leq \sqrt{\pi} e^{1/12m} \left(\frac{m}{2}\right)^{\frac{M-1}{2}} \sum_{z=0}^{M-1} \binom{M}{z} \frac{1}{(\pi m)^{z/2}\, \Gamma\left(\frac{M-z}{2}\right)}. \tag{6}$$

We show this by upper bounding each $f(\boldsymbol{k})$ individually using Stirling's formula

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}, \quad \forall n \geq 1.$$

Since we cannot use this to upper bound $1/k_j!$ when $k_j = 0$, we partition the sum above according to the number of coordinates of $\boldsymbol{k}$ at which $k_j = 0$. Let z index the number of such coordinates. Since f is symmetric under permutations of its arguments, we can write the sum above as

$$\sum_{\boldsymbol{k}\in S_{m,M}} f(\boldsymbol{k}) = \sum_{z=0}^{M-1} \binom{M}{z} \sum_{\boldsymbol{k}\in S^{>0}_{m,M-z}} f(\boldsymbol{k}). \tag{7}$$

For $\boldsymbol{k} \in S^{>0}_{m,M}$ we can now use Stirling's formula to bound $f(\boldsymbol{k})$ as follows

$$f(\boldsymbol{k}) \leq \prod_{j=1}^{M} \frac{k_j^{k_j}}{\sqrt{2\pi k_j}\left(\frac{k_j}{e}\right)^{k_j}} = \prod_{j=1}^{M} \frac{e^{k_j}}{\sqrt{2\pi k_j}} = \frac{e^m}{(2\pi)^{M/2}} \prod_{j=1}^{M} \frac{1}{\sqrt{k_j}}.$$

An application of Lemma 8 now gives

$$\sum_{\boldsymbol{k}\in S^{>0}_{m,M-z}} f(\boldsymbol{k}) \leq \frac{e^m}{(2\pi)^{M/2}} \sum_{\boldsymbol{k}\in S^{>0}_{m,M-z}} \prod_{j=1}^{M} \frac{1}{\sqrt{k_j}} \leq \frac{e^m}{(2\pi)^{\frac{M}{2}}} \frac{\pi^{\frac{M-z}{2}} m^{\frac{M-z-2}{2}}}{\Gamma\left(\frac{M-z}{2}\right)} = \frac{e^m m^{\frac{M-2}{2}}}{2^{\frac{M}{2}} (\pi m)^{z/2} \Gamma\left(\frac{M-z}{2}\right)}.$$

Substituting this into equation (7) and bounding $m!$ using Stirling's formula, we have

$$\begin{aligned}
\frac{m!}{m^m} \sum_{\boldsymbol{k}\in S_{m,M}} f(\boldsymbol{k}) &\leq \frac{\sqrt{2\pi m} e^{1/12m}}{e^m} \sum_{z=0}^{M-1} \binom{M}{z} \frac{e^m m^{\frac{M-2}{2}}}{2^{M/2} (\pi m)^{z/2} \Gamma\left(\frac{M-z}{2}\right)} \\
&= \sqrt{\pi} e^{1/12m} \left(\frac{m}{2}\right)^{\frac{M-1}{2}} \sum_{z=0}^{M-1} \binom{M}{z} \frac{1}{(\pi m)^{z/2} \Gamma\left(\frac{M-z}{2}\right)}.
\end{aligned}$$

which is (6), establishing (4), completing the proof. ■

4. Implied Bounds and Construction of a Differentiable Training Objective

As already discussed, a multitude of bounds can be derived from Theorem 3 and Corollary 7, all of which then hold simultaneously with high probability. For example, suppose after a use of Corollary 7 we have a bound of the form $\mathrm{kl}(\boldsymbol{R}_S(Q)||\boldsymbol{R}_D(Q)) \leq B$. The following proposition then yields the bounds $L_j \leq R^j_D(Q) \leq U_j$, where $L_j := \inf\{p \in [0,1] : \mathrm{kl}(R^j_S(Q)\|p) \leq B\}$ and $U_j := \sup\{p \in [0,1] : \mathrm{kl}(R^j_S(Q)\|p) \leq B\}$. Moreover, since in the worst case we have $\mathrm{kl}(\boldsymbol{R}_S(Q)||\boldsymbol{R}_D(Q)) = B$, the proposition shows that the lower and upper bounds L_j and U_j are the tightest possible, since if $R^j_D(Q) \notin [L_j, U_j]$ then $\mathrm{kl}(R^j_S(Q)\|R^j_D(Q)) > B$ implying $\mathrm{kl}(\boldsymbol{R}_S(Q)||\boldsymbol{R}_D(Q)) > B$. For a more precise version of this argument and a proof of Proposition 9, see Appendix B.3.

Proposition 9 *Let $\boldsymbol{q}, \boldsymbol{p} \in \triangle_M$. Then $\mathrm{kl}(q_j\|p_j) \leq \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p})$ for all $j \in [M]$, with equality when $p_i = \frac{1-p_j}{1-q_j} q_i$. for all $i \neq j$.*

As a second much more interesting example, suppose we can quantify how bad an error of each type is by means of a loss vector $\boldsymbol{\ell} \in [0,\infty)^M$, where ℓ_j is the loss we attribute to an error of type E_j.

We may then be interested in bounding the *total risk* $R_D^T(Q) \in [0,\infty)$ of Q which, recall is defined by $R_D^T(Q) := \sum_{j=1}^M \ell_j R_D^j(Q)$. Indeed, given a bound of the form $\mathrm{kl}(\boldsymbol{R}_S(Q)||\boldsymbol{R}_D(Q)) \leq B$, we can derive

$$R_D^T(Q) \leq \sup\left\{\sum_{j=1}^M \ell_j r_j : \boldsymbol{r} \in \triangle_M, \mathrm{kl}(\boldsymbol{R}_S(Q)||\boldsymbol{r}) \leq B\right\}. \tag{8}$$

This motivates the following definition of $\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$. To see that this is indeed well-defined (at least when $\boldsymbol{u} \in \triangle_M^{>0}$), see the discussion at the beginning of Appendix B.4.

Definition 10 *For $\boldsymbol{u} \in \triangle_M, c \in [0,\infty)$ and $\boldsymbol{\ell} \in [0,\infty)^M$, define $\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ to be an element $\boldsymbol{v} \in \triangle_M$ solving the constrained optimisation problem*

$$\textit{Maximise:} \quad f_{\boldsymbol{\ell}}(\boldsymbol{v}) := \sum_{j=1}^M \ell_j v_j, \tag{9}$$

$$\textit{Subject to:} \quad \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v}) \leq c. \tag{10}$$

Can we calculate $\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ and hence $f_{\boldsymbol{\ell}}(\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c))$ in order to evaluate the bound on the total risk given by (8)? Additionally, if we wish to use the bound on the total risk as a training objective, can we calculate the partial derivatives of $f_{\boldsymbol{\ell}}^*(\boldsymbol{u},c) := f_{\boldsymbol{\ell}}(\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c))$ with respect to the u_j and c so that we can use gradient descent? Our Proposition 11 answers both of these questions in the affirmative, at least in the sense that it provides a speedy method for approximating these quantities to arbitrary precision provided $u_j > 0$ for all $j \in [M]$ and $c > 0$. Indeed, the only approximation step required is that of approximating the unique root of a continuous and strictly increasing scalar function. Thus, provided the u_j themselves are differentiable, Corollary 7 combined with Proposition 11 yields a tractable and fully differentiable objective that can be used for training. More details on how this can be done, including an algorithm written in pseudocode, can be found in Appendix A. While somewhat analogous to the technique used in Clerico et al. (2021) to obtain derivatives of the one-dimensional kl-inverse, our proposition directly yields derivatives on the total risk by (implicitly) employing the envelope theorem (see for example Takayama and Akira, 1985). Since the proof of Proposition 11 is rather long and technical, we defer it to Appendix B.4.

Proposition 11 *Fix $\boldsymbol{\ell} \in [0,\infty)^M$ such that not all ℓ_j are equal, and define $f_{\boldsymbol{\ell}} : \triangle_M \to [0,\infty)$ by $f_{\boldsymbol{\ell}}(\boldsymbol{v}) := \sum_{j=1}^M \ell_j v_j$. For all $\tilde{\boldsymbol{u}} = (\boldsymbol{u}, c) \in \triangle_M^{>0} \times (0,\infty)$, define $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) := \mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c) \in \triangle_M$ and let $\mu^*(\tilde{\boldsymbol{u}}) \in (-\infty, -\max_j \ell_j)$ be the unique solution to $c = \phi_{\boldsymbol{\ell}}(\mu)$, where $\phi_{\boldsymbol{\ell}} : (-\infty, -\max_j \ell_j) \to \mathbb{R}$ is the continuous and strictly increasing function*

$$\phi_{\boldsymbol{\ell}}(\mu) := \log\left(-\sum_{j=1}^M \frac{u_j}{\mu + \ell_j}\right) + \sum_{j=1}^M u_j \log\big(-(\mu + \ell_j)\big).$$

Then $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) = \mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ is given by

$$\boldsymbol{v}^*(\tilde{\boldsymbol{u}})_j = \frac{\lambda^*(\tilde{\boldsymbol{u}}) u_j}{\mu^*(\tilde{\boldsymbol{u}}) + \ell_j} \quad \textit{for } j \in [M], \quad \textit{where} \quad \lambda^*(\tilde{\boldsymbol{u}}) = \left(\sum_{j=1}^M \frac{u_j}{\mu^*(\tilde{\boldsymbol{u}}) + \ell_j}\right)^{-1}.$$

Further, defining $f^*_{\boldsymbol{\ell}} : \triangle_M^{>0} \times (0,\infty) \to [0,\infty)$ *by* $f^*_{\boldsymbol{\ell}}(\tilde{\boldsymbol{u}}) := f_{\boldsymbol{\ell}}(\boldsymbol{v}^*(\tilde{\boldsymbol{u}}))$*, we have that*

$$\frac{\partial f^*_{\boldsymbol{\ell}}}{\partial u_j}(\tilde{\boldsymbol{u}}) = \lambda^*(\tilde{\boldsymbol{u}}) \left(1 + \log \frac{u_j}{\boldsymbol{v}^*(\tilde{\boldsymbol{u}})_j}\right) \qquad \textit{and} \qquad \frac{\partial f^*_{\boldsymbol{\ell}}}{\partial c}(\tilde{\boldsymbol{u}}) = -\lambda^*(\tilde{\boldsymbol{u}}).$$

5. Perspectives

We have established a novel type of PAC-Bayes generalisation bound by abstracting to a general setting of discretised error types. We intend to carry on adapting this to different learning problems, including structured output prediction (as investigated by Cantelobre et al., 2020, in the PAC-Bayes setting), multi-task learning and the learning-to-learn framework (among many references, see *e.g.* Maurer et al., 2016). Besides these exciting theoretical developments which we will address in follow-up works, we aim to put our bounds to the test, in particular adapting and applying the algorithm given in Appendix A, obtaining numerical results for real-world learning problems.

Acknowledgments

Reuben Adams acknowledges support from the Foundational AI Centre for Doctoral Training at University College London, funded by EPSRC (under grant number 2408098). Benjamin Guedj and John Shawe-Taylor acknowledge partial support by the U.S. Army Research Laboratory and the U.S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1. Benjamin Guedj acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02.

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. 2021. URL https://www.arxiv.org/abs/2110.11216.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 9–16. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/3f5ee243547dee91fbd053c1c4a845a

Loubna Benabbou and Pascal Lang. PAC-Bayesian generalization bound for multi-class learning. In *NIPS 2017 Workshop. (Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights*, 2017. URL https://bguedj.github.io/nips2017/pdf/PAC-Bayes_2017_paper_3.pdf.

Felix Biggs and Benjamin Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10):1280, 2021. doi: 10.3390/e23101280. URL https://doi.org/10.3390/e23101280.

Felix Biggs and Benjamin Guedj. On margins and derandomisation in PAC-Bayes. In *AISTATS*, 2022a. URL https://arxiv.org/abs/2107.03955.

Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. Submitted., 2022b. URL https://arxiv.org/abs/2202.01627.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. In *Séminaire de Probabilités XII*, pages 342–363. Springer, 1978. URL http://www.numdam.org/item/SPS_1978__12__342_0.pdf.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Bounds based on the Rényi Divergence. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 435–444, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/begin16.html.

Théophile Cantelobre, Benjamin Guedj, María Pérez-Ortiz, and John Shawe-Taylor. A PAC-Bayesian perspective on structured prediction with implicit loss embeddings. Submitted, 2020. URL https://arxiv.org/abs/2012.03780.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics (IMS) Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, 2007. ISBN 9780940600720. URL https://books.google.fr/books?id=acnaAAAAMAAJ.

Eugenio Clerico, George Deligiannidis, and Arnaud Doucet. Conditional Gaussian PAC-Bayes. *arXiv preprint arXiv:2110.11886*, 2021.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

MD Donsker and SRS Varadhan. Large deviations for markov processes and the asymptotic evaluation of certain markov process expectations for large times. In *Probabilistic Methods in Differential Equations*, pages 82–88. Springer, 1975.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence [UAI]*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1376–1385. PMLR, 2018. URL http://proceedings.mlr.press/v80/dziugaite18a.html.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 604–612. PMLR, 2021. URL http://proceedings.mlr.press/v130/karolina-dziugaite21a.html.

Vasilii Feofanov, Emilie Devijver, and Massih-Reza Amini. Transductive bounds for the multi-class majority vote classifier. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3566–3573, 2019. doi: 10.1609/aaai.v33i01.33013566. URL https://ojs.aaai.org/index.php/AAAI/article/view/4236.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019. URL https://arxiv.org/abs/1901.05353.

Sokol Koço and Cécile Capponi. On multi-class classification through the minimization of the confusion matrix norm. In Cheng Soon Ong and Tu Bao Ho, editors, *Proceedings of the 5th Asian Conference on Machine Learning*, volume 29 of *Proceedings of Machine Learning Research*, pages 277–292, Australian National University, Canberra, Australia, 13–15 Nov 2013. PMLR. URL https://proceedings.mlr.press/v29/Koco13.html.

Charles Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publication in Statistics*, 2:125–141, 1955.

François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*, 219:15–25, 2017. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2016.09.016. URL https://www.sciencedirect.com/science/article/pii/S0925231216310177.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and Francois Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6872–6882. Curran Associates, Inc., 2019.

Guy Lever, François Laviolette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *International Conference on Algorithmic Learning Theory*, pages 119–133. Springer, 2010.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, February 2013. ISSN 0304-3975. doi: 10.1016/j.tcs.2012.10.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0304397512009346.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17:81:1–81:32, 2016. URL http://jmlr.org/papers/v17/15-242.html.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.

Emilie Morvant, Sokol Koço, and Liva Ralaivola. PAC-Bayesian generalization bound on confusion matrix for multi-class classification. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/434.pdf.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13:3507–3531, 2012. URL http://dl.acm.org/citation.cfm?id=2503353.

Maria Perez-Ortiz, Omar Rivasplata, Benjamin Guedj, Matthew Gleeson, Jingyu Zhang, John Shawe-Taylor, Miroslaw Bober, and Josef Kittler. Learning PAC-Bayes priors for probabilistic neural networks. Submitted, 2021a. URL https://arxiv.org/abs/2109.10304.

Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021b. URL http://jmlr.org/papers/v22/20-879.html.

Omar Rivasplata, Csaba Szepesvári, John Shawe-Taylor, Emilio Parrado-Hernández, and Shiliang Sun. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9234–9244, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/386854131f58a556343e056f03626e0

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

Ton Steerneman. On the total variation and Hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society*, 88(4):684–688, 1983.

Akira Takayama and Takayama Akira. *Mathematical economics*. Cambridge university press, 1985.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500. URL https://doi.org/10.1109/TIT.2014.2320500.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJgqqsAct7.

Appendix A. Recipe for implementing and deploying our strategy

We here outline more explicitly how Corollary 7 and Proposition 11 may be used to formulate a fully differentiable objective by which a model may be trained.

First, if one wishes to make hard labels, namely $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, it will first be necessary to use a surrogate class of soft hypotheses $\mathcal{H}' \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ during training, before reverting to hard labels for example by taking the mean label or the one with highest probability. Using soft hypotheses during training is necessary to ensure that the empirical j-risks $R_S^j(Q)$ are differentiable in the model parameters. Since how one chooses to do this will depend on the specific use case, we restrict our attention here to the case of soft hypotheses. Specifically, we consider a class of soft hypotheses $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^N\} \subseteq \mathcal{M}(\mathcal{Y})^{\mathcal{X}}$ parameterised by the weights $\theta \in \mathbb{R}^N$ of some neural network of a given architecture with N parameters in such a way that the $R_S^j(h_\theta)$ are differentiable in θ. A concrete example would be multiclass classification using a fully connected neural network with output being softmax probabilities on the classes so that the $R_S^j(h_\theta)$ are differentiable.

Second, it is necessary to restrict the prior and posterior $P, Q \in \mathcal{M}(\mathcal{H})$ to a parameterised subset of $\mathcal{M}(\mathcal{H})$ in which $\mathrm{KL}(Q\|P)$ has a closed form which is differentiable in the parameterisation. A simple choice for our case of a neural network with N parameters is $P, Q \in \{\mathcal{N}(\boldsymbol{w}, \mathrm{diag}(\boldsymbol{s})) : \boldsymbol{w} \in \mathbb{R}^N, \boldsymbol{s} \in \mathbb{R}^N_{>0}\}$. For prior a $P_{\boldsymbol{v},\boldsymbol{r}} = \mathcal{N}(\boldsymbol{v}, \mathrm{diag}(\boldsymbol{r}))$ and posterior $Q_{\boldsymbol{w},\boldsymbol{s}} = \mathcal{N}(\boldsymbol{w}, \mathrm{diag}(\boldsymbol{s}))$ we have the closed form

$$\mathrm{KL}(Q_{\boldsymbol{w},\boldsymbol{s}}\|P_{\boldsymbol{v},\boldsymbol{r}}) = \frac{1}{2}\left[\sum_{n=1}^{N}\left(\frac{s_n}{r_n} + \frac{(w_n - v_n)^2}{r_n} + \ln\frac{r_n}{s_n}\right) - N\right],$$

which is indeed differentiable in $\boldsymbol{v}, \boldsymbol{r}, \boldsymbol{w}$ and $\boldsymbol{s}$. While $Q_{\boldsymbol{w},\boldsymbol{s}}$ and $P_{\boldsymbol{v},\boldsymbol{r}}$ are technically distributions on $\mathbb{R}^D$ rather than $\mathcal{H}$, the KL-divergence between the distributions they induce on $\mathcal{H}$ will be at most as large as the expression above. Thus, substituting the expression above into the bounds we prove in Section 3 can only increase the value of the bounds, meaning the enlarged bounds certainly still hold with probability at least $1 - \delta$.

Third, in all but the simplest cases $R_S^j(Q_{\boldsymbol{w},\boldsymbol{s}})$ will not have a closed form, much less one that is differentiable in $\boldsymbol{w}$ and $\boldsymbol{s}$. A common solution to this is to use the so-called pathwise gradient estimator. In our case, this corresponds to drawing $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I})$, where $\mathbb{I}$ is the $N \times N$ identity matrix, and estimating

$$\nabla_{\boldsymbol{w},\boldsymbol{s}} R_S^j(Q_{\boldsymbol{w},\boldsymbol{s}}) = \nabla_{\boldsymbol{w},\boldsymbol{s}}\left[\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(\boldsymbol{0},\mathbb{I})} R_S^j(h_{\boldsymbol{w}+\boldsymbol{\epsilon}'\odot\sqrt{\boldsymbol{s}}})\right] \approx \nabla_{\boldsymbol{w},\boldsymbol{s}} R_S^j(h_{\boldsymbol{w}+\boldsymbol{\epsilon}\odot\sqrt{\boldsymbol{s}}}),$$

where $h_{\boldsymbol{w}}$ denotes the function expressed by the neural network with parameters $\boldsymbol{w}$. For a proof that this is an unbiased estimator, and for other methods for estimating the gradients of expectations, see the survey Mohamed et al. (2020).

Fourth, one must choose the prior. Designing priors which are optimal in some sense (*i.e.*, minimising the Kullback-Leibler term in the right-hand side of generalisation bounds) has been at the core of an active line of work in the PAC-Bayesian literature. For the sake of simplicity, and since it is out of the scope of our contributions, we assume here that the prior is given beforehand, although we stress that practitioners should pay great attention to its tuning. For our purposes, it suffices to say that if one is using a data-dependent prior then it is necessary to partition the sample into $S = S_{\text{Prior}} \cup S_{\text{Bound}}$, where S_{Prior} is used to train the prior and S_{Bound} is used to evaluate the bound. Since our bound holds uniformly over posteriors $Q \in \mathcal{M}(\mathcal{H})$, the entire sample S is free to be used to train the posterior Q. For a more in-depth discussion on the choice of prior, we refer to the following body of work: Ambroladze et al. (2006), Lever et al. (2010, 2013), Parrado-Hernández et al. (2012), Dziugaite and Roy (2017, 2018), Rivasplata et al. (2018), Letarte et al. (2019), Perez-Ortiz et al. (2021a), Dziugaite et al. (2021), Biggs and Guedj (2021, 2022a,b).

Finally, given a confidence level $\delta \in (0, 1]$, one may use Algorithm 1 to obtain a posterior $Q_{\boldsymbol{w},\boldsymbol{s}}$ with minimal upper bound on the total risk. Note we take the pointwise logarithm of the variances $\boldsymbol{r}$ and $\boldsymbol{s}$ to obtain unbounded parameters on which to perform stochastic gradient descent or some other minimisation algorithm. We use $\oplus$ to denote vector concatenation. The algorithm can be straightforwardly adapted to permit mini-batches by, for each epoch, sequentially repeating the steps with S equal to each mini-batch.

Algorithm 1: Calculating a posterior with minimal bound on the total risk.

Input:

$\mathcal{X}, \mathcal{Y}$ /* Arbitrary input and output spaces */
$\bigcup_{j=1}^{M} E_j = \mathcal{Y}^2$ /* A finite partition into error types */
$\boldsymbol{\ell} \in [0, \infty)^M$ /* A vector of losses, not all equal */
$S = S_{\text{Prior}} \cup S_{\text{Bound}} \in (\mathcal{X} \times \mathcal{Y})^m$ /* A partitioned i.i.d. sample */
$N \in \mathbb{N}$ /* The number of model parameters */
$P_{\boldsymbol{v},\boldsymbol{r}}$, $\boldsymbol{v}(S_{\text{Prior}}) \in \mathbb{R}^N, \boldsymbol{r}(S_{\text{Prior}}) \in \mathbb{R}^N_{\geq 0}$ /* A (data-dependent) prior */
$Q_{\boldsymbol{w}_0,\boldsymbol{s}_0}$, $\boldsymbol{w}_0 \in \mathbb{R}^N, \boldsymbol{s}_0 \in \mathbb{R}^N_{\geq 0}$ /* An initial posterior */
$\delta \in (0, 1]$ /* A confidence level */
$\lambda > 0$ /* A learning rate */
T /* The number of epochs to train for */

Output:

$Q_{\boldsymbol{w},\boldsymbol{s}}$, $\boldsymbol{w} \in \mathbb{R}^N, \boldsymbol{s} \in \mathbb{R}^N_{\geq 0}$ /* A trained posterior */

Procedure:

$\boldsymbol{\zeta}_0 \leftarrow \log \boldsymbol{s}_0$ /* Transform to unbounded scale parameters */
$\boldsymbol{p} \leftarrow \boldsymbol{w}_0 \oplus \boldsymbol{\zeta}_0$ /* Collect mean and scale parameters */
for $t \leftarrow 1$ **to** T **do**
  Draw $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$
  $\boldsymbol{u} \leftarrow \boldsymbol{R}_S\left(h_{\boldsymbol{w}+\boldsymbol{\epsilon}\odot\sqrt{\exp(\boldsymbol{\zeta})}}\right)$
  $B \leftarrow \frac{1}{m}\left[\mathrm{KL}\left(Q_{\boldsymbol{w},\exp(\boldsymbol{\zeta})} \middle\| P_{\boldsymbol{v},\boldsymbol{r}}\right) + \ln\left(\frac{1}{\delta}\sqrt{\pi}e^{1/12m}\left(\frac{m}{2}\right)^{\frac{M-1}{2}}\sum_{z=0}^{M-1}\binom{M}{z}\frac{1}{(\pi m)^{z/2}\Gamma\left(\frac{M-z}{2}\right)}\right)\right]$
  $\tilde{\boldsymbol{u}} \leftarrow (u_1, \ldots, u_M, B)$
  $\boldsymbol{G} \leftarrow \mathbf{0}_{2N\times(M+1)}$ /* Initialise gradient matrix */
  $\boldsymbol{F} \leftarrow \mathbf{0}_{M+1}$ /* Initialise gradient vector */
  for $j \leftarrow 1$ **to** $M+1$ **do**
    $\boldsymbol{F}_j \leftarrow \frac{\partial f^*_{\boldsymbol{\ell}}}{\partial \tilde{u}_j}(\tilde{\boldsymbol{u}})$ /* Gradients of total loss from Prop 11 */
    for $i \leftarrow 1$ **to** $2N$ **do**
      $\boldsymbol{G}_{i,j} \leftarrow \frac{\partial \tilde{u}_j}{\partial p_i}(\boldsymbol{p})$ /* Gradients of empirical risks and bound */
    end
  end
  $\boldsymbol{H} \leftarrow \boldsymbol{G}\boldsymbol{F}$ /* Gradients of total loss w.r.t. parameters */
  $\boldsymbol{p} \leftarrow \boldsymbol{p} - \lambda \boldsymbol{H}$ /* Gradient step */
end
$\boldsymbol{w} = (p_1, \ldots, p_N)$
$\boldsymbol{s} = (\exp(p_{N+1}), \ldots, \exp(p_{2N}))$
return $\boldsymbol{w}, \boldsymbol{s}$

Appendix B. Proofs

B.1. Proof of Lemma 5

Let $\boldsymbol{E}_M := \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M\}$, namely the set of M-dimensional basis vectors. We will denote a typical element of $\boldsymbol{E}_M^m$ by $\boldsymbol{\eta}^{(m)} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m)$. For any $\boldsymbol{x}^{(m)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \in \triangle_M^m$, a straightforward induction on m yields

$$\sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^m \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) = 1. \tag{11}$$

To see this, for $m = 1$ we have $\boldsymbol{E}_M^1 = \{(\boldsymbol{e}_1,), \ldots, (\boldsymbol{e}_M,)\}$, where we have been pedantic in using 1-tuples to maintain consistency with larger values of m. Thus, for any $\boldsymbol{x}^{(1)} = (\boldsymbol{x}_1,) \in \triangle_M^1$, the left hand side of equation (11) can be written as

$$\sum_{j=1}^M \boldsymbol{x}_1 \cdot \boldsymbol{e}_j = \sum_{j=1}^M (\boldsymbol{x}_1)_j = 1.$$

Now suppose that equation (11) holds for any $\boldsymbol{x}^{(m)} \in \triangle_M^m$ and let $\boldsymbol{x}^{(m+1)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m+1}) \in \triangle_M^{m+1}$. Then the left hand side of equation (11) can be written as

$$\begin{aligned} \sum_{\boldsymbol{\eta}^{(m+1)} \in \boldsymbol{E}_M^{m+1}} \left(\prod_{i=1}^{m+1} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) &= \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \sum_{j=1}^M \left(\prod_{i=1}^m \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) \\ &= \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^m \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) \sum_{j=1}^M (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) = 1. \end{aligned}$$

We now show that any $\boldsymbol{x}^{(m)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \in \triangle_M^m$ can be written as a convex combination of the elements of $\boldsymbol{E}_M^m$ in the following way

$$\boldsymbol{x}^{(m)} = \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^m \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) \boldsymbol{\eta}^{(m)}. \tag{12}$$

We have already shown that the weights sum to one, and they are clearly elements of $[0, 1]$, so the right hand side of equation (12) is indeed a convex combination of the elements of $\boldsymbol{E}_M^m$. We now show that equation (12) holds, again by induction.

For $m = 1$ and any $\boldsymbol{x}^{(1)} = (\boldsymbol{x}_1,) \in \triangle_M^1$, the right hand side of equation (12) can be written as

$$\sum_{j=1}^M (\boldsymbol{x}_1 \cdot \boldsymbol{e}_j)(\boldsymbol{e}_j,) = (\boldsymbol{x}_1,) = \boldsymbol{x}.$$

For the inductive hypothesis, suppose equation (12) holds for some arbitrary $m \geq 1$, and denote elements of $\boldsymbol{E}_M^{m+1}$ by $\boldsymbol{\eta}^{(m)} \oplus (\boldsymbol{e},)$ for some $\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m$ and $\boldsymbol{e} \in \boldsymbol{E}_M$, where $\oplus$ denotes vector concatenation. Then for any $\boldsymbol{x}^{(m+1)} = \boldsymbol{x}^{(m)} \oplus (\boldsymbol{x}_{m+1},) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m+1}) \in \triangle_M^{m+1}$, the right hand side of equation (12) can be written as

$$\sum_{\boldsymbol{\eta}^{(m+1)} \in \boldsymbol{E}_M^{m+1}} \left(\prod_{i=1}^{m+1} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) \boldsymbol{\eta}^{(m+1)} = \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \sum_{j=1}^M \left(\prod_{i=1}^m \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) \boldsymbol{\eta}^{(m)} \oplus (\boldsymbol{e}_j,)$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \sum_{j=1}^{M} \left(\prod_{i=1}^{m} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) \boldsymbol{\eta}^{(m)} \\
&\oplus \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \sum_{j=1}^{M} \left(\prod_{i=1}^{m} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) (\boldsymbol{e}_j,) \\
&= \sum_{j=1}^{M} (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^{m} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) \boldsymbol{\eta}^{(m)} \\
&\oplus \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^{m} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) \sum_{j=1}^{M} (\boldsymbol{x}_{m+1} \cdot \boldsymbol{e}_j) (\boldsymbol{e}_j,) \\
&= 1 \cdot \boldsymbol{x}^{(m)} \oplus 1 \cdot (\boldsymbol{x}_{m+1},) = \boldsymbol{x}^{(m+1)},
\end{aligned}
$$

where in the penultimate inequality we have used the inductive hypothesis and (twice) the result of the previous induction.

We can now prove the statement of the Lemma. Applying Jensen's inequality to equation (12) with the convex function f, we have that

$$
\begin{aligned}
f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) &= f \left(\sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^{m} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) \boldsymbol{\eta}^{(m)} \right) \\
&\leq \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^{m} \boldsymbol{x}_i \cdot \boldsymbol{\eta}_i \right) f \left(\boldsymbol{\eta}^{(m)} \right).
\end{aligned}
$$

Let $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_1]$ denote the mean of the i.i.d. random vectors X_i. Then the above inequality implies

$$
\begin{aligned}
\mathbb{E}[f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)] &\leq \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^{m} \boldsymbol{\mu} \cdot \boldsymbol{\eta}_i \right) f \left(\boldsymbol{\eta}^{(m)} \right) \\
&= \sum_{\boldsymbol{\eta}^{(m)} \in \boldsymbol{E}_M^m} \left(\prod_{i=1}^{m} \mathbb{P}(\boldsymbol{X}_i' = \boldsymbol{\eta}_i) \right) f \left(\boldsymbol{\eta}^{(m)} \right) \\
&= \mathbb{E}[f(\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')].
\end{aligned}
$$

B.2. Proof of Lemma 8

The proof of Lemma 8 itself requires two technical helping lemmas which we now state and prove.

Lemma 12 *For any integers $n \geq 2$ and $p \geq -1$,*

$$
\sum_{k=1}^{n-1} \frac{(n-k)^{p/2}}{\sqrt{k}} \leq n^{\frac{p+1}{2}} \int_0^1 \frac{(1-x)^{p/2}}{\sqrt{x}} dx.
$$

Proof The case of $p = -1$, namely

$$\sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}} \leq \int_0^1 \frac{1}{\sqrt{x(1-x)}} dx,$$

has already been demonstrated in Maurer (2004). For $p > -1$, let

$$f_p(x) := \frac{(1-x)^{p/2}}{\sqrt{x}}.$$

We will show that each $f_p(\cdot)$ is monotonically decreasing on $(0, 1)$. Indeed,

$$\frac{df_p}{dx}(x) = -\frac{(1-x)^{\frac{p}{2}-1}(px+1-x)}{2x^{3/2}} \leq -\frac{(1-x)^{p/2}}{2x^{3/2}} < 0,$$

where for the inequalities we have used the fact that $p > -1$ and $x \in (0, 1)$. We therefore see that

$$\begin{aligned}
\sum_{k=1}^{n-1} \frac{(n-k)^{p/2}}{\sqrt{k}} &= \sum_{k=1}^{n-1} \frac{n^{p/2}(1-\frac{k}{n})^{p/2}}{\sqrt{n}\sqrt{\frac{k}{n}}} \\
&= n^{\frac{p+1}{2}} \sum_{k=1}^{n-1} \frac{1}{n} \frac{(1-\frac{k}{n})^{p/2}}{\sqrt{\frac{k}{n}}} \\
&= n^{\frac{p+1}{2}} \sum_{k=1}^{n-1} \frac{1}{n} f_p\left(\frac{k}{n}\right) \\
&\leq n^{\frac{p+1}{2}} \sum_{k=1}^{n-1} \int_{\frac{k-1}{n}}^{\frac{k}{n}} f_p(x) dx \\
&= n^{\frac{p+1}{2}} \int_0^{1-\frac{1}{n}} f_p(x) dx \\
&\leq n^{\frac{p+1}{2}} \int_0^1 f_p(x) dx.
\end{aligned}$$

■

Intuitively, the proof of the above lemma works by bounding the integral below by a Riemann sum. In the following lemma we actually calculate this integral, yielding a more explicit bound on the sum in Lemma 12. We found it is easier to calculate a slightly more general integral, where the 1 in the limit and the integrand is replaced by a positive constant a.

Lemma 13 *For any real number $a > 0$ and integer $n \geq -1$,*

$$\int_0^a \frac{(a-x)^{n/2}}{\sqrt{x}} dx = \sqrt{\pi} \frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+3}{2})} a^{\frac{n+1}{2}}.$$

Proof Define

$$\mathrm{I}_n(a) := \int_0^a \frac{(a-x)^{n/2}}{\sqrt{x}} dx \qquad \text{and} \qquad f_n(a) := \sqrt{\pi}\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+3}{2})} a^{\frac{n+1}{2}}.$$

We proceed by induction, increasing n by 2 each time. This means we need two base cases. First, for $n = -1$, we have

$$\mathrm{I}_{-1}(a) = \int_0^a \frac{1}{\sqrt{x(a-x)}} dx = \left[2 \arcsin \sqrt{\frac{x}{a}}\right]_0^a = \pi = f_{-1}(a),$$

since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\Gamma(1) = 1$. Second, for $n = 0$,

$$\mathrm{I}_0(a) = \int_0^a \frac{1}{\sqrt{x}} dx = \left[2\sqrt{x}\right]_0^a = 2\sqrt{a} = f_0(a),$$

since $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$. Now, by the Leibniz integral rule, we have

$$\frac{d}{da}\mathrm{I}_{n+2}(a) = \int_0^a \frac{\partial}{\partial a} \frac{(a-x)^{\frac{n+2}{2}}}{\sqrt{x}} dx = \frac{n+2}{2} \int_0^a \frac{(a-x)^{\frac{n}{2}}}{\sqrt{x}} dx = \frac{n+2}{2}\mathrm{I}_n(a).$$

Thus

$$\mathrm{I}_{n+2}(a) = \frac{n+2}{2}\left[\int_0^a \mathrm{I}_n(t)dt + \mathrm{I}_n(0)\right] = \frac{n+2}{2}\int_0^a \mathrm{I}_n(t)dt,$$

since $\mathrm{I}_n(0) = 0$.

Now, for the inductive step, suppose $\mathrm{I}_n(a) = f_n(a)$ for some $n \geq -1$. Then, using the previous calculation, we have

$$\begin{aligned}
\mathrm{I}_{n+2}(a) &= \frac{n+2}{2}\int_0^a f_n(t)dt \\
&= \frac{n+2}{2}\int_0^a \sqrt{\pi}\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+3}{2})} t^{\frac{n+1}{2}} dt \\
&= \sqrt{\pi}\frac{\frac{n+2}{2}\Gamma(\frac{n+2}{2})}{\frac{n+3}{2}\Gamma(\frac{n+3}{2})} a^{\frac{n+3}{2}} \\
&= \sqrt{\pi}\frac{\Gamma(\frac{n+2}{2}+1)}{\Gamma(\frac{n+3}{2}+1)} a^{\frac{n+3}{2}} \\
&= \sqrt{\pi}\frac{\Gamma\left(\frac{(n+2)+2}{2}\right)}{\Gamma\left(\frac{(n+2)+3}{2}\right)} a^{\frac{(n+2)+1}{2}} \\
&= f_{n+2}(a).
\end{aligned}$$

This completes the proof. ■

We are now ready to prove Lemma 8 which, for ease of reference, we restate here. For integers $M \geq 1$ and $m \geq M$,

$$\sum_{\boldsymbol{k} \in S^{>0}_{m,M}} \frac{1}{\prod_{j=1}^M \sqrt{k_j}} \leq \frac{\pi^{\frac{M}{2}} m^{\frac{M-2}{2}}}{\Gamma(\frac{M}{2})}.$$

Proof (of Lemma 8) We proceed by induction on M. For $M = 1$, the set $S_{m,M}$ contains a single element, namely the one-dimensional vector $\boldsymbol{k} = (k_1,) = (m,)$. In this case, the left hand side is $1/\sqrt{m}$ while the right hand side is $\sqrt{\pi}/(\sqrt{m}\Gamma(1/2)) = 1/\sqrt{m}$, since $\Gamma(1/2) = \sqrt{\pi}$.

Now, as the inductive hypothesis, assume (8) holds for some fixed $M \geq 1$ and all $m \geq M$. Then for all $m \geq M+1$, we have

$$\begin{aligned}
\sum_{\boldsymbol{k}\in S^{>0}_{m,M+1}} \frac{1}{\prod_{j=1}^{M+1}\sqrt{k_j}} &= \sum_{k_1=1}^{m-M} \frac{1}{\sqrt{k_1}} \sum_{\boldsymbol{k}'\in S^{>0}_{m-k_1,M}} \frac{1}{\prod_{j=1}^{M}\sqrt{k'_j}} \\
&\leq \sum_{k_1=1}^{m-M} \frac{1}{\sqrt{k_1}} \frac{\pi^{\frac{M}{2}}(m-k_1)^{\frac{M-2}{2}}}{\Gamma(\frac{M}{2})} \qquad \text{(by the inductive hypothesis)} \\
&= \frac{\pi^{\frac{M}{2}}}{\Gamma(\frac{M}{2})} \sum_{k_1=1}^{m-M} \frac{(m-k_1)^{\frac{M-2}{2}}}{\sqrt{k_1}} \\
&\leq \frac{\pi^{\frac{M}{2}}}{\Gamma(\frac{M}{2})} \sum_{k_1=1}^{m-1} \frac{(m-k_1)^{\frac{M-2}{2}}}{\sqrt{k_1}} \qquad \text{(enlarging the sum domain)} \\
&\leq \frac{\pi^{\frac{M}{2}}}{\Gamma(\frac{M}{2})} m^{\frac{M-1}{2}} \int_0^1 \frac{(1-x)^{\frac{M-2}{2}}}{\sqrt{x}} dx \qquad \text{(by Lemma 12)} \\
&= \frac{\pi^{\frac{M}{2}}}{\Gamma(\frac{M}{2})} m^{\frac{M-1}{2}} \sqrt{\pi} \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{M+1}{2})} \qquad \text{(by Lemma 13)} \\
&= \frac{\pi^{\frac{M+1}{2}} m^{\frac{M-1}{2}}}{\Gamma(\frac{M+1}{2})},
\end{aligned}$$

as required. ■

B.3. Proof of Proposition 9

Proof The case where $q_j = 1$ or $p_j = 1$ can be dealt with trivially by splitting into the three following subcases

- $q_j = p_j = 1 \implies \mathrm{kl}(q_j \| p_j) = \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) = 0$
- $q_j = 1, p_j \neq 1 \implies \mathrm{kl}(q_j \| p_j) = \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) = -\log p_j$
- $q_j \neq 1, p_j = 1 \implies \mathrm{kl}(q_j \| p_j) = \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) = \infty.$

For $q_j \neq 1$ and $p_j \neq 1$ define the distributions $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}} \in \triangle_M$ by $\tilde{q}_j = \tilde{p}_j = 0$ and

$$\tilde{q}_i = \frac{q_i}{1-q_j} \qquad \text{and} \qquad \tilde{p}_i = \frac{p_i}{1-p_j}$$

for $i \neq j$. Then

$$\sum_{i\neq j} q_i \log \frac{q_i}{p_i} = \sum_{i\neq j} (1-q_j)\tilde{q}_i \log \frac{(1-q_j)\tilde{q}_i}{(1-p_j)\tilde{p}_i}$$

$$
\begin{aligned}
&= (1-q_j)\sum_{i\neq j}\tilde{q}_i \log\frac{\tilde{q}_i}{\tilde{p}_i} + \tilde{q}_i\log\frac{1-q_j}{1-p_j} \\
&= (1-q_j)\mathrm{kl}(\tilde{\boldsymbol{q}}\|\tilde{\boldsymbol{p}}) + (1-q_j)\log\frac{1-q_j}{1-p_j} \\
&\geq (1-q_j)\log\frac{1-q_j}{1-p_j}.
\end{aligned}
$$

The final inequality holds since $\mathrm{kl}(\tilde{\boldsymbol{q}}\|\tilde{\boldsymbol{p}}) \geq 0$. Further, note that we have equality if and only if $\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{p}}$, which, by their definitions, translates to

$$p_i = \frac{1-p_j}{1-q_j}q_i$$

for all $i \neq j$. If we now add $q_j \log \frac{q_j}{p_j}$ to both sides, we obtain

$$\mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) \geq (1-q_j)\log\frac{1-q_j}{1-p_j} + q_j\log\frac{q_j}{p_j} = \mathrm{kl}(q_j\|p_j),$$

with the same condition for equality. ■

The following proposition makes more precise the argument found at the beginning of Section 4 for how Proposition 9 can be used to derive the tightest possible lower and upper bounds on each $R_D^j(Q)$.

Proposition 14 *Suppose that $\boldsymbol{q}, \boldsymbol{p} \in \triangle_M$ are such that $\mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) \leq B$, where $\boldsymbol{q}$ is known and $\boldsymbol{p}$ is unknown. Then, in the absence of any further information, the tightest bound that can be obtained on each p_j is*

$$p_j \leq \mathrm{kl}^{-1}(q_j, B).$$

Proof Suppose $p_j > \mathrm{kl}^{-1}(q_j, B)$. Then, by definition of kl^{-1}, we have that $\mathrm{kl}(q_j\|p_j) > B$. By Proposition 9, this would then imply $\mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) > B$, contradicting our assumption. Therefore $p_j \leq \mathrm{kl}^{-1}(q_j, B)$. Now, with the information we have, we cannot rule out that

$$p_i = \frac{1-p_j}{1-q_j}q_i$$

for all $i \neq j$ and thus, by Proposition 9, that $\mathrm{kl}(q_j\|p_j) = \mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p})$. Further, we cannot rule out that $\mathrm{kl}(\boldsymbol{q}\|\boldsymbol{p}) = B$. Thus, it is possible that $\mathrm{kl}(q_j\|p_j) = B$, in which case $p_j = \mathrm{kl}^{-1}(q_j, B)$. We therefore see that $\mathrm{kl}^{-1}(q_j, B)$ is the tightest possible upper bound on p_j, for each $j \in [M]$. ■

B.4. Proof of Proposition 11

Before proving the proposition, we first argue that $\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ given by Definition 10 is well-defined. First, note that $A_{\boldsymbol{u}} := \{\boldsymbol{v} \in \triangle_M : \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v}) \leq c\}$ is compact (boundedness is clear and it is closed because it is the preimage of the closed set $[0, c]$ under the continuous map $\boldsymbol{v} \mapsto \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v})$) and so the continuous function $f_{\boldsymbol{\ell}}$ achieves its supremum on $A_{\boldsymbol{u}}$. Further, note that $A_{\boldsymbol{u}}$ is a convex subset of $\triangle_M$ (because the map $\boldsymbol{v} \mapsto \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v})$ is convex) and $f_{\boldsymbol{\ell}}$ is linear, so the supremum of $f_{\boldsymbol{\ell}}$ over $A_{\boldsymbol{u}}$ is

achieved and is located on the boundary of $A_{\boldsymbol{u}}$. This means we can replace the inequality constraint $\mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v}) \leq c$ in Definition 10 with the equality constraint $\mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v}) = c$. Finally, if $\boldsymbol{u} \in \triangle_M^{>0}$ then $A_{\boldsymbol{u}}$ is a *strictly* convex subset of $\triangle_M$ (because the map $\boldsymbol{v} \mapsto \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v})$ is then *strictly* convex) and so the supremum of $f_{\boldsymbol{\ell}}$ occurs at a *unique* point on the boundary of $A_{\boldsymbol{u}}$. In other words, if $\boldsymbol{u} \in \triangle_M^{>0}$ then $\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ is defined *uniquely*.

Proof (of Proposition 11) We start by deriving the implicit expression for $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) = \mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ given in the proposition by solving a transformed version of the optimisation problem given by (9) and (10) using the method of Lagrange multipliers. We obtain two solutions to the Lagrangian equations, which must correspond to the maximum and minimum total risk over the set $A_{\boldsymbol{u}} := \{\boldsymbol{v} \in \triangle_M : \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v}) \leq c\}$ because, as argued in the main text (see the discussion after Definition 10), $A_{\boldsymbol{u}}$ is compact and so the linear total risk $f_{\boldsymbol{\ell}}(\boldsymbol{v})$ attains its maximum and minimum on $A_{\boldsymbol{u}}$.

By definition of $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) = \mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$, we know that $\mathrm{kl}(\boldsymbol{v}^*(\tilde{\boldsymbol{u}})\|\boldsymbol{u}) \leq c$. Since, by assumption, $u_j > 0$ for all j, we see that $\boldsymbol{v}^*(\tilde{\boldsymbol{u}})_j > 0$ for all j, otherwise we would have $\mathrm{kl}(\boldsymbol{v}^*(\tilde{\boldsymbol{u}})\|\boldsymbol{u}) = \infty$, a contradiction. Thus $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) \in \triangle_M^{>0}$ and we are permitted to instead optimise over the unbounded variable $\boldsymbol{t} \in \mathbb{R}^M$, where $t_j := \ln v_j$. With this transformation, the constraint $\boldsymbol{v} \in \triangle_M$ can be replaced simply by $\sum_j e^{t_j} = 1$ and the optimisation problem becomes

$$\begin{aligned} \text{Maximise:} \quad & F(\boldsymbol{t}) := \sum_{j=1}^{M} \ell_j e^{t_j} \\ \text{Subject to:} \quad & g(\boldsymbol{t}; \boldsymbol{u}, c) := \mathrm{kl}(\boldsymbol{u}\|e^{\boldsymbol{t}}) - c = 0, \\ & h(\boldsymbol{t}) := \sum_{j=1}^{M} e^{t_j} - 1 = 0, \end{aligned}$$

where $e^{\boldsymbol{t}} \in \mathbb{R}^M$ is defined by $(e^{\boldsymbol{t}})_j := e^{t_j}$. Note that $F(\boldsymbol{t}) = f_{\boldsymbol{\ell}}(e^{\boldsymbol{t}})$. Following the terminology of mathematical economics, we call the t_j the *optimisation variables*, and the $\tilde{u}_j$ (namely the u_j and c) the *choice variables*. The vector $\boldsymbol{\ell}$ is considered fixed—we neither want to optimise over it nor differentiate with respect to it—which is why we occasionally suppress it from the notation henceforth.

For each $\tilde{\boldsymbol{u}}$, let $\boldsymbol{v}^*(\tilde{\boldsymbol{u}})$ and $\boldsymbol{t}^*(\tilde{\boldsymbol{u}})$ be the solutions to the original and transformed optimisation problems respectively. Since the map $\boldsymbol{v} = e^{\boldsymbol{t}}$ is one-to-one, it is clear that since $\boldsymbol{v}^*(\tilde{\boldsymbol{u}})$ exists uniquely, so does $\boldsymbol{t}^*(\tilde{\boldsymbol{u}})$, and that they are related by $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) = e^{\boldsymbol{t}^*(\tilde{\boldsymbol{u}})}$. We therefore have the identity

$$f_{\boldsymbol{\ell}}(\boldsymbol{v}^*(\tilde{\boldsymbol{u}})) \equiv F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})).$$

Recalling that $f_{\boldsymbol{\ell}}^*(\tilde{\boldsymbol{u}}) := f_{\boldsymbol{\ell}}(\boldsymbol{v}^*(\tilde{\boldsymbol{u}}))$, we see that

$$\nabla_{\tilde{\boldsymbol{u}}} f_{\boldsymbol{\ell}}^*(\tilde{\boldsymbol{u}}) \equiv \nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})). \tag{13}$$

the derivatives of $f_{\boldsymbol{\ell}}(\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c))$ with respect to $\boldsymbol{u}$ and c are given by $\nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))$.

Using the method of Lagrange multipliers, there exist real numbers $\lambda^* = \lambda^*(\tilde{\boldsymbol{u}})$ and $\mu^* = \mu^*(\tilde{\boldsymbol{u}})$ such that $(\boldsymbol{t}^*, \lambda^*, \mu^*)$ is a stationary point (with respect to $\boldsymbol{t}$, λ and μ) of the Lagrangian function

$$\mathcal{L}(\boldsymbol{t}, \lambda, \mu; \tilde{\boldsymbol{u}}) := F(\boldsymbol{t}) + \lambda g(\boldsymbol{t}; \tilde{\boldsymbol{u}}) + \mu h(\boldsymbol{t}).$$

Let $F_{\boldsymbol{t}}(\cdot)$ and $h_{\boldsymbol{t}}(\cdot)$ denote the gradient vectors of F and h respectively, and let $g_{\boldsymbol{t}}(\,\cdot\,;\tilde{\boldsymbol{u}})$ and $g_{\tilde{\boldsymbol{u}}}(\boldsymbol{t};\,\cdot\,)$ denote the gradient vectors of g with respect to $\boldsymbol{t}$ only and $\tilde{\boldsymbol{u}}$ only, respectively. Simple calculation yields

$$\begin{aligned} g_{\boldsymbol{t}}(\boldsymbol{t};\tilde{\boldsymbol{u}}) &= \left(\frac{\partial g}{\partial t_1}(\boldsymbol{t};\tilde{\boldsymbol{u}}),\ldots,\frac{\partial g}{\partial t_M}(\boldsymbol{t};\tilde{\boldsymbol{u}})\right) = -\boldsymbol{u} \quad \text{and} \\ g_{\tilde{\boldsymbol{u}}}(\boldsymbol{t};\tilde{\boldsymbol{u}}) &= \left(\frac{\partial g}{\partial \tilde{u}_1}(\boldsymbol{t};\tilde{\boldsymbol{u}}),\ldots,\frac{\partial g}{\partial \tilde{u}_{M+1}}(\boldsymbol{t};\tilde{\boldsymbol{u}})\right) = \Big(1-t_1+\log u_1,\ldots,1-t_M+\log u_M,-1\Big). \end{aligned} \tag{14}$$

Then, taking the partial derivatives of $\mathcal{L}$ with respect to λ, μ and the t_j, we have that $(\boldsymbol{t},\lambda,\mu) = (\boldsymbol{t}^*(\tilde{\boldsymbol{u}}),\lambda^*(\tilde{\boldsymbol{u}}),\mu^*(\tilde{\boldsymbol{u}}))$ solves the simultaneous equations

$$F_{\boldsymbol{t}}(\boldsymbol{t}) + \lambda g_{\boldsymbol{t}}(\boldsymbol{t};\tilde{\boldsymbol{u}}) + \mu h_{\boldsymbol{t}}(\boldsymbol{t}) = \mathbf{0}, \tag{15}$$

$$g(\boldsymbol{t};\tilde{\boldsymbol{u}}) = 0, \quad \text{and}$$

$$h(\boldsymbol{t}) = 0,$$

where the last two equations recover the constraints. Substituting the gradients $F_{\boldsymbol{t}}, g_{\boldsymbol{t}}$ and $h_{\boldsymbol{t}}$, the first equation reduces to

$$\boldsymbol{\ell} \odot e^{\boldsymbol{t}} - \lambda \boldsymbol{u} + \mu e^{\boldsymbol{t}} = \mathbf{0},$$

which implies that for all $j \in [M]$

$$e^{t_j} = \frac{\lambda u_j}{\mu + \ell_j}. \tag{16}$$

Substituting this into the constraints $g = h = 0$ yields the following simultaneous equations in λ and μ

$$c = \mathrm{kl}(\boldsymbol{u}\|e^{\boldsymbol{t}}) = \sum_{j=1}^{M} u_j \log \frac{u_j}{e^{t_j}} = \sum_{j=1}^{M} u_j \log \frac{\mu+\ell_j}{\lambda} \quad \text{and} \quad \lambda \sum_{j=1}^{M} \frac{u_j}{\mu+\ell_j} = 1.$$

Substituting the second into the first and rearranging the second, this is equivalent to solving

$$c = \sum_{j=1}^{M} u_j \log\left((\mu+\ell_j)\sum_{k=1}^{M}\frac{u_k}{\mu+\ell_k}\right) \quad \text{and} \quad \lambda = \left(\sum_{j=1}^{M}\frac{u_j}{\mu+\ell_j}\right)^{-1}. \tag{17}$$

It has already been established in the discussion after Definition 10 that $f_{\boldsymbol{\ell}}(\boldsymbol{v})$ attains its maximum on the set $A_{\boldsymbol{u}} := \{\boldsymbol{v} \in \triangle_M : \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{v}) \le c\}$. Therefore $F(\boldsymbol{t})$ also attains its maximum on $\mathbb{R}^M$ and one of the solutions to these simultaneous equations corresponds to this maximum. We first show that there is a single solution to the first equation in the set $(-\infty, -\max_j \ell_j)$, referred to as $\mu^*(\tilde{\boldsymbol{u}})$ in the proposition. Second, we show that any other solution corresponds to a smaller total risk, so that $\mu^*(\tilde{\boldsymbol{u}})$ corresponds to the maximum total risk and yields $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) = \mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$ when $\mu^*(\tilde{\boldsymbol{u}})$ and the associated $\lambda^*(\tilde{\boldsymbol{u}})$ are substituted into Eq. (16).

For the first step, note that since the e^{t_j} are probabilities, we see from Eq. (16) that either $\mu + \ell_j > 0$ for all j (in the case that $\lambda > 0$), or $\mu + \ell_j < 0$ for all j (in the case that $\lambda < 0$). Thus any solutions μ to the first equation must be in $(-\infty, -\max_j \ell_j)$ or $(-\min_j \ell_j, \infty)$. If $\mu \in$

$(-\infty, -\max_j \ell_j)$ then the first equation can be written as $c = \phi_{\boldsymbol{\ell}}(\mu)$, with $\phi_{\boldsymbol{\ell}}$ as defined in the statement of the proposition. We now show that $\phi_{\boldsymbol{\ell}}$ is strictly increasing in μ, and that $\phi_{\boldsymbol{\ell}}(\mu) \to 0$ as $\mu \to -\infty$ and $\phi_{\boldsymbol{\ell}}(\mu) \to \infty$ as $\mu \to -\max_j \ell_j$, so that $c = \phi_{\boldsymbol{\ell}}(\mu)$ does indeed have a single solution in the set $(-\infty, -\max_j \ell_j)$. Straightforward differentiation and algebra shows that

$$\begin{aligned}\phi'_{\boldsymbol{\ell}}(\mu) &= \sum_{j=1}^M \frac{u_j}{(\mu+\ell_j)\sum_{k=1}^M \frac{u_k}{\mu+\ell_k}} \left(\sum_{k'=1}^M \frac{u_{k'}}{\mu+\ell_{k'}} - (\mu+\ell_j) \sum_{k'=1}^M \frac{u_{k'}}{(\mu+\ell_{k'})^2} \right) \\ &= \frac{\left(\sum_{j=1}^M \frac{u_j}{\mu+\ell_j}\right)^2 - \sum_{j=1}^M \frac{u_j}{(\mu+\ell_j)^2}}{\sum_{k=1}^M \frac{u_k}{\mu+\ell_k}}.\end{aligned}$$

Jensen's inequality demonstrates that the numerator is strictly negative, where strictness is due to the assumption that the ℓ_j are not all equal. Further, since the denominator is strictly negative (since we are dealing with the case where $\mu \in (-\infty, -\max_j \ell_j)$), we see that $\phi_{\boldsymbol{\ell}}$ is strictly increasing for $\mu \in (-\infty, -\max_j \ell_j)$.[3] Turning to the limits, we first show that $\phi_{\boldsymbol{\ell}}(\mu) \to \infty$ as $\mu \to -\max_j \ell_j$.

We now determine the left hand limit. Define $J = \{j \in [M] : \ell_j = \max_k \ell_k\}$, noting that this is a strict subset of $[M]$ since by assumption the ℓ_j are not all equal. We then have that for $\mu \in (-\infty, \max_j \ell_j)$

$$\begin{aligned}e^{\phi_{\boldsymbol{\ell}}(\mu)} &= \left(-\sum_{j=1}^M \frac{u_j}{\mu+\ell_j} \right) \left(\prod_{k=1}^M \big(-(\mu+\ell_k) \big)^{u_k} \right) \\ &= \left(-\sum_{j\in J} \frac{u_j}{\mu+\ell_j} - \sum_{j'\notin J} \frac{u_{j'}}{\mu+\ell_{j'}} \right) \prod_{k\in J} \big(-(\mu+\ell_k) \big)^{u_k} \prod_{k'\notin J} \big(-(\mu+\ell_{k'}) \big)^{u_{k'}} \\ &\geq \left(-\sum_{j\in J} \frac{u_j}{\mu+\ell_j} \right) \prod_{k\in J} \big(-(\mu+\ell_k) \big)^{u_k} \prod_{k'\notin J} \big(-(\mu+\ell_{k'}) \big)^{u_{k'}} \\ &= \frac{\left(\sum_{j\in J} u_j\right) \left(\prod_{k'\notin J} \big(-(\mu+\ell_{k'}) \big)^{u_{k'}} \right)}{\big(-(\mu+\max_j \ell_j) \big)^{1-\sum_{k\in J} u_k}}.\end{aligned}$$

The first term in the numerator is a positive constant, independent of μ. The second term in the numerator tends to a finite positive limit as $\mu \uparrow -\max_j \ell_j$. Since $[M] \setminus J$ is non-empty, the power in the denominator is positive and the term in the outer brackets is positive and tends to zero as $\mu \uparrow -\max_j \ell_j$. Thus $e^{\phi_{\boldsymbol{\ell}}(\mu)} \to \infty$ as $\mu \uparrow -\max_j \ell_j$ and, by the continuity of the logarithm, $\phi_{\boldsymbol{\ell}}(\mu)$ as $\mu \uparrow -\max_j \ell_j$.

We now determine $\lim_{\mu\to-\infty} \phi_{\boldsymbol{\ell}}(\mu)$ by sandwiching $\phi(\mu)$ between two functions that both tend to zero as $\mu \to -\infty$. First, since $\ell_j \geq 0$ for all j, for $\mu \in (-\infty, -\max_j \ell_j)$ we have

$$\log\left(-\sum_{j=1}^M \frac{u_j}{\mu+\ell_j} \right) \geq \log\left(-\sum_{j=1}^M \frac{u_j}{\mu} \right) = -\log(-\mu) = -\sum_{j=1}^M u_j \log(-\mu),$$

3. Incidentally, this argument also shows that there is at most one solution to the first equation in (17) in the range $(-\min_j \ell_j, \infty)$. There indeed exists a unique solution, which corresponds to the minimum total risk, but we do not prove this.

and so

$$\phi_{\boldsymbol{\ell}}(\mu) \geq -\sum_{j=1}^{M} u_j \log(-\mu) + \sum_{j=1}^{M} u_j \log\left(-(\mu+\ell_j)\right) = \sum_{j=1}^{M} u_j \log\left(1+\frac{\ell_j}{\mu}\right) \to 0 \quad \text{as} \quad \mu \to -\infty.$$

Similarly,

$$\sum_{j=1}^{M} u_j \log\left(-(\mu+\ell_j)\right) \leq \sum_{j=1}^{M} u_j \log(-\mu) = \log(-\mu),$$

and so

$$\phi_{\boldsymbol{\ell}}(\mu) \leq \log\left(\mu \sum_{j=1}^{M} \frac{u_j}{\mu+\ell_j}\right) = \log\left(\sum_{j=1}^{M} \frac{u_j}{1+\frac{\ell_j}{\mu}}\right) \to 0 \quad \text{as} \quad \mu \to -\infty.$$

This completes the first step, namely showing that there does indeed exist a unique solution $\mu^*(\tilde{\boldsymbol{u}})$ in the set $(-\ell_1, \infty)$ to the first equation in line (17).

We now turn to the second step, namely showing that this solution corresponds to the maximum total risk. Given a value of the Lagrange multiplier μ, substitution into Eq. (16) gives

$$e^{t_j}(\mu) = \frac{\frac{u_j}{\mu+\ell_j}}{\sum_{k=1}^{M} \frac{u_k}{\mu+\ell_k}}$$

and therefore total risk

$$R(\mu) = \frac{\sum_{j=1}^{M} \frac{u_j \ell_j}{\mu+\ell_j}}{\sum_{k=1}^{M} \frac{u_k}{\mu+\ell_k}}.$$

To prove that the solution $\mu^*(\tilde{\boldsymbol{u}}) \in (-\infty, -\max_j \ell_j)$ is the solution to the first equation in line (17) that maximises R, it suffices to show that $R(\mu) \to \sum_{j=1}^{M} u_j \ell_j$ as $|\mu| \to \infty$ and $R'(\mu) \geq 0$ for all $\mu \in (-\infty, -\max_j \ell_j) \cup (-\min_j \ell_j, \infty)$, so that

$$\inf_{\mu \in (-\infty, -\max_j \ell_j)} R(\mu) \geq \sup_{\mu \in (-\min_j \ell_j, \infty)} R(\mu).$$

This suffices as we have already proved that $\mu^*(\tilde{\boldsymbol{u}})$ is the only solution in $(-\infty, -\max_j \ell_j)$ to the first equation in line (17), and that no solutions exists in the set $[-\max_j \ell_j, -\min_j \ell_j]$.

The limit can be easily evaluated by first rewriting $R(\mu)$ and then taking the limit as $|\mu| \to \infty$ as follows

$$R(\mu) = \frac{\sum_{j=1}^{M} \frac{u_j \ell_j}{1+\frac{\ell_j}{\mu}}}{\sum_{k=1}^{M} \frac{u_k}{1+\frac{\ell_k}{\mu}}} \to \frac{\sum_{j=1}^{M} u_j \ell_j}{\sum_{k=1}^{M} u_k} = \sum_{j=1}^{M} u_j \ell_j.$$

To show that $R'(\mu) \geq 0$, let $\ell_{(j)}$ denote the j'th smallest component of $\boldsymbol{\ell}$ (breaking ties arbitrarily), so that $\ell_{(1)} \leq \cdots \leq \ell_{(M)}$, and use the quotient rule to see that

$$R'(\mu) \geq 0 \iff \frac{\left(\sum_{k=1}^{M} \frac{u_k}{\mu+\ell_k}\right)\left(\sum_{j=1}^{M} \frac{-u_j \ell_j}{(\mu+\ell_j)^2}\right) - \left(\sum_{j=1}^{M} \frac{u_j \ell_j}{\mu+\ell_j}\right)\left(\sum_{k=1}^{M} \frac{-u_k}{(\mu+\ell_k)^2}\right)}{\left(\sum_{p=1}^{M} \frac{u_p}{\mu+\ell_p}\right)^2} \geq 0$$

$$\iff \sum_{j=1}^{M}\sum_{k=1}^{M} \frac{u_j u_k \ell_j}{(\mu+\ell_j)(\mu+\ell_k)}\left(\frac{1}{\mu+\ell_k}-\frac{1}{\mu+\ell_j}\right) \geq 0$$

$$\iff \sum_{\substack{j,k\in[M]\\ k<j}} \frac{u_j u_k \ell_{(j)}}{(\mu+\ell_{(j)})(\mu+\ell_{(k)})}\left(\frac{1}{\mu+\ell_{(k)}}-\frac{1}{\mu+\ell_{(j)}}\right)$$

$$+ \sum_{\substack{j,k\in[M]\\ k>j}} \frac{u_j u_k \ell_{(j)}}{(\mu+\ell_{(j)})(\mu+\ell_{(k)})}\left(\frac{1}{\mu+\ell_{(k)}}-\frac{1}{\mu+\ell_{(j)}}\right) \geq 0,$$

where in the final line we have dropped the summands where $k = j$ since they equal zero as the terms in the bracket cancel. This final inequality holds since the first sum can be bounded below by the negative of the second sum as follows

$$\sum_{\substack{j,k\in[M]\\ k<j}} \frac{u_j u_k \ell_{(j)}}{(\mu+\ell_{(j)})(\mu+\ell_{(k)})}\left(\frac{1}{\mu+\ell_{(k)}}-\frac{1}{\mu+\ell_{(j)}}\right)$$

$$\geq \sum_{\substack{j,k\in[M]\\ k<j}} \frac{u_j u_k \ell_{(k)}}{(\mu+\ell_{(j)})(\mu+\ell_{(k)})}\left(\frac{1}{\mu+\ell_{(k)}}-\frac{1}{\mu+\ell_{(j)}}\right) \quad \text{(since } \ell_{(k)} \leq \ell_{(j)} \text{ for } k<j\text{)}$$

$$= \sum_{\substack{j,k\in[M]\\ k>j}} \frac{u_k u_j \ell_{(j)}}{(\mu+\ell_{(k)})(\mu+\ell_{(j)})}\left(\frac{1}{\mu+\ell_{(j)}}-\frac{1}{\mu+\ell_{(k)}}\right) \quad \text{(swapping dummy variables } j,k\text{)}.$$

We now turn to finding the partial derivatives of $F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))$ with respect the $\tilde{u}_j$, which in turn will allow us to find the partial derivatives of $\mathrm{kl}_{\boldsymbol{\ell}}^{-1}(\boldsymbol{u}|c)$. Let $\nabla_{\tilde{\boldsymbol{u}}}$ denote the gradient operator with respect to $\tilde{\boldsymbol{u}}$. Then the quantity we are after is $\nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) \in \mathbb{R}^{M+1}$, the j'th component of which is

$$\left(\nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))\right)_j = \sum_{k=1}^{M+1} \frac{\partial F}{\partial t_k}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))\frac{\partial t_k^*}{\partial \tilde{u}_j}(\tilde{\boldsymbol{u}}) = F_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) \cdot \frac{\partial \boldsymbol{t}^*}{\partial \tilde{u}_j}(\tilde{\boldsymbol{u}}) \in \mathbb{R}.$$

Thus the full gradient vector is

$$\nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) = F_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))\nabla_{\tilde{\boldsymbol{u}}} \boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tag{18}$$

where $\nabla_{\tilde{\boldsymbol{u}}} \boldsymbol{t}^*(\tilde{\boldsymbol{u}})$ is the $M \times (M+1)$ matrix given by

$$\left(\nabla_{\tilde{\boldsymbol{u}}} \boldsymbol{t}^*(\tilde{\boldsymbol{u}})\right)_{j,k} = \frac{\partial t_k^*}{\partial \tilde{u}_j}(\tilde{\boldsymbol{u}}).$$

Finding an expression for this matrix is difficult. Fortunately we can avoid needing to by using a trick from mathematical economics referred to as the envelope theorem, as we now show.

First, note that since, for all $\tilde{\boldsymbol{u}}$, the constraints $g = h = 0$ are satisfied by $\boldsymbol{t}^*(\tilde{\boldsymbol{u}})$, we have the identities

$$g(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) \equiv 0 \quad \text{and} \quad h(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) \equiv 0.$$

Differentiating these identities with respect to $\tilde{u}_j$ then yields

$$g_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) \cdot \frac{\partial \boldsymbol{t}^*}{\partial \tilde{u}_j}(\tilde{\boldsymbol{u}}) + g_{\tilde{u}_j}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) \equiv 0 \quad \text{and} \quad h_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) \cdot \frac{\partial \boldsymbol{t}^*}{\partial \tilde{u}_j}(\tilde{\boldsymbol{u}}) \equiv 0.$$

As before, we can write these $M+1$ pairs of equations as the following pair of matrix equations

$$g_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) \nabla_{\tilde{\boldsymbol{u}}} \boldsymbol{t}^*(\tilde{\boldsymbol{u}}) + g_{\tilde{\boldsymbol{u}}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) \equiv \mathbf{0} \quad \text{and} \quad h_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) \nabla_{\tilde{\boldsymbol{u}}} \boldsymbol{t}^*(\tilde{\boldsymbol{u}}) \equiv \mathbf{0}.$$

Multiplying these identities by $\lambda^*(\tilde{\boldsymbol{u}})$ and $\mu^*(\tilde{\boldsymbol{u}})$ respectively, and combining with equation (18), yields

$$\begin{aligned} \nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) &= \Big(F_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) + \lambda^*(\tilde{\boldsymbol{u}}) g_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) + \mu^*(\tilde{\boldsymbol{u}}) h_{\boldsymbol{t}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) \Big) \nabla_{\tilde{\boldsymbol{u}}} \boldsymbol{t}^*(\tilde{\boldsymbol{u}}) \\ &\quad + \lambda^*(\tilde{\boldsymbol{u}}) g_{\tilde{\boldsymbol{u}}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}) \\ &= \lambda^*(\tilde{\boldsymbol{u}}) g_{\tilde{\boldsymbol{u}}}(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{u}}), \end{aligned}$$

where the final equality comes from noting that the terms in the large bracket vanish due to equation (15). Recalling the expression for $g_{\tilde{\boldsymbol{u}}}(\boldsymbol{t}; \tilde{\boldsymbol{u}})$ given by Eq. (14) and that $\boldsymbol{v}^*(\tilde{\boldsymbol{u}}) = \exp(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))$ we obtain

$$\begin{aligned} \nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}})) &= \lambda^*(\tilde{\boldsymbol{u}}) \Big(1 - \boldsymbol{t}^*(\tilde{\boldsymbol{u}})_1 + \log u_1, \ldots, 1 - \boldsymbol{t}^*(\tilde{\boldsymbol{u}})_M + \log u_M, -1 \Big) \\ &= \lambda^*(\tilde{\boldsymbol{u}}) \left(1 + \log \frac{u_1}{\boldsymbol{v}^*(\tilde{\boldsymbol{u}})_1}, \ldots, 1 + \log \frac{u_M}{\boldsymbol{v}^*(\tilde{\boldsymbol{u}})_M}, -1 \right) \end{aligned}$$

Finally, recalling Equivalence (13), namely $\nabla_{\tilde{\boldsymbol{u}}} f^*_{\boldsymbol{\ell}}(\tilde{\boldsymbol{u}}) \equiv \nabla_{\tilde{\boldsymbol{u}}} F(\boldsymbol{t}^*(\tilde{\boldsymbol{u}}))$, we see that the above expression gives the derivatives $\frac{\partial f^*_{\boldsymbol{\ell}}}{\partial u_j}(\tilde{\boldsymbol{u}})$ and $\frac{\partial f^*_{\boldsymbol{\ell}}}{\partial c}(\tilde{\boldsymbol{u}})$ stated in the proposition, thus completing the proof. ■