Binary Forecast and Decision Rules via PAC Bayesian Model Aggregation

Daniel F. Pellatt and Yixiao Sun*
Department of Economics
University of California, San Diego

Abstract

We consider a PAC-Bayesian model aggregation approach to binary decision or forecast rules when different decision-outcome pairs may have asymmetric payoffs that can vary with observed covariates. The approach estimates a probability distribution over a class of models from which majority vote or stochastic decision rules can be derived. Adopting a utility-based measure of loss considered in Granger and Machina (2006), we show the PAC-Bayesian methodology is well suited to this setting. Non-asymptotic training sample bounds and oracle inequalities familiar in form to counterparts from the 0/1-loss literature are derived for the utility-based setting. The decision rules perform competitively in simulation experiments, achieving higher expected utility than several methods proposed in recent literature. The approach is also well suited to data-rich modeling environments; a constrained version of the learning algorithm produces utility-oriented decision rules with similarities to support vector machines.

JEL Classification:

Keywords:

1 Introduction

Forecasting an uncertain binary outcome arises in a variety of economic decision-making problems. Predicting whether or not a loan will be repaid or which direction an asset price will move are examples where a decision maker's action may vary in tandem with a binary forecast. In general, a decision maker may incur costs or benefits that vary depending on the prediction-outcome pair when making a decision such as to grant or decline a loan. Additionally, payoffs may vary with covariates observed prior to realizing the outcome of interest and these covariates may also influence the likelihood of the outcome. For example, as noted in Elliott and Lieli (2013), development finance institutions may view failing to grant a loan that would be repaid as being more costly when the entity is deemed beneficial to a vulnerable population. At the same time, observable characteristics that quantify this need could be correlated with whether or not a loan will be repaid.

*Email: dpellatt@ucsd.edu; yisun@ucsd.edu. Address correspondence to Department of Economics, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0508, USA

It is well known that there are many successful classification algorithms suitable to a variety of applications. However, asymmetric loss can be a crucial element to decision making and most popular classifiers are not designed around this feature. Maximizing a likelihood function or minimizing a zero-one loss function, or a convex surrogate, does not typically weigh the relative costs of false negatives and false positives according to the preferences of the decision maker. Recently, Elliott and Lieli (2013) proposed a maximum-utility approach. Given a class of parametric decision rules, $\{a(x,\theta) : \mathbb{R}^d \to \{-1,1\},\ \theta \in \Theta\}$, which map covariates $X \in \mathbb{R}^d$ to a binary decision or forecast, the parameters $\hat{\theta} \in \Theta$ are selected as those that maximize the empirical expected utility of the decision maker. Here the binary action or forecast a is associated with an uncertain outcome Y with aligned categories $\{-1,1\}$. The utility maximization framework will be the starting point for our analysis.

There is not a large econometric literature geared at this setting for data-rich environments. First, we point out some recent developments. Su (2020) notes that the trade-off between model class complexity and the propensity to over-fit carries through from empirical risk minimization to the utility maximization setting. He situates the maximum-utility problem in the structural risk minimization paradigm of Vapnik (1982). Building on the analysis of Bartlett et al. (2002), Koltchinskii (2001), and others, he considers a hierarchy of potential model classes with increasing complexity and derives distribution-free and data-driven penalties to select an appropriate model class and decision rule. Another approach was recently considered by Babii et al. (2020) who replace non-convex objects that arise in the utility-maximization problem with convex surrogates.

Here we approach utility-based decision rules from the PAC-Bayesian framework. For a collection (or collections) of decision models associated with a measurable parameter space, this will entail estimating a probability distribution over the model parameters. Then decisions are made by aggregating over all possible decision rules, placing the greatest weight on subsets of the parameter or model space associated with the lowest empirical loss. We adopt a utility-based measure of loss considered in Granger and Machina (2006). Several prior works consider binary classification from the PAC-Bayesian point of view including McAllester (1999), Langford and Shawe-Taylor (2003), McAllester (2003b), Catoni (2007), Germain et al. (2015), and others. We build in particular on the work of Catoni (2007), Germain et al. (2009), and Alquier et al. (2016). When the utility function is bounded, a lemma of Maurer (2004) enables several key steps of the analysis to proceed as one would in the 0/1 loss setting. This trick is also noted in Germain et al. (2015). In the non-bounded case, our setting turns out to be well suited to higher level assumptions like those in Alquier et al. (2016), where the PAC-Bayesian analysis allows for more general loss functions.

Although estimating a probability measure over a parameter space to form decision rules may seem unfamiliar, it is possible to view a variety of decision rules or classifiers in this light. For example, given covariates $X \in \mathbb{R}^d$, a set of transformations $\phi_j(X) : \mathbb{R}^d \to \mathbb{R}$ for $j = 1, \ldots, M$, and some estimated parameter vector $\hat{\theta} \in \Theta = \mathbb{R}^M$, consider predicting $Y \in \{-1,1\}$ with

$$\hat{Y} = \text{sign}\left[\sum_{j=1}^{M} \phi_j(X)\hat{\theta}_j\right].$$

The estimated parameter vector $\hat{\theta}$ could come from the method of support vector machine (SVM) or the MU procedure of Elliott and Lieli (2013). Both SVM and MU can result in predictions of the above form. Alternatively, consider the probability distribution $\hat{\rho}(\theta)$ over Θ given by the

multivariate normal $N(\hat{\theta}, I_M)$ distribution. In this case, it holds that

$$\text{sign}\left[\int_{\Theta} \text{sign}\left\{\sum_{j=1}^{M} \phi_j(X)\theta_j\right\} d\hat{\rho}(\theta)\right] = \text{sign}\left[\sum_{j=1}^{M} \phi_j(X)\hat{\theta}_j\right], \tag{1}$$

as can be seen in Section 3.2. The left-hand side above can be interpreted as taking a weighted majority vote over the class of models of the form

$$\text{sign}\left[\sum_{j=1}^{M} \phi_j(X)\theta_j\right], \quad \theta \in \Theta,$$

where $\hat{\rho}$ determines the weights that different regions of Θ receive. On the other hand, the right-hand side of (1) takes the same form as the SVM and MU rules. The PAC-Bayesian approach provides a tractable path to analyzing useful theoretical attributes of decision rules centered around data-dependent distributions $\hat{\rho}$, including those for the SVM and MU methods. This analysis guides the choice of distributions that we focus on in this paper. More broadly, while the PAC-Bayesian framework is useful for deriving competitive learning models (our focus here), this tractable path to analyzing potentially complicated models is a key point of interest itself in the machine learning literature. For example, Neyshabur et al. (2017) derive generalization bounds for deep neural networks in a PAC-Bayesian framework.

There are several attractive characteristics of the PAC-Bayesian approach to utility-oriented decision rules. It allows for a very flexible selection of the decision model class (or classes). Almost any classification model with real parameters can be accommodated. Rather than estimating these parameters by minimizing a 0/1 loss, convex surrogate, or likelihood function, a probability distribution over the parameters that is dependent on a measure of empirical utility is constructed. This puts the greatest weight on regions of the parameter space with high empirical utility and then one can aggregate over potential models in a way that favors these regions. Although the analysis is frequentist in nature, estimation tools from the Bayesian literature such as Markov Chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) can be applied, sidestepping potential difficulties with computational complexity. By utilizing variational or change-of-measure formulas, the approach allows us to derive training sample bounds that hold with high probability. Additionally, model aggregation can alleviate model misspecification and estimation noise; see, for example, Jiang and Tanner (2008) and Freund et al. (2004) where this is analyzed in different 0/1-loss-based classification settings. Both of these papers utilize exponentially weighted aggregators similar to that employed here. Lastly, as pointed out in Elliott and Lieli (2013), the utility-maximizing decision rule is not unique. It is not unusual in many settings to identify several models with identical or similar in-sample performances but with different out-of-sample performances. Model aggregation makes sense in the context of multiple solutions or multiple near-solutions.

The main contributions of this paper are as follows. We add to the toolbox available for estimating binary choice or forecast rules when the decision maker faces asymmetric payoffs that may depend on the value of observable covariates. The methodology is well suited to data-rich environments and the decision/forecast rules perform very competitively against existing alternatives, exhibiting noticeable gains in expected utility in the simulation environments also studied in Elliott and Lieli (2013) and Su (2020). We develop training sample bounds and oracle inequalities for the decision rules. These are similar in form to existing PAC-Bayesian bounds in

alternative settings such as the 0/1 loss which is nested by the utility-based loss adopted here. We show that the theoretical insights, training sample bounds, and modeling guidance of the PAC-Bayesian classification literature can be applied to the utility-oriented setting. Additionally, we illustrate how these concepts and decision rules can be adapted to accommodate the situation with multiple model classes of interest and provide practical guidance regarding implementation. Finally, we try to keep the presentation self-contained and expand on details of the approach. While the PAC-Bayesian methodology has not gained a lot of traction or exposure in the econometric literature, its flexibility and analytical tractability in a variety of machine learning problems suggest that it may prove useful in future econometric applications.

The paper is structured as follows. In Section 2 we introduce the decision model and PAC-Bayesian framework. In Section 3, we establish the theoretical properties of the PAC-Bayesian decision rules and derive a constrained version of the model, which is easier to implement and provides insight to the PAC-Bayesian machinery in this setting. In Section 4 we discuss implementation and estimation, and in Section 5 we carry out a simulation study. Section 6 concludes. Proofs are given in the appendix.

2 Forecasting Framework

2.1 Model

To frame the decision problem, we adopt the setting of Elliott and Lieli (2013), tying a binary action or decision to forecasting a binary outcome. This is the standard decision-theoretic framework analyzed in, for example, Granger and Machina (2006). In addition to the discussion below, we refer the reader to Granger and Machina (2006), Elliott and Lieli (2013), and the references therein for further theoretical considerations and additional applications of our setting to problems in economics and other sciences.

The decision maker's problem is to choose an action $a \in \{-1, 1\}$ given an observable vector of covariates $X \in \mathbb{R}^d$ with support $\mathcal{X} \subset \mathbb{R}^d$. The actions are defined in a broad sense and are categorically aligned with a binary outcome variable $Y \in \{-1, 1\}$ that is not observable at the time of decision making. Conditional on $X = x$, the outcome variable Y follows a Bernoulli distribution with parameter $P(x)$ where

$$P(x) = \Pr(Y = 1|X = x). \tag{2}$$

The payoff or utility function of the decision maker is $U(a, Y, X)$. $U : \{-1, 1\} \times \{-1, 1\} \times \mathcal{X} \to \mathbb{R}$ represents the preferences of the decision maker and is assumed known. We allow that the payoff $U(a, y, x)$ is a nontrivial function of x. The table below illustrates the payoff function under $X = x$ with different combinations of (a, Y).

	State	
Action	$Y = 1$	$Y = -1$
$a = 1$	$U(1, 1, x)$	$U(1, -1, x)$
$a = -1$	$U(-1, 1, x)$	$U(-1, -1, x)$

As a primary application of this setting, we may regard a as a forecast of the outcome of a future random variable Y, or alternatively as an action taken based on the predicted binary outcome of Y. Then $U(a, y, x)$ is the payoff when the forecast or action is a, the realized value

of Y is y, and the covariate vector is equal to x. In this application, we expect that

$$U(1,1,x) > U(1,-1,x) \text{ and } U(-1,-1,x) > U(-1,1,x) \text{ for all } x \in \mathcal{X}. \tag{3}$$

That is, a correct prediction delivers a higher payoff than an incorrect prediction.

As a second application, our setting can be cast as a 2×2 game where Nature plays Y and the decision maker plays a. More specifically, Nature plays a mixed strategy: for a given $X = x$, Nature plays $Y = 1$ with probability $P(x)$ and plays $Y = -1$ with probability $1 - P(x)$. In this case, (3) states that there is no dominating strategy for the decision maker.

We formalize (3) along with a self-explanatory technical condition as an assumption below.

Assumption 1 *(i) For all $x \in \mathcal{X}$, $U(1,1,x) - U(-1,1,x) > 0$ and $U(-1,-1,x) - U(1,-1,x) > 0$; (ii) for all $(a,y) \in \{-1,1\}^2$, $U(a,y,\cdot)$ is Borel measurable.*

2.2 Utility Maximizing Actions

Given $X = x$, a decision maker's action is optimal if it maximizes her conditional expected utility, i.e., a^* is optimal if

$$a^* \in \arg\max_a E\left[U(a,Y,X) \,|X = x\right]. \tag{4}$$

Here a^* depends on the observed value x. To signify this, we write it as $a^*(x)$. We can alternatively formulate the decision maker's problem in terms of a loss function. We think about the loss of an action a as the amount by which the resulting utility differs from that of a perfect forecast if Y were known when the decision is made. Given Assumption 1(i), a perfect forecast would entail setting the category of action a to that of Y; we denote this unobtainable action based on the realization of Y by a_R. To motivate the form of the loss function, note that (4) is equivalent to

$$a^* \in \arg\min_a E\left[U(a_R,Y,X) - U(a,Y,X) | X = x\right]. \tag{5}$$

With $a_R = Y$ by Assumption 1(i), we define the loss function $\ell : \{-1,1\}^2 \times \mathcal{X} \to \mathbb{R}$ by

$$\ell(a,y,x) = U(y,y,x) - U(a,y,x). \tag{6}$$

This utility-induced loss function is called the *point-forecast/point-realization loss function* in Granger and Machina (2006). Clearly,

$$\ell(a,y,x) = \begin{cases} 0, & \text{if } a = y \\ U(y,y,x) - U(a,y,x) > 0, & \text{if } a \neq y. \end{cases}$$

In general, $\ell(a,y,x) \neq \ell(y,a,x)$, and so the loss function is not symmetric. In terms of the loss function, we have

$$a^* \in \arg\min_a E\left[\ell(a,Y,X) | X = x\right]. \tag{7}$$

We can now derive a solution of (7) (equation (9) below), which is also obtained in Elliott and Lieli (2013). When $X = x$ and $a = 1$, the expected loss is

$$E[\ell(1,Y,X)|X = x] = (1 - P(x))\ell(1,-1,x).$$

When $X = x$ and $a = -1$, the expected loss is

$$E[\ell(-1,Y,X)|X = x] = P(x)\ell(-1,1,x).$$

Now, if we let

$$b(x) = \ell(1,-1,x) + \ell(-1,1,x),$$
$$c(x) = \frac{\ell(1,-1,x)}{b(x)} = \frac{\ell(1,-1,x)}{\ell(1,-1,x)+\ell(-1,1,x)}, \tag{8}$$

then a little algebra shows that an optimal decision rule, i.e., the one that obtains the lowest possible expected loss, is to set $a^*(x) = 1$ if and only if $P(x) > c(x)$. This can be written as

$$a^*(x) = \text{sign}\left[P(x) - c(x)\right], \tag{9}$$

where $\text{sign}(z) = 1$ for $z > 0$ and $\text{sign}(z) = -1$ for $z \leq 0$.

For intuition, under Assumption 1 and provided that $P(x) < 1$, $a^*(x)$ in (9) can be restated as setting $a = 1$ if and only if

$$\frac{P(x)}{1-P(x)} > \frac{\ell(1,-1,x)}{\ell(-1,1,x)} = \frac{U\left(-1,-1,x\right) - U\left(1,-1,x\right)}{U\left(1,1,x\right) - U\left(-1,1,x\right)}.$$

If we think of a as a prediction of Y based on $X = x$, then $\ell(1,-1,x) = U(-1,-1,x) - U(1,-1,x)$ is the *ex post* missed utility from a false positive prediction (i.e., take $a = 1$ when $Y = -1$) and $\ell(-1,1,x) = U(1,1,x) - U(-1,1,x)$ is the *ex post* missed utility from a false negative prediction (i.e., take $a = -1$ when $Y = 1$). The optimal decision rule sets $a = 1$ only when the odds ratio of the event $Y = 1$ relative to the event $Y = 0$ is greater than the false positive to false negative loss ratio. As the relative cost of a false positive gets larger, a greater odds ratio is required for an optimal utility-based decision rule to permit the action $a = 1$.

In terms of $b\left(x\right)$ and $c\left(x\right)$, the point-realization loss function in (6) can be written as

$$\ell(a,y,x) = \psi(x,y) \cdot 1\left\{y \neq a\right\}, \tag{10}$$

where

$$\psi(x,y) = b(x)\left[\frac{y+1}{2} - yc(x)\right] = U(y,y,x) - U(-y,y,x) > 0. \tag{11}$$

This can be easily verified. Therefore,

$$a^* \in \arg\min_a E\left[\psi(X,Y) 1\left\{Y \neq a\right\} | X = x\right]. \tag{12}$$

The decision maker knows the payoff function $U\left(a,y,x\right)$ and hence $b\left(x\right), c\left(x\right)$, and $\psi(x,y)$. She does not know $P(x)$, the only piece of information that is still missing in solving the above minimization problem. To make an optimal decision, she has to estimate $P\left(x\right)$ based on the sample $\{(X_i,Y_i)\}_{i=1}^n$. One of her options would be to choose a proxy $m(x)$ for the unknown $P(x)$ from some class of functions. Her task is then to learn the most suitable m for a decision rule of the form $a(x) = \text{sign}\left[m(x) - c(x)\right]$. In considering such options, we will maintain the following additional sampling and distributional assumptions.

Assumption 2 *(i)* $\{(X_i,Y_i)\}_{i=1}^n$ *is an iid sample; (ii)* $X_i \in \mathcal{X}$ *and* $Y_i \in \{-1,1\}$*; (iii) The joint distribution function of* (X,Y) *is* $P(X,Y)$ *where* $P(X,Y)$ *is a probability measure over* $(\mathcal{X} \times \{-1,1\}, \mathcal{B}_x \otimes \mathcal{B}_y)$ *where* $\mathcal{B}_x$ *is the Borel* σ*-algebra associated with* $\mathcal{X}$ *and* $\mathcal{B}_y$ *consists of all subsets of* $\{-1,1\}$*; (iv) There exists some* $K_\psi > 0$ *such that*

$$E\exp\left\{\lambda^2 \psi(X,Y)^2\right\} \leq \exp\left\{K_\psi^2 \lambda^2\right\} \text{ for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_\psi}.$$

The condition on the moment generating function in Assumption 2(iv) specifies that the random variable $\psi(X,Y)$ is sub-Gaussian (c.f. Proposition 2.5.2 (iii) and Definition 2.5.6 of Vershynin (2018)). Given that $\psi(x,y) = U(y,y,x) - U(-y,y,x)$, this assumption requires that the payoffs from a correct decision (or alternatively, the costs from an incorrect decision) must be sub-Gaussian. This is a fairly mild requirement and accommodates, for example, any underlying utility function that is bounded, a condition that is assumed in Elliott and Lieli (2013) and Su (2020). Here benefits and costs of correct or incorrect decisions do not have to be bounded provided that the tails of the distribution decay exponentially.

In terms of the resulting action rule $a_{m^*}(x) = \text{sign}\left[m^*(x) - c(x)\right]$, the conditional optimization problem (12) is equivalent to the unconditional optimization problem

$$m^* \in \arg\min_{m\in\mathcal{M}} E\left\{\psi\left(X,Y\right) 1\left\{Y \neq \text{sign}\left[m(X) - c(X)\right]\right\}\right\}, \tag{13}$$

where $\mathcal{M}$ is the space of all measurable functions from $\mathcal{X}$ to $\mathbb{R}$. To implement the optimal m^*, the decision maker could solve the sample version of the above problem,

$$\hat{m}^* \in \arg\min_{m\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n} \psi\left(X_i,Y_i\right) 1\left\{Y_i \neq \text{sign}\left[m(X_i) - c(X_i)\right]\right\},$$

and let

$$a_{\hat{m}^*}\left(x\right) = \text{sign}\left[\hat{m}^*(x) - c(x)\right].$$

The M estimator $\hat{m}^*$ is motivated from utility maximization, and we will refer to it as the maximum utility (MU) estimator. The MU estimator is clearly different from the maximum likelihood estimator defined as

$$\hat{m}^*_{MLE} = \arg\max_{m\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i+1}{2}\log m\left(X_i\right) + \left(1 - \frac{Y_i+1}{2}\right)\log\left[1 - m(X_i)\right]\right\},$$

where we have assumed that $m(X_i) \in (0,1)$.[1] The likelihood function is motivated statistically without accounting for the payoff differences under different actions and states of the world.

Implementation of the optimal strategy requires searching over the whole space of measurable functions $\mathcal{M}$. This is a formidable task. In addition, such a method may not generalize well. In practice, we restrict attention to a parameterized subclass of $\mathcal{M}$. Denote this collection of models by $\mathcal{M}_\Theta \subset \mathcal{M}$ where each model $m(x,\theta) \in \mathcal{M}_\Theta$ is determined by parameters $\theta \in \Theta$ and $\Theta \subset \mathbb{R}^q$ is the parameter space with potentially $q \neq d$, where d is the dimension of $\mathcal{X}$. We delay specifying the functional form of $m(x,\theta)$ for now. The MU estimator over $\mathcal{M}_\Theta$ selects the model parameter $\hat{\theta}$ by solving

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \psi\left(X_i,Y_i\right) 1\left\{Y_i \neq \text{sign}[m(X_i,\theta) - c(X_i)]\right\}.$$

Such an estimator has been considered in Elliott and Lieli (2013). In the special case that the loss functions $\ell(1,-1,x)$ and $\ell(-1,1,x)$ are equal to the same constant function, we have $c\left(x\right) = 1/2$ and $\psi(x,y) = \left[\ell(1,-1,x) + \ell(-1,1,x)\right]/2$, which is also a constant function. Hence,

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} 1\left\{Y_i \neq \text{sign}[m(X_i,\theta) - c(X_i)]\right\}.$$

[1] If this is not the case, we can take a transform such as the logistic transform so that the transformed version is in $(0,1)$.

In this case, the MU estimator reduces to the maximum score estimator of Manski (1975, 1985). Su (2020) considers model selection in the MU framework. There, model selection is based on a penalized MU estimator where the additive penalty regularizes the complexity of the model class and controls the generalization error.

A key observation from Elliott and Lieli (2013) is that m^* and $\hat{m}^*$ may not be unique. Consider the sample problem as an example. If $\hat{m}^*$ is a solution, then any function $\hat{m}(x)$ that satisfies

$$\text{sign}\left[\hat{m}^*(x) - c(x)\right] = \text{sign}\left[\hat{m}(x) - c(x)\right]$$

is also a solution. Each crossing point of $P(x)$ and $c(x)$ corresponds to a region of $\mathcal{X}$ where $\hat{m}^*$ and $\hat{m}$ may disagree out of sample even if both achieve the same in-sample empirical utility. This provides an incentive to consider ensemble methods. In the presence of multiple solutions, it is reasonable to average or aggregate models with high empirical utility rather than trying to select a solution.

2.3 PAC-Bayesian Framework

Instead of model selection, we consider model aggregation in this paper. We do so within the PAC-Bayesian framework. In this subsection, we introduce some definitions and concepts central to this approach before considering statistical properties of the resulting decision rules in Section 3.

Most generally, we work with $\mathcal{R}_\Theta$, a parameterized subclass of the set of measurable functions from $\mathcal{X}$ to $\{-1, 1\}$, characterized by a parameter space Θ. The typical example we deal with here and in our simulations is the setting where

$$\mathcal{R}_\Theta = \left\{\text{sign}\left[m(x,\theta) - c(x)\right] : m \in \mathcal{M}_\Theta\right\}, \tag{14}$$

where again $\mathcal{M}_\Theta$ is a parameterized subclass of the space of measurable functions $m : \mathcal{X} \to \mathbb{R}$ that are characterized by the parameter space $\Theta \subset \mathbb{R}^q$ associated with q model parameters.

For actions $a(x,\theta) \in \mathcal{R}_\Theta$ (determined by $\theta \in \Theta$), with some abuse of notation, we denote the utility-induced, point-realization loss by

$$\ell(\theta, y, x) = \psi(x,y)1\left\{y \neq a(x,\theta)\right\},$$

where $\psi(x,y)$ is defined in (11). Additionally, for any $\theta \in \Theta$, define the risk function $R(\theta)$ and its empirical counterpart $R_n(\theta)$ by

$$R(\theta) = E[\ell(\theta, Y, X)], \tag{15}$$

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Y_i, X_i). \tag{16}$$

Whereas the MU approach selects a single $\hat{\theta} \in \Theta$ by minimizing $R_n(\theta)$ over Θ, here we will place a non-negative weighting on each θ in the form of a probability measure on Θ and then take actions based on aggregation over all possible models. The goal is to construct a probability measure $\rho(\cdot)$ on Θ that may depend on the sample $\{(X_i, Y_i)\}_{i=1}^n$. The PAC-Bayesian framework allows us to identify bounds on functionals of $R(\theta)$ that depend on $R_n(\theta)$ and hold with high probability. These bounds can then guide the choice of ρ. We will need to integrate over both the sample space and the parameter space, and we make the following assumption.

Assumption 3 *(i)* $(\Theta, \mathcal{B}_\theta)$ *is a measurable space where* $\mathcal{B}_\theta$ *is the standard* σ*-algebra on* Θ *and is countably generated; (ii)* $(\theta, x) \mapsto a(x, \theta) : (\Theta \times \mathcal{X}, \mathcal{B}_\theta \otimes \mathcal{B}_x) \to (\{-1, 1\}, \mathcal{B}_a)$ *is a measurable function where* $\mathcal{B}_a = \mathcal{B}_y$.

Assumption 3 contains some technical conditions that address measurability concerns. By a probability measure $\rho(\cdot)$ on Θ that may be sample dependent, we mean a regular conditional probability measure $\rho(z, \cdot)$ where $z \in (\mathcal{X} \times \{-1, 1\})^{\times n}$. That is, for any fixed $S \in \mathcal{B}_\theta$, $\rho(z, S) : ((\mathcal{X} \times \{-1, 1\})^{\times n}, (\mathcal{B}_x \otimes \mathcal{B}_y)^{\otimes n}) \to \mathbb{R}_+$ is measurable in z and for any fixed z, the map $S \mapsto \rho(z, S) : \mathcal{B}_\theta \to \mathbb{R}_+$ is a probability measure. For conciseness, we suppress the potential reliance of ρ on the particular sample set z. Given some deterministic probability measure π, we will work with the Kullback-Leibler (KL) divergence between π and ρ,

$$D_{\mathrm{KL}}(\rho, \pi) = \begin{cases} \int_\Theta \log\left[\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta), & \text{if } \rho \ll \pi \\ \infty, & \text{else.} \end{cases}$$

We will consider only the case that $\rho \ll \pi$ (a.s.) in this paper. The requirement in Assumption 3 that $\mathcal{B}_\theta$ is countably generated serves to ensure that objects such as $D_{\mathrm{KL}}(\rho, \pi)$ are measurable. For further measure-theoretic consideration, we refer the reader to Catoni (2004), in particular Proposition 1.7.1 and its proof on pages 50-54. There the measurability of $D_{\mathrm{KL}}(\rho, \pi)$ when ρ and π may be regular conditional probability measures is demonstrated under conditions that are met by our assumptions.

Given a probability measure $\rho(\cdot)$ over Θ, there are a few ways to form a decision rule. Among them, the Gibbs method and the majority vote method are widely used. The Gibbs method associated with ρ draws a value, say $\theta_\circ$, randomly according to ρ and then takes the action based on $\theta_\circ$. Mathematically, we let $\theta_\circ \sim \rho$ and we take

$$a_{G,\rho}(x) = a_{\theta_\circ}(x).$$

That is, we play a mixed strategy based on the distribution ρ. With some abuse of the notation, the average risk of the Gibbs method associated with ρ is

$$R(a_{G,\rho}) = \int_\Theta R(\theta) d\rho(\theta) = E_{\theta\sim\rho} E_{X,Y\sim P(X,Y)} \psi(X, Y) 1\{Y \neq a(X, \theta)\}, \tag{17}$$

which is referred to as the Gibbs risk in the literature. Above, $E_{\theta\sim\rho}$ is the expectation with respect to the distribution of θ, and $E_{X,Y\sim P(X,Y)}$ is the expectation with respect to the distribution of (X, Y). We adopt the same convention hereafter.

The Gibbs risk $R(a_{G,\rho})$ is the expectation of the risk function $R(\theta)$ under measure $\rho(\cdot)$. It is thus a linear functional of $\rho(\cdot)$. More precisely, if $\rho = \alpha\rho_1 + (1 - \alpha)\rho_2$ for some ρ_1 and ρ_2 and a constant α, then $R(a_{G,\rho}) = \alpha R(a_{G,\rho_1}) + (1 - \alpha) R(a_{G,\rho_2})$. The linearity makes the Gibbs risk more amenable to theoretical analysis.

For the majority vote method, which is also called the Bayes method, the action is defined according to

$$a_{B,\rho}(x) = \operatorname{sign}\{E_{\theta\sim\rho} a(x, \theta)\}.$$

Such a method aggregates the actions $\{a(x, \theta) : \theta \in \Theta\}$ to obtain the prevailing action. For intuition, suppose that $\theta_1, \ldots, \theta_N$ are N i.i.d. draws from ρ and consider the action

$$a_{B,N}(x) = \operatorname{sign}\left\{\frac{1}{N}\sum_{j=1}^N a(x, \theta_j)\right\}.$$

Provided that $E_{\theta\sim\rho}a(x,\theta) \neq 0$, we have $a_{B,N}(x) \overset{\text{a.s.}}{\to} a_{B,\rho}(x)$ as $N \to \infty$ for each x. Note that $a_{B,N}(x) = 1$ if and only if more than half of the actions $\{a(x,\theta_j)\}_{j=1}^N$ are equal to 1 so this is akin to a weighted majority vote of the parameter values in Θ. The risk of the majority vote (also called the Bayes risk) associated with ρ is defined by

$$\begin{aligned} R(a_{B,\rho}) &= E_{X,Y\sim P(X,Y)}\psi(X,Y)1\{Y \neq a_{B,\rho}(X)\} \\ &= E_{X,Y\sim P(X,Y)}\psi(X,Y)1\{Y \neq \text{sign}\{E_{\theta\sim\rho}a(X,\theta)\}\}. \end{aligned}$$

The Bayes risk is clearly not linear in ρ.

In practice, the majority vote method or the Bayes method delivers numerically more stable results than the Gibbs method, but the latter is easier to analyze. However, the Bayes risk is upper bounded by twice the Gibbs risk as shown in the following lemma.

Lemma 1 *Let Assumption 1, Assumptions 2(ii) and (iii), and Assumption 3 hold. Then, for any probability measure ρ on Θ,*

$$R(a_{B,\rho}) \leq 2R(a_{G,\rho}).$$

Lemma 1 extends the "factor 2" bound for the majority vote method in the machine learning literature to the utility-based, point-realization loss setting. This property is well documented in the case of 0/1 loss (e.g., Langford and Shawe-Taylor (2003), McAllester (2003a), and Germain et al. (2015)). Here, we use Lemma 1 only to justify using the Gibbs risk as a surrogate for the majority vote risk. The loose bound in the lemma is enough for this purpose. Langford and Shawe-Taylor (2003) show that the factor of 2 can sometimes be reduced to $(1+\epsilon)$ for some small $\epsilon > 0$. Lacasse et al. (2006) and Germain et al. (2015) show that tighter bounds on $R(a_{B,\rho})$ can be obtained in the 0/1 loss setting and in a related loss variant.

To choose ρ to guide our decisions, we follow the PAC-Bayesian approach. Let $\mathcal{P}(\Theta)$ be the set of probability measures on $(\Theta, \mathcal{B}_\theta)$. The first ingredient is a reference or prior probability measure π. We make the following assumption:

Assumption 4 *$\pi \in \mathcal{P}(\Theta)$ is a (deterministic) probability measure that does not depend on the sample.*

We will denote the set of probability measures on $(\Theta, \mathcal{B}_\theta)$ that are absolutely continuous with respect to π by $\mathcal{P}_\pi(\Theta)$. Assumption 4 is essential in PAC-Bayesian analysis. For example, our analysis will involve the sample version of the Gibbs risk, defined for $\rho \in \mathcal{P}(\Theta)$ by

$$R_n(a_{G,\rho}) = \int_\Theta R_n(\theta)d\rho = \sum_{i=1}^n \int_\Theta \ell(\theta, Y_i, X_i)d\rho(\theta). \tag{18}$$

If ρ is derived from $\{X_i, Y_i\}_{i=1}^n$, (18) is difficult to work with because $\int_\Theta \ell(\theta, Y_i, X_i)d\rho(\theta)$ is not iid and so $R_n(a_{G,\rho})$ is not a sum of iid terms. However, for any measurable function $A(\theta)$, the so-called change-of-measure inequality states that for *any* $\rho \in \mathcal{P}_\pi(\Theta)$,

$$\int_\Theta A(\theta)d\rho(\theta) \leq \log\left[\int_\Theta \exp(A(\theta))\, d\pi(\theta)\right] + D_{\text{KL}}(\rho, \pi), \tag{19}$$

provided the integrals are well defined. When both $\rho(\cdot)$ and $A(\cdot)$ depend on the sample and exhibit complicated dependence, it may not be easy to control $\int_\Theta A(\theta)d\rho(\theta)$. But when π does

not depend on the sample, (19) can provide a manageable upper bound. Although the change of measure inequality is simple and easy to prove, it is foundational to the PAC-Bayesian approach. See McAllester (2003b) and references therein for further discussion. (19) is stated below as Corollary 3(b), and a proof is given in the appendix. Some choices for π are discussed in Sections 3.2 and 4.

Given the pre-specified π, we choose ρ to minimize the sample Gibbs risk $R_n(a_{G,\rho})$ in (18), subject to the constraint that ρ is not too different from π. We utilize the KL divergence to measure the difference between two probability measures. Mathematically, we solve the constrained minimization problem:

$$\min_{\rho\in\mathcal{P}_\pi(\Theta)} \int_\Theta R_n(\theta)d\rho(\theta) \ \text{ s.t. } D_{\mathrm{KL}}(\rho,\pi)\le C,$$

for some constant C. Alternatively, we use the Lagrangian form and solve the unconstrained minimization problem

$$\min_{\rho\in\mathcal{P}_\pi(\Theta)} \left[\int_\Theta R_n(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right], \tag{20}$$

where $\lambda > 0$ is a constant. Theoretical justification for this choice of optimization problem is given in Section 3.

Let $\mathcal{M}(\Theta)$ be the set of measurable functions on $(\Theta,\mathcal{B}_\theta)$ and

$$\mathcal{M}_b^\pi(\Theta) = \left\{A : A\in\mathcal{M}(\Theta) \text{ and } \int_\Theta \exp(A(\theta))\,d\pi(\theta) < \infty\right\},$$

which is a subset of $\mathcal{M}(\Theta)$ that has a finite exponential moment under π. To obtain a closed-form solution to 20, we provide the following lemma and corollary, which will also be used repeatedly for establishing other results.

Lemma 2 *For $\pi\in\mathcal{P}(\Theta)$ and $A\in\mathcal{M}(\Theta)$ such that $-A\in\mathcal{M}_b^\pi(\Theta)$, let $\rho_{A,\pi}\in\mathcal{P}_\pi(\Theta)$ be the probability measure on Θ with the Radon–Nikodym (RN) derivative with respect to π given by*

$$\frac{d\rho_{A,\pi}(\theta)}{d\pi(\theta)} = \frac{\exp(-A(\theta))}{\int_\Theta \exp\left(-A\left(\tilde\theta\right)\right)d\pi\left(\tilde\theta\right)}.$$

Then for any probability measure $\rho\in\mathcal{P}_\pi(\Theta)$ we have

$$\log\left[\int_\Theta \exp(-A(\theta))\,d\pi(\theta)\right] = -\left[\int_\Theta A(\theta)\,d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] + D_{\mathrm{KL}}(\rho,\rho_{A,\pi}). \tag{21}$$

Corollary 3 *(a) For A, π, ρ, and $\rho_{A,\pi}$ as in Lemma 2, we have*

$$\rho_{A,\pi} = \arg\min_{\rho\in\mathcal{P}_\pi(\Theta)}\left[\int_\Theta A(\theta)\,d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] \tag{22}$$

and

$$\min_{\rho\in\mathcal{P}_\pi(\Theta)}\left[\int_\Theta A(\theta)\,d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] = -\log\left[\int_\Theta \exp(-A(\theta))\,d\pi(\theta)\right].$$

(b) For any $\mathcal{A}(\cdot)\in\mathcal{M}_b^\pi(\Theta)$, $\pi\in\mathcal{P}(\Theta)$, $\rho\in\mathcal{P}_\pi(\Theta)$,

$$\int_\Theta \mathcal{A}(\theta)\,d\rho(\theta) \le \log\left[\int_\Theta \exp(\mathcal{A}(\theta))\,d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho,\pi).$$

Lemma 2 and Corollary 3(a) provide a closed-form solution to the minimization problem in (20). Let

$$\hat{\rho}_\lambda := \arg \min_{\rho \in \mathcal{P}_\pi(\Theta)} \left[\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right]. \tag{23}$$

Then it follows from Corollary 3(a) that $\hat{\rho}_\lambda = \rho_{\lambda R_n, \pi}$. Given λ and π, $\hat{\rho}_\lambda$ will be our primary choice of probability measure for deriving decision rules through a majority vote or Gibbs method. We present this as a definition.

Definition 4 *$\hat{\rho}_\lambda$ is a (random) probability measure on Θ with the following RN derivative with respect to π :*

$$\frac{d\hat{\rho}_\lambda}{d\pi}(\theta) = \frac{\exp[-\lambda R_n(\theta)]}{\int_\Theta \exp\left[-\lambda R_n\left(\tilde{\theta}\right)\right] d\pi\left(\tilde{\theta}\right)}.$$

$\hat{\rho}_\lambda$ is sometimes called the Gibbs posterior. From a Bayesian perspective, we may regard π as the prior distribution for the parameter $\theta \in \Theta$ and $\hat{\rho}_\lambda$ as the posterior distribution. Such a Bayesian interpretation may help us understand the approach, but it is not necessary. In fact, this interpretation is valid only if $\exp[-\lambda R_n(\theta)]$ is proportional to a likelihood function. The approach we use is a frequentist one, and $\exp[-\lambda R_n(\theta)]$ does not have to be a likelihood function. The definition of $\hat{\rho}_\lambda$ is motivated from the minimization problem in (23), not from any Bayesian principle. In particular, there does not have to be a likelihood function or a complete model. All we need is the empirical risk based on the utility-based loss function. Also, π does not have to be a prior distribution. It can be any distribution that does not depend on the sample. However, for easy references, we may still refer to π as the prior and $\hat{\rho}_\lambda$ as the posterior. More generally, any ρ determined from the sample may be referred to as a posterior distribution.

The probability measure $\hat{\rho}_\lambda$ can be regarded as an adjusted version of π. Consider two parameters $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$. If $R_n(\theta_1) < R_n(\theta_2)$, then $\exp[-\lambda R_n(\theta_1)] > \exp[-\lambda R_n(\theta_2)]$ for any $\lambda > 0$. Hence, relative to π, $\hat{\rho}_\lambda$ assigns more weights to θ_1 than to θ_2. The distributional adjustment, therefore, favors the parameter value that delivers a smaller in-sample empirical risk. The degree of adjustment is determined by the tuning parameter λ. On the one hand, if λ approaches zero, then $\hat{\rho}_\lambda$ approaches π, and there will be no adjustment. On the other hand, if $\lambda \to +\infty$, then $\hat{\rho}_\lambda$ assigns all weights to the minimizers of $R_n(\theta)$, provided that the minimizers are in the support of the prior π. We will investigate the choice of λ in subsequent sections.

3 PAC-Bayesian Analysis Under Utility-Based Loss

In this section, we derive PAC-Bayesian bounds on the Gibbs risk and oracle inequalities for decision rules based on $\hat{\rho}_\lambda$ in Definition 4 for the utility-induced loss setting. The bounds provide justification for focusing on the minimization problem in (20). They are non-asymptotic training set bounds that hold for a user-specified confidence level. The oracle inequalities illustrate a sense in which $\hat{\rho}_\lambda$ is close to the probability measure we would select if $R(\theta)$ were known. We also consider a constrained version of the problem in (20) which illustrates the mechanics of the methodology and produces decision rules with similarities to support vector machines. Lastly, we consider the formulation when one is interested in aggregating multiple decision model classes.

For a probability measure ρ on Θ that may depend on the sample, an integral step in PAC-Bayesian analysis is to establish an upper bound for $D[R(a_{G,\rho}), R_n(a_{G,\rho})]$ where $D : \mathbb{R}^2_+ \to \mathbb{R}$ is a measure of the difference between the Gibbs risk $R(a_{G,\rho})$ defined in (17) and its empirical

counterpart $R_n(a_{G,\rho})$ defined in (18). We will often focus on the case $D(r_1, r_2) = r_1 - r_2$, i.e., when

$$D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] = \int_\Theta R\left(\theta\right) d\rho(\theta) - \int_\Theta R_n\left(\theta\right) d\rho(\theta).$$

Let $\epsilon > 0$ be a small constant. The initial aim is to establish the following result: for some upper bound $B_n(\pi, \rho, \epsilon)$ we have

$$\Pr\left\{D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] \leq B_n(\pi, \rho, \epsilon) \text{ for all } \rho \in \mathcal{P}_\pi(\Theta) \text{ simultaneously}\right\} \geq 1 - \epsilon. \tag{24}$$

We can use such a bound to choose $\hat{\rho} \in \mathcal{P}_\pi(\Theta)$ so that the Gibbs risk of $\hat{\rho}$, $R(a_{G,\hat{\rho}})$, is minimized with high probability. We will see that this leads to the minimization problem in (20).

For a given $D\left(\cdot, \cdot\right)$, we can regard $D[R(a_{G,\hat{\rho}}), R_n(a_{G,\hat{\rho}})]$ as a measure of the generalization error under the Gibbs method for $\hat{\rho}$. If $B_n(\pi, \hat{\rho}, \epsilon)$ decays to zero for any $\epsilon > 0$ as n increases, then the above inequality implies a low generalization error with high probability (i.e., with probability at least $1 - \epsilon$ for any small ϵ). In this case, we say that $R_n(a_{G,\hat{\rho}}) = \int_\Theta R_n(\theta) d\hat{\rho}(\theta)$ is probably (the high probability part) and approximately correct (the low generalization error part) for $R(a_{G,\hat{\rho}}) = \int_\Theta R(\theta) d\hat{\rho}(\theta)$. The PAC framework, introduced by Valiant (1984), evaluates learning mechanisms via the probability (prescribing a confidence level) that the resulting rule will approximate an optimal rule at some level of accuracy. As noted in Shalev-Shwartz and Ben-David (2014), which includes an excellent introduction to PAC analysis, this framework has broad appeal, has been extended in scope (e.g. Haussler (1992)), and has been utilized in several foundational analyses (e.g. Vapnik (1982), Vapnik (1992), and Vapnik (2013)). In the PAC-Bayesian framework, rather than centering attention on learning mechanisms that settle on a particular instance in the parameter space, the focus rests on PAC statements for objects concerning distributions over models or model parameters. The approach then has flavors of both Probably Approximately Correct (PAC) learning and Bayesian learning. Hence it can be called PAC-Bayesian learning. As we discussed previously, the Bayesian part is a misnomer, and we use "PAC-Bayesian" in the absence of a better term.

3.1 Bounds and Oracle Inequalities for the Decision Rule

Here we establish PAC-Bayesian and oracle bounds under Assumptions 1 – 4. We begin with the following bound of the form in (24).

Theorem 5 *Let Assumptions 1, 2, and 3 hold. Let $D : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ be convex over the range of $(\psi(x, y), \psi(x, y))$ where ψ is defined in (11) and depends on the utility function $U(a, y, x)$. Assume there exists a function $f(\lambda, n)$ and an interval $I \subseteq \mathbb{R}_+^* = \{\lambda \in \mathbb{R} : \lambda > 0\}$ such that for all $\lambda \in I$,*

$$\int_\Theta E \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi(\theta) \leq \exp(f(\lambda, n)). \tag{25}$$

Then for any $\epsilon \in (0, 1]$,

$$\Pr\left\{D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] \leq \frac{f\left(\lambda, n\right) + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}\left(\rho, \pi\right)}{\lambda} \text{ for all } \rho \in \mathcal{P}_\pi(\Theta) \text{ simultaneously}\right\}$$
$$\geq 1 - \epsilon. \tag{26}$$

There is a fairly well established path to results like Theorem5 in the literature. For example, Bégin et al. (2016) lays out a blueprint for deriving such bounds in the 0/1 loss setting that is general enough to encompass many results identified in the previous literature. The above bound combines elements of Theorem 4.2 in Alquier et al. (2016) and Theorem 18 in Germain et al. (2015). Theorem 5 is proved in the Appendix. Alquier et al. (2016) refer to condition (25) as the Hoeffding assumption. In situations where $D[R(\theta), R_n(\theta)]$ may become unbounded almost surely for certain values of θ, such a condition can allow for valid and nontrivial PAC-Bayesian bounds provided that $\pi(\theta)$ is chosen judiciously. We will also note that D in Theorem 7 may depend on λ provided that for each $\lambda \in I$ it is convex over the range of $(\psi(x,y), \psi(x,y))$. In our analysis, when D depends on λ in this way, it will be the case that the resulting high probability inequality simplifies so that the left-hand-side contains an object of interest and does not depend on λ.

To produce the main bounds and oracle inequalities of interest, we combine the above theorem with the following lemma.

Lemma 6 *Let Assumptions 1 – 4 hold.*

(a) For $s \in \{-1, 1\}$, let $D(r_1, r_2) = s(r_1 - r_2)$, so that

$$D\left[R\left(\theta\right), R_n\left(\theta\right)\right] = s\left(R\left(\theta\right) - R_n\left(\theta\right)\right).$$

Then for $\lambda > 0$, (25) holds with

$$f(\lambda, n) = \frac{\lambda^2 \left[K_\psi^2 + \mu_\psi^2\right]}{n},$$

where K_ψ is the constant in Assumption 2 and $\mu_\psi = E\psi(X, Y)$. Additionally, if

$$U_{\max} = \sup_{a,y,x} |U(a, y, x)| < \infty, \tag{27}$$

then for $\lambda > 0$, (25) holds with

$$f(\lambda, n) = \frac{\lambda^2 U_{\max}^2}{2n}.$$

(b) Assume (27) holds. Let

$$D(r_1, r_2) = \mathcal{F}(r_1) - r_2,$$

where

$$\mathcal{F}(r) := \mathcal{F}_{n,\lambda}\left(r\right) = -\frac{n}{\lambda} \log\left\{1 - \frac{r}{2U_{\max}}\left[1 - \exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right]\right\}. \tag{28}$$

Then, for $\lambda > 0$, (25) holds with

$$f(\lambda, n) = 0.$$

(c) Assume (27) holds. Let

$$D(r_1, r_2) = \max\left\{r_1 - \frac{\lambda U_{\max}^2}{2n} - r_2,\ \mathcal{F}(r_1) - r_2\right\},$$

where $\mathcal{F}$ is defined as in (28). Then, for $\lambda > 0$, (25) holds with

$$f(\lambda, n) = 0.$$

Theorem 5 combined with Lemma 6 produces the following result.

Theorem 7 *Under Assumptions 1 – 4, for $\lambda > 0$ and $\epsilon \in (0, 1]$ we have the following properties.*
(a) For $s \in \{-1, 1\}$, the following event occurs with probability at least $1 - \epsilon$ for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously:

$$\int_\Theta s\left[R\left(\theta\right) - R_n\left(\theta\right)\right] d\rho(\theta) \leq \frac{1}{\lambda}\left[\frac{\lambda^2}{n}\left(K_\psi^2 + \mu_\psi^2\right) + D_{\mathrm{KL}}\left(\rho, \pi\right) + \log\frac{1}{\epsilon}\right]$$

where K_ψ is the constant in Assumption 2 and $\mu_\psi = E\psi(X, Y)$. If (27) holds, then the term $(K_\psi^2 + \mu_\psi^2)$ can be replaced by $U_{\max}^2/2$.
(b) If (27) holds, then the following event occurs with probability at least $1 - \epsilon$ for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously:

$$\int_\Theta R\left(\theta\right) d\rho(\theta) \leq \mathcal{F}_{n,\lambda}^{-1}\left(\int_\Theta R_n\left(\theta\right) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \frac{1}{\lambda}\log\frac{1}{\epsilon}\right).$$

where $\mathcal{F}_{n,\lambda}^{-1}\left(r\right)$ is the inverse function of $\mathcal{F}_{n,\lambda}\left(r\right)$:

$$\mathcal{F}_{n,\lambda}^{-1}\left(r\right) = 2U_{\max}\frac{1 - \exp\left(-\frac{\lambda}{n}\cdot r\right)}{1 - \exp\left(-\frac{\lambda}{n}\cdot 2U_{\max}\right)}.$$

(c) Define

$$U_{\lambda,\pi,\rho}(\epsilon) = \int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n} + D_{\mathrm{KL}}\left(\rho, \pi\right) + \log\frac{1}{\epsilon}\right], \tag{29}$$

$$U_{\lambda,\pi,\rho}^{\mathcal{F}}(\epsilon) = \mathcal{F}_{n,\lambda}^{-1}\left(\int_\Theta R_n\left(\theta\right) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \frac{1}{\lambda}\log\frac{1}{\epsilon}\right). \tag{30}$$

If (27) holds, the following event occurs with probability at least $1 - \epsilon$ for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously:

$$\int_\Theta R\left(\theta\right) d\rho(\theta) \leq \min\left\{U_{\lambda,\pi,\rho}(\epsilon), U_{\lambda,\pi,\rho}^{\mathcal{F}}(\epsilon)\right\}.$$

When $U_{\max} < \infty$, the bounds in Theorem 7(a), (b), and (c) can be computed from the sample for a learned ρ, be it of the form in Definition 4 or that in Section 3.2 or some other form. In the $U_{\max} < \infty$ setting, part (c) can provide an improvement over the bounds in part (a) and (b) which are, respectively, similar in form to bounds in Alquier et al. (2016) and Catoni (2007). Setting $s = 1$ in Theorem 7(a), we obtain, with probability at least $1 - \epsilon$ for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously:

$$\int_\Theta R\left(\theta\right) d\rho(\theta) \leq \left[\int_\Theta R_n\left(\theta\right) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\rho, \pi\right)\right] + \frac{1}{\lambda}\left[\frac{\lambda^2}{n}\left(K_\psi^2 + \mu_\psi^2\right) + \log\frac{1}{\epsilon}.\right]$$

The above bound and the bound in 7(b) are slight variants of one another. Note that for a given λ, if we choose ρ to minimize the upper bound for $R(a_{G,\rho}) = \int_\Theta R(\theta) d\rho(\theta)$ in either of the inequalities we are led back to the minimization problem in (20). The bound in Theorem 7(b) is similar in form to Theorem 1.2.6 in Catoni (2007) for the 0/1 loss. It is recovered from the distance measure D in Lemma 6(b) similarly to Germain et al. (2009) who focus on the 0/1 loss setting.

When $s = 1$, Theorem 7(a) gives us

$$\Pr\left\{\int_\Theta \left[R\left(\theta\right) - R_n\left(\theta\right)\right] d\rho(\theta) \leq B_n(\pi, \rho, \epsilon) \text{ for all } \rho \in \mathcal{P}_\pi(\Theta) \text{ simultaneously}\right\} \geq 1 - \epsilon,$$

for

$$B_{n,\lambda}(\pi, \rho, \epsilon) = \frac{\lambda}{n}\left(K_\psi^2 + \mu_\psi^2\right) + \frac{1}{\lambda}\left[\log\frac{1}{\epsilon} + D_{\mathrm{KL}}\left(\rho, \pi\right)\right].$$

Setting λ proportional to $n^{1/2}$ yields the following best rate of the PAC bound $B_{n,\lambda}(\pi, \rho, \epsilon)$:

$$B_{n,\lambda}(\pi, \rho, \epsilon) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

On the other hand, for the function $\mathcal{F}_{n,\lambda}^{-1}\left(\cdot\right)$ in Theorem 7(b), we have, using $\exp(x) \geq 1 + x$ for all $x \in \mathbb{R}$,

$$\mathcal{F}_{n,\lambda}^{-1}\left(r\right) = 2U_{\max}\frac{1 - \exp\left(-\frac{\lambda}{n}\cdot r\right)}{1 - \exp\left(-\frac{\lambda}{n}\cdot 2U_{\max}\right)} \leq \frac{C_n}{1 - \exp\left(-C_n\right)}r$$

where $C_n = \frac{\lambda}{n}\cdot 2U_{\max}$. Hence, Theorem 7(b) implies that

$$\Pr\left\{\int_\Theta \left[R\left(\theta\right) - R_n\left(\theta\right)\right] d\rho(\theta) \leq B_{n,C}(\pi, \rho, \epsilon) \text{ for all } \rho \in \mathcal{P}_\pi(\Theta) \text{ simultaneously}\right\} \geq 1 - \epsilon,$$

where

$$\begin{aligned} B_{n,C_n}(\pi, \rho, \epsilon) = &\left[\frac{C_n}{1 - \exp\left(-C_n\right)} - 1\right]\int_\Theta R_n\left(\theta\right) d\rho(\theta) \\ &+ \frac{2U_{\max}}{n}\frac{1}{1 - \exp\left(-C_n\right)}\left[\log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\rho, \pi)\right]. \end{aligned}$$

When $R_n\left(\theta\right) > 0$, setting C_n proportional to $\left(n\int_\Theta R_n\left(\theta\right) d\rho(\theta)\right)^{-1/2}$ yields the following best rate of the PAC bound $B_{n,C_n}(\pi, \rho, \epsilon)$:

$$B_{n,C_n}(\pi, \rho, \epsilon) = O_p\left(\sqrt{\frac{\int_\Theta R_n\left(\theta\right) d\rho(\theta)}{n}}\right).$$

It should be noted, however, that we cannot choose λ in C_n according to the data for the bounds in Theorem 7. We consider valid bounds when λ is data-dependent, for example when it is chosen via cross-validation in Theorems 9 and 11.

When $P(X, Y)$ and ρ are such that $R_n(a_{G,\rho}) = \int_\Theta R_n\left(\theta\right) d\rho(\theta)$ is very small, the PAC bound from Theorem 7(b) can be smaller than that in Theorem 7(a). For a given λ, Theorem 7(c) says that we can take the better of the two, without applying any union bound arguments that require a reduction in ϵ. On the other hand, Theorem 7(b) and (c) only provide upper bounds for $\int_\Theta R\left(\theta\right) d\rho(\theta)$ while 7(a) provides both an upper bound and a lower bound.

Note that Theorem 7 holds for all ρ simultaneously. Setting $\rho(\cdot)$ equal to $\hat{\rho}_\lambda(\cdot)$ in Theorem 7(a) and (c), we can obtain the following theorem.

Theorem 8 *Let Assumptions 1 – 4 hold. Then for $\epsilon \in (0,1]$ each of the following holds with probability at least $1-\epsilon$:*
(a)

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda \leq \int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda + \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{1}{\epsilon}\right],$$

(b)

$$\left|\int_\Theta R(\theta)\, d\hat{\rho}_\lambda - \int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda\right| \leq \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right]$$

(c)

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda \leq \min_{\rho \in \mathcal{P}_\pi(\Theta)}\left[\int_\Theta R(\theta) d\rho(\theta) + \frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi)\right] + \frac{2}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right].$$

If (27) holds, $(K_\psi^2 + \mu_\psi^2)$ can be replaced by $U_{\max}^2/2$ in (a)-(c).

(d) When (27) holds,

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \min\left\{U_{\lambda,\pi,\hat{\rho}_\lambda}(\epsilon), U^{\mathcal{F}}_{\lambda,\pi,\hat{\rho}_\lambda}(\epsilon)\right\},$$

where $U_{\lambda,\pi,\hat{\rho}_\lambda}(\epsilon)$ and $U^{\mathcal{F}}_{\lambda,\pi,\hat{\rho}_\lambda}(\epsilon)$ are given by (29) and (30) with ρ set to $\hat{\rho}_\lambda$.

Theorem 8(a) provides a PAC-Bayesian bound for the generalization error of the Gibbs method. When $U_{\max} < \infty$, choosing the rate-optimal $\lambda = \kappa\sqrt{n}$ for some constant $\kappa > 0$ gives us

$$\Pr\left\{\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda(\theta) + \frac{1}{\kappa\sqrt{n}}\left[D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) + \log\frac{1}{\epsilon}\right] + \frac{\kappa U_{\max}^2}{2\sqrt{n}}\right\} \geq 1-\epsilon. \quad (31)$$

Therefore, the PAC generalization error decays to zero at the rate of $1/\sqrt{n}$.

Theorem 8(b) allows us to construct a $(1-\epsilon)$ confidence interval $CI_{\lambda,\pi}(\epsilon)$ for $\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta)$:

$$CI_{\lambda,\pi}(\epsilon) = \left[L_{\lambda,\pi}(\epsilon), U_{\lambda,\pi}(\epsilon)\right],$$

where

$$L_{\lambda,\pi}(\epsilon) = \int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda - \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) - \frac{1}{\lambda}\left(\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right),$$

$$U_{\lambda,\pi}(\epsilon) = \int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda + \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) + \frac{1}{\lambda}\left(\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right).$$

Let

$$U^{\mathcal{F}}_{\lambda,\pi}(\epsilon) = \mathcal{F}^{-1}_{n,\lambda}\left(\int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda + \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) + \frac{1}{\lambda}\log\frac{2}{\epsilon}\right).$$

Then the upper limit of $CI_{\lambda,\pi}(\epsilon)$ can be replaced by $\min(U_{\lambda,\pi}(\epsilon), U^{\mathcal{F}}_{\lambda,\pi}(\epsilon))$, leading to a shorter confidence interval. This follows from a union bound argument as in the proof of Theorem 8(a). Note that $U^{\mathcal{F}}_{\lambda,\pi}(\epsilon)$ above is equal to $U^{\mathcal{F}}_{\lambda,\pi,\hat{\rho}_\lambda}(\epsilon/2)$ in equation (30). If there is a natural bound for $\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta)$, such as 0 for the lower bound or $2U_{\max}$ for the upper bound, we should make an obvious modification to the above interval.

Theorem 8(c) shows that the estimated probability measure $\hat{\rho}_\lambda(\cdot)$ strikes almost the best trade-off between the average risk $\int_\Theta R(\theta)\, d\rho(\theta)$ and the regularization term $\frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi)$. The best trade-off that solves the minimization problem is given by the distribution $\rho_{\lambda R/2}$ with the following RN derivative

$$\frac{d\rho_{\lambda R/2}}{d\pi}(\theta) = \frac{\exp\left[-\frac{\lambda}{2} R(\theta)\right]}{\int_\Theta \exp\left[-\frac{\lambda}{2} R(\theta)\right] d\pi(\theta)}.$$

This follows from Corollary 3(a). Note that $R(\theta)$ is not feasible and is only known to an oracle. Hence, $\rho_{\lambda R/2}$ is not feasible and the bound in the theorem is an oracle-type risk bound.

Theorem 8(c) can be interpreted as selecting the best probability measure in $\mathcal{P}_\pi(\Theta)$. Ideally, we select $\rho(\theta) \in \mathcal{P}_\pi(\Theta)$ to minimize the average risk $\int_\Theta R(\theta)\, d\rho(\theta)$. An oracle who knows $R(\theta)$ can solve for the best $\rho^*(\theta)$, namely, $\rho^* = \arg\min_{\rho \in \mathcal{P}_\pi(\Theta)} \int_\Theta R(\theta)\, d\rho(\theta)$. Not knowing $R(\theta)$, we replace it by the empirical estimator $R_n(\theta)$ and add a regularization term to the objective function. That is, we solve the optimization problem in (23). The selected $\hat{\rho}_\lambda$ can not be expected to be as good as ρ^*. However, Theorem 8(c) shows that it is almost as good as a second best oracle solution $\rho_{\lambda R/2}$.

In practice, λ will be chosen by cross validation. However, cross validating λ inhibits the use of Theorems 7 and 8 for deriving risk bounds or confidence intervals. We mention two methods for dealing with this. First, we can employ an idea from Catoni (2007) for deriving bounds that do not rely on λ. This entails combining a union-bound argument with Theorem 7 and leads to the following theorem.

Theorem 9 *Let Assumptions 1 – 4 hold and let $\alpha > 1$ and $\epsilon \in (0, 1]$. Assume (27) holds. Each event below holds with probability at least $1 - \epsilon$.*

(a) For $s \in \{-1, 1\}$ and for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously,

$$\int_\Theta s\left[R(\theta) - R_n(\theta)\right] d\rho(\theta) \leq \inf_{\lambda > 1} \left\{ \frac{\alpha}{\lambda} \left[\frac{\lambda^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\rho, \pi) + 2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha} \right] \right\}$$

(b) For $s \in \{-1, 1\}$ and any $\tilde{\lambda} > 1$ which may be chosen based on the sample,

$$\int_\Theta s\left[R(\theta) - R_n(\theta)\right] d\hat{\rho}_{\tilde{\lambda}}(\theta) \leq \frac{\alpha}{\tilde{\lambda}} \left[\frac{\tilde{\lambda}^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}}, \pi) + 2\log\frac{\log\left(\alpha^2\tilde{\lambda}\right)}{\log\alpha} \right]$$

(c) Let

$$\mathcal{F}^{-1}_{n,\lambda,\alpha}(r) = 2U_{\max} \frac{1 - \exp\left(-\frac{\lambda}{n} \cdot r\right)}{1 - \exp\left(-\frac{\lambda}{\alpha n} \cdot 2U_{\max}\right)},$$

and define

$$\overline{U}_{\lambda,\pi,\rho,\alpha}(\epsilon) = \int_\Theta R_n(\theta) d\rho(\theta) + \frac{\alpha}{\lambda} \left[\frac{\lambda^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\rho, \pi) + 2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha} \right],$$

$$\overline{U}^{\mathcal{F}}_{\lambda,\pi,\rho,\alpha}(\epsilon) = \mathcal{F}^{-1}_{n,\lambda,\alpha}\left(\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\left[\log\frac{1}{\epsilon} + 2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right]\right).$$

For all $\rho \in \mathcal{P}_\pi(\Theta)$ *simultaneously,*

$$\int_\Theta R(\theta) d\rho(\theta) \leq \inf_{\lambda>1}\left\{\min\left[\overline{U}_{\lambda,\pi,\rho,\alpha}(\epsilon), \overline{U}^{\mathcal{F}}_{\lambda,\pi,\rho,\alpha}(\epsilon)\right]\right\}.$$

(d) For any $\tilde{\lambda} > 1$ *that may be chosen based on the sample,*

$$\int_\Theta R(\theta) d\hat{\rho}_{\tilde{\lambda}}(\theta) \leq \min\left[\overline{U}_{\tilde{\lambda},\pi,\hat{\rho}_{\tilde{\lambda}},\alpha}(\epsilon), \overline{U}^{\mathcal{F}}_{\tilde{\lambda},\pi,\hat{\rho}_{\tilde{\lambda}},\alpha}(\epsilon)\right].$$

Theorem 9 is stated for the case where $U_{\max} < \infty$, i.e., a setting where the bounds can be computed without knowledge of the DGP. However, the bounds in parts (a) and (b) have valid counterparts in the more general case where we would replace $U^2_{\max}/2$ by $K^2_\psi + \mu^2_\psi$. Following similar arguments to those producing the confidence interval $CI_{\lambda,\pi}(\epsilon)$ after Theorem 8, a confidence interval for $\int_\Theta R(\theta) d\hat{\rho}_{\tilde{\lambda}}$ can be derived from Theorem (9) that is valid when $\hat{\rho}_{\tilde{\lambda}}$ is such that $\tilde{\lambda}$ is data-dependent. Note that in parts (a) and (c) the infimum is taken over all $\lambda > 1$. The condition that $\lambda > 1$ is fairly reasonable in relation to the bounds that motivate the decision rules. To see this, in the bounded utility setting, suppose that $U : \{-1,1\}^2 \times \mathcal{X} \to [-U_{\max}, U_{\max}]$ is replaced with the normalized utility $\tilde{U} = U/(2U_{\max})$, which of course does not alter the underlying preferences. Then $\tilde{U}_{\max} = \sup_{a,y,x}|\tilde{U}(a,y,x)| = 1/2$, so that the point-forecast loss based on this utility function satisfies $0 \leq \tilde{\ell}(\theta,y,x) \leq 1$. With this normalization, any observed loss is then a percentage of the largest possible loss rather than relying on potentially arbitrary utils. With this normalization, for any $0 < \lambda \leq 1$, both the bounds in parts (a) and (b) of Theorem 7 are such that the right-hand side is trivial (i.e., it is at least 1) whenever $\epsilon < \exp(-1)$. Focusing on $\lambda > 1$ restricts attention to values for which confidence in the bounds is more reasonable.

A second method for obtaining bounds or confidence intervals when $\tilde{\lambda}$ is data-dependent is to build from bounds in the literature where the PAC-Bayesian analysis does not utilize this temperature parameter. For example, the following result is also obtained in Maurer (2004) and Germain et al. (2015). While these authors do not explicitly consider loss functions that vary with X, some results there carry through when the utility function is bounded.

Lemma 10 *Let Assumptions 1 – 4 hold and assume that* (27) *holds. Let*

$$D(r_1,r_2) = \frac{n}{\lambda}\left[\mathrm{kl}\left(\frac{r_2}{2U_{\max}}, \frac{r_1}{2U_{\max}}\right)\right], \text{ where } \mathrm{kl}(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}.$$

Then, for $\lambda > 0$*, condition* (25) *in Theorem 5 holds with*

$$f(\lambda,n) = \log\xi(n), \text{ where } \xi(n) := \sum_{k=1}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}.$$

That $\mathrm{kl}(\cdot,\cdot)$ is convex follows from Theorem 2.7.2 of Cover and Thomas (2006) and we adopt the convention that $0\log 0 = 0$, $a\log\frac{a}{0} = \infty$ if $a > 0$ and $0\log\frac{0}{0} = 0$. Note that $\mathrm{kl}(a,b)$ is the KL-divergence between two Bernoulli random variables with success probabilities a and b respectively. It can be shown (c.f. Lemma 19 in Germain et al. (2015) and references therein) that $\sqrt{n} \leq \xi(n) \leq 2\sqrt{n}$. Theorem 5 combined with Lemma 10 produce the first part of the following theorem. The second part follows from an application of Pinsker's inequality, $2(a-b)^2 \leq \mathrm{kl}(a,b)$.

Theorem 11 *Let Assumptions 1 – 4 hold and assume that* (27) *holds. For* $\epsilon > 0$*, each of the following holds with probability at least* $1 - \epsilon$*.*

(a) for all $\rho \in \mathcal{P}_\pi(\Theta)$ *simultaneously,*

$$\mathrm{kl}\left(\frac{R_n(a_{G,\rho})}{2U_{\max}}, \frac{R\left(a_{G,\rho}\right)}{2U_{\max}}\right) \leq \frac{1}{n}\left[\log \xi(n) + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\rho, \pi)\right].$$

(b) for all $\rho \in \mathcal{P}_\pi(\Theta)$ *simultaneously,*

$$\left|\int_\Theta R(\theta)d\rho(\theta) - \int_\Theta R_n(\theta)d\rho(\theta)\right| \leq 2U_{\max}\sqrt{\frac{1}{2n}\left(\log \xi(n) + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\rho, \pi)\right)}.$$

As discussed in Germain et al. (2015), (a) is a slight improvement over similar bounds that have arisen in earlier PAC-Bayesian literature. One option to derive a bound for $\int_\Theta R_n(\theta)d\hat{\rho}_{\tilde{\lambda}}(\theta)$ is to solve the inequality in (a) numerically. As the bounds in Theorem 11 do not depend on λ and are valid for any $\rho \in \mathcal{P}_\pi(\Theta)$, they produces bounds for $\hat{\rho}_{\tilde{\lambda}}$ when $\tilde{\lambda}$ is data dependent.

Lastly, the generalization bounds for the loss function can be used to obtain generalization bounds for the utility function directly. To this end, denote

$$\begin{aligned}\mathbb{U}_n(\theta) &= \frac{1}{n}\sum_{i=1}^n [U(a(X,\theta), Y, X)] = \frac{1}{n}\sum_{i=1}^n U(Y_i, Y_i, X_i) - R_n\left(\theta\right)\\ \text{and } \mathbb{U}(\theta) &= E[U(a(X,\theta), Y, X)] = EU(Y, Y, X) - R\left(\theta\right).\end{aligned}$$

Also let

$$B_U = U_{\max}\sqrt{\frac{2\log\frac{2}{\epsilon}}{n}}.$$

Then we have the following corollary of earlier bounds for $\int_\Theta [R(\theta) - R_n(\theta)]d\hat{\rho}(\theta)$.

Corollary 12 *For* $\epsilon > 0$*, let* $\hat{\rho}$ *be a probability distribution over* Θ *and let* $B_R(\hat{\rho})$ *be a high probability (at least* $1 - \epsilon/2$*) bound for* $\int_\Theta [R(\theta) - R_n(\theta)]d\hat{\rho}(\theta)$*, i.e.* $B_R(\hat{\rho})$ *satisfies*

$$\Pr\left\{\int_\Theta \left[R(\theta) - R_n(\theta)\right] d\hat{\rho}(\theta) \leq B_R\left(\hat{\rho}\right)\right\} \geq 1 - \frac{\epsilon}{2}.$$

Then

$$\Pr\left(\int_\Theta \mathbb{U}(\theta)d\hat{\rho}\left(\theta\right) \geq \int_\Theta \mathbb{U}_n(\theta)d\hat{\rho}\left(\theta\right) - \left(B_U + B_R\left(\hat{\rho}\right)\right)\right) \geq 1 - \epsilon.$$

For example, if we are considering decision rules using $\hat{\rho}_{\tilde{\lambda}}$ with data dependent $\tilde{\lambda} > 1$, under the assumptions of Theorem 9(a) and for $\alpha > 1$ we can take

$$B_R(\hat{\rho}_{\tilde{\lambda}}) = \frac{\alpha}{\tilde{\lambda}}\left[\frac{\tilde{\lambda}^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}}, \pi) + 2\log\frac{\log\left(\alpha^2\tilde{\lambda}\right)}{\log\alpha}\right].$$

3.2 Linear Decision Rules in the Utility Setting

By definition, the estimator $\hat{\rho}_\lambda$ solves

$$\hat{\rho}_\lambda := \underset{\rho \in \mathcal{P}_\pi(\Theta)}{\arg\min} \left[\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right].$$

The distribution is not standard and must be approximated by numerical methods such as MCMC or tempered SMC (the latter is discussed in Section 4). Here, we consider a restrictive class of posteriors from a parametric family. In particular, we consider the case that both ρ and π are normal. Specifically, we assume that under π

$$\theta = (\theta_1, \theta_2, ..., \theta_q)' \sim N(\mu_\pi, \Sigma_\pi),$$

and under ρ

$$\theta = (\theta_1, \theta_2, ..., \theta_q)' \sim N(\mu_\rho, \Sigma_\rho),$$

where μ_π and μ_ρ are the mean vectors and Σ_π and Σ_ρ are the covariance matrices.

Lemma 13 *The KL divergence between $\rho : N(\mu_\rho, \Sigma_\rho)$ and $\pi : N(\mu_\pi, \Sigma_\pi)$ on $\mathbb{R}^q$ is*

$$D_{\mathrm{KL}}(\rho, \pi) = \frac{1}{2}(\mu_\rho - \mu_\pi)' \Sigma_\pi^{-1} (\mu_\rho - \mu_\pi) + \frac{1}{2}\left[\mathrm{tr}\left(\Sigma_\rho \Sigma_\pi^{-1}\right) - q\right] - \frac{1}{2}\log\frac{\det(\Sigma_\rho)}{\det(\Sigma_\pi)}.$$

We further assume that $\mathcal{R}_\Theta$ is described by (14) and that for $x \in \mathcal{X}$,

$$m(x, \theta) = \sum_{j=1}^{q} \phi_j(x)\theta_j = \phi(x)'\theta,\ \theta \in \mathbb{R}^q \tag{32}$$

for some set of feature transformations $\{\phi_1(x), \ldots, \phi_q(x)\}$ where $\phi_j(x) : \mathcal{X} \to \mathbb{R}$. For example, $\{\phi_1(x), \ldots, \phi_q(x)\}$ can consist of transforms of the observable variables using any set of basis functions. Another case of interest would be the setting where $\mathcal{R}_\Theta$ is specified by

$$\mathcal{R}_\Theta = \left\{a(x, \theta) = \mathrm{sign}\left(\phi(x)'\theta\right) : \theta \in \mathbb{R}^q\right\} \tag{33}$$

This is analogous to the setting of Germain et al. (2009) in the 0/1 loss setting. For example, one could take $\{\phi_1(x), \ldots, \phi_q(x)\}$ to be a set of decision stumps, with a fixed number of stumps and predetermined thresholds for each component of $x \in \mathbb{R}^d$. We focus on (32) below, but the results are easily adjusted to the setting of (33), simply drop the term $c(x)$.

Before proceeding, we note that the majority vote or Bayes method in this setting takes a particularly convenient form. For any fixed X, note that under $\theta \sim N(\mu_\rho, \Sigma_\rho)$ we have

$$\phi(X)'\theta - c(X) \sim N\left(\phi(X)'\mu_\rho - c(X), \phi(X)'\Sigma_\rho\phi(X)\right),$$

and therefore it follows that

$$E_{\theta\sim\rho}\mathrm{sign}\left[\phi(X)'\theta - c(X)\right] = 2\Phi\left(\frac{\phi(X)'\mu_\rho - c(X)}{\sqrt{\phi(X)'\Sigma_\rho\phi(X)}}\right) - 1.$$

Hence the majority vote takes the form

$$\begin{aligned} a_{B,\rho}(X) &= \text{sign}\left\{E_{\theta\sim\rho}\text{sign}\left[\phi(X)'\theta - c(X)\right]\right\} \\ &= \text{sign}\left\{2\Phi\left(\frac{\phi(X)'\mu_\rho - c(X)}{\sqrt{\phi(X)'\Sigma_\rho\phi(X)}}\right) - 1\right\} = \text{sign}\left[\phi(X)'\mu_\rho - c(X)\right]. \end{aligned}$$

That is, the decision rule in this case is straightforward to calculate and depends directly on a linear combination of a set of mappings from $\mathcal{X}$ to $\mathbb{R}$. Additionally, we will utilize the following lemma.

Lemma 14 *Under the normal prior and posterior setting described above,*

$$\int_\Theta R_n(\theta)\, d\rho(\theta) = \frac{1}{n}\sum_{i=1}^n \psi(X_i, Y_i)\Phi\left(-\frac{V(X_i, Y_i, \mu_\rho)}{\sqrt{\phi(X_i)'\Sigma_\rho\phi(X_i)}}\right).$$

where

$$V(X_i, Y_i, \mu_\rho) = Y_i\left[\phi(X_i)'\mu_\rho - c(X_i)\right].$$

Given Lemma 14, the minimization problem then reduces to the following problem:

$$\left(\hat{\mu}_\rho, \hat{\Sigma}_\rho\right) := \arg\min_{\mu_\rho, \Sigma_\rho}\left\{\frac{1}{n}\sum_{i=1}^n \psi(X_i, Y_i)\,\Phi\left(-\frac{V(X_i, Y_i, \mu_\rho)}{\sqrt{\phi(X_i)'\Sigma_\rho\phi(X_i)}}\right) + \frac{1}{\lambda}D_{\text{KL}}(\rho, \pi)\right\}.$$

When $\mu_\pi = 0$, $\Sigma_\pi = diag\left(\sigma^2_{\pi,j}\right)$, and $\Sigma_\rho = diag\left(\sigma^2_{\rho,j}\right)$, we have

$$\begin{aligned} D_{\text{KL}}(\rho, \pi) &= \frac{1}{2}\sum_{j=1}^q \frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}} + \frac{1}{2}\left[\sum_{j=1}^q \frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} - q\right] - \frac{1}{2}\sum_{j=1}^q \log\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} \\ &= \frac{1}{2}\left[\sum_{j=1}^q \frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}} + \sum_{j=1}^q\left(\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} - \log\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}}\right) - q\right], \end{aligned}$$

and the minimization problem becomes

$$\left(\hat{\mu}_\rho, \hat{\sigma}^2_\rho\right) := \arg\min_{\mu_\rho, \sigma^2_\rho}\frac{1}{n}\sum_{i=1}^n \psi(X_i, Y_i)\,\Phi\left(-\frac{V(X_i, Y_i, \mu_\rho)}{\sqrt{\sum_{j=1}^q \sigma^2_{\rho,j}\phi_j(X_i)^2}}\right) + \frac{1}{2\lambda}\sum_{j=1}^q\left(\frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}} + \frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} - \log\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}}\right). \tag{34}$$

Given the estimator $\hat{\sigma}^2_\rho = (\hat{\sigma}^2_{\rho,1}, ..., \hat{\sigma}^2_{\rho,q})$, we have

$$\hat{\mu}_\rho := \arg\min_{\mu_\rho}\left\{\frac{1}{n}\sum_{i=1}^n \psi(X_i, Y_i)\,\Phi\left(-\frac{V(X_i, Y_i, \mu_\rho)}{\sqrt{\sum_{j=1}^q \hat{\sigma}^2_{\rho,j}\phi_j(X_i)^2}}\right) + \frac{1}{2\lambda}\sum_{j=1}^q \frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}}\right\}.$$

The first term can be regarded as the empirical loss function for μ_ρ, and the second term is a weighted L_2 regularizer. If all of $\left\{\hat{\sigma}^2_{\rho,j}\right\}$ converge to zero, which is expected, then

$$\Phi\left(-\frac{V(X_i, Y_i, \mu_\rho)}{\sqrt{\sum_{j=1}^q \hat{\sigma}^2_{\rho,j}\phi_j(X_i)^2}}\right) \approx 1\left\{V(X_i, Y_i, \mu_\rho) < 0\right\}.$$

In addition to the weighted L_2 regularization, the PAC-Bayesian approach, therefore, also replaces the indicator $1\{V(X,Y,\mu_\rho)<0\}$, which is not smooth, by a smooth function $\Phi(-V(X,Y,\mu_\rho)/h)$ for a small h. Smoothing and regularizations are two built-in features of the PAC-Bayesian approach.

In the econometric literature, smoothing has been proposed to overcome the technical difficulties behind the maximum score estimator. See, for example, Horowitz (1992). In instrumental variable quantile regressions where an indicator function is present in the criterion function, Kaplan and Sun (2017) discuss several benefits of smoothing, including variance reduction and computational convenience. The PAC-Bayesian approach provides another justification for smoothing.

Lastly, we consider a particular form of the restrictive model considered here that will be utilized in the simulation section and is easier to implement. When $\Sigma_\pi = \Sigma_\rho = I_M$ and $\mu_\pi = 0$, we seek only to estimate μ_ρ and the optimization problem is now equivalent to

$$\hat{\mu}_\rho = \arg\min_{\mu_\rho} \frac{\lambda}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)\Phi\left(-\frac{V(X_i,Y_i,\mu_\rho)}{||\phi(X_i)||}\right)+\frac{1}{2}||\mu_\rho||^2. \tag{35}$$

The resulting decision rule is given by

$$a(x,\hat{\mu}_\rho) = \text{sign}[\phi(x)'\hat{\mu}_\rho - c(x)].$$

Here λ is a hyperparameter we will choose via cross validation. Alternatively, the version corresponding to the model class in (33) would drop the term $c(X_i)$, and is just a weighted version of the model derived in Germain et al. (2009), where the only difference in the objective function above is the weighting term $\psi(X_i,Y_i)$. Germain et al. (2009) utilizes the 0/1 based loss version of this model with $\{\phi_1(X),\ldots,\phi_M(X)\}$ taken as a set of weak learning decision stumps and show that the estimator performs competitively against AdaBoost in terms of misclassification rates on several real world data sets.

Note that (35) exhibits similarities with the soft-margin support vector machine, which selects $\hat{\mu}_{svm}$ to minimize the objective function

$$C\sum_{i=1}^{n}\left[1-Y_i\phi(X_i)'\mu_\rho\right]_+ + \frac{1}{2}||\mu_\rho||^2$$

for some constant $C>0$ and has classification rule $a(x,\hat{\mu}_{svm}) = \text{sign}[\phi(x)'\hat{\mu}_{svm}]$. In the restrictive PAC-Bayesian objective function in (35), a bounded and smooth "sigmoid" loss replaces the hinge loss of the SVM and now the terms in the objective function are weighted by $\psi(X_i,Y_i)$, the missed payoff from an incorrect decision.

3.3 PAC-Bayesian Multi-Model Aggregation

Here we consider the situation where there are multiple binary decision model classes of interest. Section 3.1 is general enough to encompass this setting with only some notational changes and reinterpretations. Here we detail the changes in the model space, prior specification, and posterior distribution, and present some implications relevant to implementation in this setting.

Suppose there are now K models indexed by $k=1,2,...,K$. Let $\theta_{(k)}\in\mathbb{R}^{q_k}$ be the parameter vector for model k. The number of parameters q_k can be different for a different model. For example, different decision boundaries may consist of a different subset of covariates, and the size

of the subset can be different. Denote $\theta = \left(k, \theta_{(k)}\right)$. The first component of θ signifies the model class, and the second component signifies the model parameter given the model class in the first component. The parameter space for θ is

$$\Theta = \cup_{k=1}^{K} \left(k \times \Theta_{(k)}\right),$$

where $\Theta_{(k)}$ is the parameter space for $\theta_{(k)}$. Given $\theta = \left(k, \theta_{(k)}\right) \in \Theta$, the action function, now denoted by $a_{(k)}(x, \theta_{(k)})$, maps the covariate space $\mathcal{X}$ to a binary action. The single model setting in Section 3.1 can be regarded as a special case here with $k = K = 1$.

As before, we equip Θ with the standard σ-algebra denoted by $\mathcal{B}_\theta$. PAC-Bayesian learning for model aggregation works in the same way as before. We need to specify a "prior" distribution π over the (model, parameter)-pairs $\left\{\left(k, \theta_{(k)}\right)\right\}$ in the measurable space $(\Theta, \mathcal{B}_\theta)$ and then use the performances of different pairs to update π to obtain an "evidence-based" distribution. The final decision rule involves aggregating the actions of all (model, parameter)-pairs using the evidence-based distribution.

To specify a distribution π over Θ, we first specify the distribution $\pi(k)$ over the model classes $k = 1, ..., K$ and then specify the distribution $\pi\left(\theta_{(k)}|k\right)$ over $\theta_{(k)} \in \Theta_{(k)}$ given the model class k. Let $\mathcal{K}^\circ$ be a subset of $\mathcal{K} := \{1, 2, ..., K\}$ and $\Theta^\circ_{(k)}$ be a measurable subset of $\Theta_{(k)}$. Then $\Theta^\circ = \cup_{k \in \mathcal{K}^\circ}(k \times \Theta^\circ_{(k)})$ is a measurable subset of Θ. Based on $\pi(k)$ and $\pi\left(\theta_{(k)}|k\right)$, $\pi(\Theta^\circ)$ is defined as

$$\pi(\Theta^\circ) = \sum_{k \in \mathcal{K}^\circ} \left[\pi(k) \cdot \int_{\Theta^\circ_{(k)}} d\pi\left(\theta_{(k)}|k\right)\right].$$

With some abuse of notation[2], we write the measure π as

$$\pi(\theta) := \pi\left(\left(k, \theta_{(k)}\right)\right) = \pi(k)\,\pi\left(\theta_{(k)}|k\right) \text{ for } \theta = \left(k, \theta_{(k)}\right). \tag{36}$$

This gives a general characterization of any distribution on $(\Theta, \mathcal{B}_\theta)$.

Given a $\pi \in \mathcal{P}(\Theta)$, we denote the family of all distributions on $(\Theta, \mathcal{B}_\theta)$ that is absolutely continuous with respect to π as $\mathcal{P}_\pi(\Theta)$. The evidence-based distribution we consider will belong to $\mathcal{P}_\pi(\Theta)$. For any $\rho \in \mathcal{P}_\pi(\Theta)$, define the Kullback–Leibler divergence between ρ and π as

$$D_{\mathrm{KL}}(\rho, \pi) = \sum_{k=1}^{K} \left\{\int_{\Theta_{(k)}} \log\left[\frac{\rho(k)}{\pi(k)} \cdot \frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right] d\rho\left(\theta_{(k)}|k\right)\right\} \rho(k).$$

This is the same definition as before but is tailored to the model aggregation setting with new interpretations of $\theta \in \Theta$ and the distribution over Θ.

Let $\mathcal{M}(\Theta)$ be the set of measurable functions on $(\Theta, \mathcal{B}_\theta)$ and

$$\mathcal{M}_b^\pi(\Theta) = \left\{A : A(\cdot, \cdot) \in \mathcal{M}(\Theta) \text{ and } \sum_{k=1}^{K}\left[\int_{\Theta_{(k)}} \exp\left(A\left(k, \theta_{(k)}\right)\right) d\pi\left(\theta_{(k)}|k\right)\right]\pi(k) < \infty\right\},$$

which is a subset of $\mathcal{M}(\Theta)$ that has a finite exponential moment under π. In this setting, Lemma 2 can be stated as follows.

[2]Here the meaning of $\pi(\cdot)$ depends on the argument supplied. We could write $\pi(\theta)$ as $\pi_{\boldsymbol{\theta}}(\theta)$, $\pi(k)$ as $\pi_{\boldsymbol{k}}(k)$ and $\pi\left(\theta_{(k)}|k\right)$ as $\pi_{\boldsymbol{\theta}_{(k)}|\boldsymbol{k}}\left(\theta_{(k)}|k\right)$ but we opt for a more economical notation. This should not cause any confusion.

Lemma 15 *For $\pi \in \mathcal{P}(\Theta)$ and $A \in \mathcal{M}(\Theta)$ such that $-A \in \mathcal{M}_b^{\pi}(\Theta)$, let $\rho_{A,\pi} \in \mathcal{P}_{\pi}(\Theta)$ be the probability measure on Θ defined by*

$$\rho_{A,\pi}(\theta) = \rho_{A,\pi}(k) \cdot \rho_{A,\pi}\left(\theta_{(k)}|k\right), \text{ for } \theta = \left(k, \theta_{(k)}\right),$$

where

$$\rho_{A,\pi}(k) = \frac{\pi(k)\,\nu_A(k)}{\sum_{j=1}^{K}\pi(j)\,\nu_A(j)},$$

$$\frac{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)} = \frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\nu_A(k)},$$

and

$$\nu_A(k) = \int_{\Theta_{(k)}} \exp\left(-A\left(k,\tilde{\theta}_{(k)}\right)\right) d\pi\left(\tilde{\theta}_{(k)}|k\right).$$

That is, for any measurable set $\Theta^{\circ} = \cup_{k\in\mathcal{K}^{\circ}}(k \times \Theta_{(k)}^{\circ}) \subseteq \Theta$,

$$\rho_{A,\pi}(\Theta^{\circ}) = \sum_{k\in\mathcal{K}^{\circ}}\left[\rho_{A,\pi}(k) \cdot \int_{\Theta_{(k)}^{\circ}} d\rho_{A,\pi}\left(\theta_{(k)}|k\right)\right].$$

Then, for any probability measure $\rho \in \mathcal{P}_{\pi}(\Theta)$ we have

$$\begin{aligned}
&\log\left[\sum_{k=1}^{K}\pi(k)\,\nu_A(k)\right] \\
&= -\left\{D_{\mathrm{KL}}(\rho,\pi) + \sum_{k=1}^{K}\left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)\right\} + D_{\mathrm{KL}}(\rho,\rho_{A,\pi}).
\end{aligned}$$

Note that $\log\left[\sum_{k=1}^{K}\pi(k)\,\nu_A(k)\right]$ does not depend on ρ. It follows from Lemma 15 that

$$\begin{aligned}
&\arg\min_{\rho\in\mathcal{P}_{\pi}(\Theta)}\left\{D_{\mathrm{KL}}(\rho,\pi) + \sum_{k=1}^{K}\left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)\right\} \\
&= \arg\min_{\rho\in\mathcal{P}_{\pi}(\Theta)} D_{\mathrm{KL}}(\rho,\rho_{A,\pi}) = \rho_{A,\pi}.
\end{aligned} \tag{37}$$

With the above details for the model aggregation setting, we can return to the optimization problem similar to that in (20). Let $R_n\left(k,\theta_{(k)}\right)$ be the empirical risk under model k with parameter $\theta_{(k)}$:

$$R_n\left(k,\theta_{(k)}\right) = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)1\left\{Y_i \neq a_{(k)}(X_i,\theta_{(k)})\right\}.$$

We now solve

$$\min_{\rho\in\mathcal{P}_{\pi}(\Theta)}\left[E_{(k,\theta_{(k)})\sim\rho}\left[R_n(k,\theta_{(k)})\right] + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right]. \tag{38}$$

To characterize the solution to the above minimization problem, we define the data-dependent measure on $\mathcal{K}$ as

$$\hat{\rho}_{\lambda}(k) = \frac{\pi(k)\,\hat{v}_{\lambda}(k)}{\sum_{j=1}^{K}\pi(j)\,\hat{\nu}_{\lambda}(j)} \tag{39}$$

and the data-dependent measure $\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)$ on $\Theta_{(k)}$ in terms of its RN derivative with respect to $\pi\left(\theta_{(k)}|k\right)$ as

$$\frac{d\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)} = \frac{\exp\left(-\lambda R_n\left(k,\theta_{(k)}\right)\right)}{\hat{\nu}_\lambda\left(k\right)},\ k=1,\ldots,K,$$

where

$$\hat{\nu}_\lambda\left(k\right) = \int_{\Theta_{(k)}} \exp\left(-\lambda R_n\left(k,\theta_{(k)}\right)\right) d\pi\left(\theta_{(k)}|k\right).$$

Based on $\hat{\rho}_\lambda\left(k\right)$ and $\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)$, we form the data-dependent measure $\hat{\rho}_\lambda\left(\theta\right) \in \mathcal{P}\left(\Theta\right)$ according to

$$\hat{\rho}_\lambda\left(\Theta^\circ\right) = \sum_{\ell\in\mathcal{L}^\circ}\left[\hat{\rho}_\lambda\left(k\right)\cdot\int_{\Theta^\circ_{(k)}} d\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)\right]. \tag{40}$$

This is our evidence-based distribution over (model, parameter)-pairs.

Letting

$$A\left(k,\theta_{(k)}\right) = \lambda R_n\left(k,\theta_{(k)}\right),$$

Lemma 15 and equation (37) thereafter show that $\hat{\rho}_\lambda\left(\theta\right)$ solves the problem in (38).

One approach to evaluate decision rules based on $\hat{\rho}_\lambda\left(\theta\right)$ is to simulate this distribution via reversible jump MCMC. Alternatively, as we consider in this paper, when the majority vote classifier is the object of interest, the form of $\hat{\rho}_\lambda\left(\theta\right)$ is amenable to the SMC method. For the single model class setting, the SMC approach is described in Section 4. One benefit of the SMC approach is that the procedure based on a single model class is easily adapted to the multiple model class setting. When the form of $\pi\left(\theta_{(k)}|k\right)$ does not depend on the choice for $\pi(k)$, this can reduce the computational burden if one is interested in choosing the prior component $\pi(k)$ over $\mathcal{K}$ from a set of potential distributions over $\mathcal{K}$ via cross-validation.

To see this, note that the majority vote (or, Bayesian) decision rule based on $\hat{\rho}_\lambda$ is

$$a_{B,\hat{\rho}_\lambda}(x) = \text{sign}\left\{E_{(k,\theta_{(k)})\sim\hat{\rho}_\lambda} a_{(k)}(x,\theta_{(k)})\right\} = \text{sign}\left\{\sum_{k=1}^{K}\hat{\rho}_\lambda\left(k\right)\hat{a}_{(k)}\left(x\right)\right\}, \tag{41}$$

where

$$\hat{a}_{(k)}\left(x\right) := \int_{\Theta_{(k)}} a_{(k)}(x,\theta_{(k)}) d\rho_{\hat{\lambda}}\left(\theta_{(k)}|k\right).$$

For a single model class $\mathcal{R}_{\Theta_{(k)}}$ with a given $\pi\left(\theta_{(k)}|k\right)$, under general conditions the SMC procedure produces accurate estimators for $\hat{a}_{(k)}\left(x\right)$ and $\hat{v}_\lambda(k)$, both of which depend only on $\pi\left(\theta_{(k)}|k\right)$. These objects can be computed separately for each $k\in\mathcal{K}$ according to the single model class SMC procedure. Then, for a given $\pi(k)$ over $\mathcal{K}$, (39) can be used to construct $\hat{\rho}_\lambda(k)$ and the majority vote rule is computed via (41). If one is interested in cross-validating the choice of $\pi(k)$ from some set of distributions on $\mathcal{K}$ and the distributions $\pi\left(\theta_{(k)}|k\right)$ do not depend on $\pi(k)$ for $k\in\mathcal{K}$, then the objects $\hat{a}_{(k)}\left(x\right)$ and $\hat{v}_\lambda(k)$ need only be computed once per cross-validation sample. This is in contrast to running a reversible jump MCMC procedure for each choice of $\pi(k)$ and can be beneficial when the number of decision model classes is not very large. If the number of model classes was very large, say, in an explanatory variable selection setting where the total number of explanatory variables is greater than the sample size, then an alternative computational strategy would be needed (to avoid running the SMC procedure independently for each model class). See, for example, Guedj (2013) for a discussion of PAC-Bayesian analysis and implementation for binary outcomes in such a setting.

4 Implementation

Here we consider implementation choices and describe some settings of the computational procedures that are applied in our simulations in Section 5. In Section 4.1 we discuss prior choices and consider examples for $\mathcal{R}_\Theta$. The $\mathcal{R}_\Theta$ considered center on decision models similar to those in Su (2020) and Elliott and Lieli (2013), some of which are used in our simulations. However, it should not be too difficult to make adjustments if a different model class is desired. In Section 4.2 we discuss the calculation of $\hat{\mu}_\rho$ in (35) associated with the linear decision rule discussed at the end of Section 3.2. We also outline an implementation of the SMC algorithm of Del Moral et al. (2006) in our setting in Section 4.2.

4.1 Model and Prior Choices

First we consider two specifications for $\mathcal{R}_\Theta$ of the form in (14) that are also considered in Su (2020). These consist of specifying a functional form for $m(x,\theta) \in \mathcal{M}_\Theta$ and the associated parameter space Θ. Then we consider potential choices for the prior probability distribution π. The $\mathcal{R}_\Theta$ specifications allow for $m(x,\theta)$ to be fairly general and are appropriate for a setting where the number of explanatory variables d is not large relative to the sample size n. If d is larger than n, the choices of function class and prior utilized in Guedj (2013) (Chapter 3) would be an option; an MCMC-based approach would be more appropriate in such a setting rather than the SMC procedure in Section 4.2.

In many empirical applications, we have a nondecreasing collection of parameterized function classes $\{\mathcal{M}_{\Theta_{(k)}}\}_{k=1}^K$ for $K \in \mathbb{N}$ where $\mathcal{M}_{\Theta_{(i)}} \subset \mathcal{M}_{\Theta_{(j)}}$ for $i < j$. In a single model class setting, we can take $\mathcal{M}_\Theta$ to be $\mathcal{M}_{\Theta_{(k)}}$ for some $k \in \mathcal{K} = \{1,\dots,K\}$ with parameter space $\Theta = \Theta_{(k)}$. In the multiple model class setting of Section 3.3, we can take $\mathcal{M}_\Theta = \cup_{k=1}^K \mathcal{M}_{\Theta_{(k)}}$ with parameter space $\Theta = \cup_{k=1}^K (k \times \Theta_{(k)})$. While the inclusion of the model class k as a component of the parameter θ may seem redundant when the model classes are nested, it simplifies the prior specification and allows for generalization when the model classes are not nested.

Example 1 *We consider polynomial transformations on $\mathcal{X}$ of order at most $k \in \mathcal{K}$. For $\mathcal{X} \subset \mathbb{R}^d$, the polynomial transformation of order at most k will have $q_k = \binom{d+k}{k}$ parameters, and it is defined as*

$$\mathcal{M}^{\text{poly}}_{\Theta_{(k)}} = \left\{ m(x,\theta) = \sum_{j=1}^{q_k} \theta_j \phi_j(x), \quad \theta_{(k)} = (\theta_1,\dots,\theta_{q_k}) \in \mathbb{R}^{q_k} \right\},$$

where the summation is over all monomials $\phi_j(x) = \prod_{\ell=1}^d x_\ell^{p_{j\ell}}$ with $\sum_{\ell=1}^d p_{j\ell} \leq k$ and $p_{j\ell} \in \mathbb{N} \cup \{0\}$. The parameter space associated with $\mathcal{M}^{\text{poly}}_{\Theta_{(k)}}$ is $\Theta_{(k)} = \mathbb{R}^{q_k}$.

Example 2 *Define $\Lambda(v) = (1+\exp(-v))^{-1}$. With the same parameter set $\Theta_{(k)} = \mathbb{R}^{q_k}$ as in Example 1, define the function space*

$$\mathcal{M}^{\text{logistic}}_{\Theta_{(k)}} = \left\{ m(x,\theta) = \Lambda\left(f(x,\theta)\right) : f(x,\theta) \in \mathcal{M}^{\text{poly}}_{\Theta_{(k)}} \right\}.$$

Now we consider some options for specifying the prior. First consider when $\mathcal{M}_\Theta$ and Θ correspond to a single model class, i.e., $\mathcal{M}_\Theta = \mathcal{M}_{\Theta_{(k)}}$ for some fixed $k \in \mathcal{K}$. In cases where it is reasonable to bound the parameter space Θ (for example, one could possibly replace $\Theta = \mathbb{R}^{q_k}$ with a bounded subset of $\mathbb{R}^{q_k}$ given some knowledge about the distribution of $P(X,Y)$), a uniform

prior over Θ is a potential choice. When $\Theta = \mathbb{R}^{q_k}$, another choice is a multivariate normal prior over Θ, for example, $N(0, \sigma_\pi^2 I_{q_k})$ for some $\sigma_\pi^2 > 0$. In the multiple model class setting with varying class complexity, a general strategy is to choose π that puts increasingly less weight on regions of the parameter space that are increasingly more complex. A prior that puts relatively more weight on very complex regions of the parameter space will tend to result in larger $D_{\mathrm{KL}}(\rho, \pi)$ terms in the bounds of Section 3.1 particularly as λ increases.

In our simulations, we use the following formulation for π in the $\Theta = \cup_{k=1}^{K}(k \times \Theta_{(k)})$ setting. We specify π as in (36), taking $\pi(\theta_k|k)$ to be the $N(0, \sigma_\pi^2 I_{q_k})$ distribution for $k \in \mathcal{K}$ with a fixed $\sigma_\pi^2 > 0$. To specify the model-class component $\pi(k)$ of the prior, in addition to simpler schemes such as equal weighting, one choice we consider is to set

$$\pi(k) = \frac{\exp\left(-\eta\xi\left(k, n\right)\right)}{z_\eta}, \quad z_\eta = \sum_{k=1}^{K} \exp(-\eta\xi(k, n)), \tag{42}$$

where $\eta \geq 0$ and $\xi(k, n) : \mathcal{K} \times \mathbb{N} \to \mathbb{R}_+$ is some measure of the complexity of model class k. Potential building blocks for $\xi(k, n)$ in the form of distribution-free model complexity measurements are as follows. For $k \in \mathcal{K}$, define $\mathcal{M}_{k,c} \equiv \{x \mapsto \mathrm{sign}(m(x, \theta) - c(x)) : m \in \mathcal{M}_{\Theta_{(k)}}\}$ and denote the growth function[3] of $\mathcal{M}_{k,c}$ by $\Pi_{k,c}(\cdot)$. Let $\psi_c(k, n)$ denote an upper bound for $\Pi_{k,c}(n)$ and $V_{k,c}$ denote an upper bound for the VC-dimension[4] of $\mathcal{M}_{k,c}$. That is, $\psi_c(k, n)$ upper bounds the maximum number of distinct ways that $\left\{\mathrm{sign}(m(x, \theta_{(k)}) - c(x)), \theta_{(k)} \in \Theta_k\right\}$ can classify any set of points in $\mathcal{X}^n$ while $V_{k,c}$ upper bounds the size of the largest sample that $\mathcal{M}_{k,c}$ could classify without error. To penalizes complexity, $\xi(k, n)$ can be taken to be an increasing function of $V_{k,c}$, $\psi_c(k, n)$, or both. In our simulations, we consider taking

$$\xi(k, n) = \sqrt{\log V_{k,c}}, \tag{43}$$

and also cross-validate η in (42) from a finite set of values.

Remark 1 below contains additional details regarding $V_{k,c}$ and $\psi_c(k, n)$ for Examples 1 and 2. These points are also noted in Su (2020); we refer the reader to their Section 3.1 and the references therein for additional discussion.

Remark 1 *When $\mathcal{M}_{\Theta_{(k)}}$ is specified as a vector space of real valued functions, the VC-dimension of $\mathcal{M}_{k,c}$ is given by the dimension of $\mathcal{M}_{\Theta_{(k)}}$ (c.f. Theorem 3.5 of Anthony and Bartlett (2009)). In particular, $\mathcal{M}^{\mathrm{poly}}_{\Theta_{(k)}}$ in Example 1 has dimension $\binom{d+k}{k}$ when $\mathcal{X}$ does not contain dummy variables, and so we can take $V_{k,c} = \binom{d+k}{k}$. For Example 2 with $\mathcal{M}^{\mathrm{logistic}}_{\Theta_{(k)}}$, Su (2020) shows that the VC-dimension of $\mathcal{M}_{k,c} = \{x \mapsto \mathrm{sign}(m(x, \theta) - c(x)) : m \in \mathcal{M}^{\mathrm{logistic}}_{\Theta_{(k)}}\}$ can be bounded above by $\binom{d+k}{k} + 1$, and hence we can take $V_{k,c} = \binom{d+k}{k} + 1$. Regarding $\psi_c(k, n)$, if $V_{k,c}$ bounds the VC-dimension of $\mathcal{M}_{k,c}$, it follows from Theorems 3.5 and 3.6 in Anthony and Bartlett (2009) that $\Pi_{k,c}(n)$ can be upper bounded by*

$$\psi_c(k, n) = \begin{cases} 2^n, & \text{if } n \leq V_{k,c} \\ \left(\frac{en}{V_{k,c}}\right)^{V_{k,c}}, & \text{if } n > V_{k,c}. \end{cases}$$

[3] For a collection $\mathcal{H}$ of functions from $\mathcal{X}$ to $\{-1, 1\}$, $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ is defined by $\Pi_{\mathcal{H}}(\ell) = \max_{(x_1,\ldots,x_\ell)\in\mathcal{X}^\ell} |\{(h(x_1), \ldots, h(x_\ell)) : h \in \mathcal{H}\}|$

[4] The VC-dimension of $\mathcal{M}_{k,c}$ is the largest integer ℓ such that $\Pi_{k,c}(\ell) = 2^\ell$.

4.2 Implementation for Methods in Section 3

In Section 5, our simulations evaluate the majority vote or Bayes method with the Gibbs posterior $\hat{\rho}_\lambda$ in Definition 4 and also with the linear decision rule in Section 3.2 associated with the optimization problem in (35). Here we first detail our approach to computing $\hat{\mu}_\rho$ in the latter case and then outline the SMC approach applied to implement $\hat{\rho}_\lambda$ in the former and more general case. We address only the single model class setting here. The discussion in Section 3.3 highlights how this can be adapted to the multiple model class setting for $\hat{\rho}_\lambda$.

First, we make a computational adjustment so that the choice of the hyperparameter λ is invariant to the units of measurement of the utility function. For $\mathcal{P}^* \subseteq \mathcal{P}_\pi(\Theta)$ and $M \in \mathbb{N}$, note that cross-validating $\lambda \in \{\lambda_1, \ldots, \lambda_M\}$ among distributions in

$$\arg\min_{\rho \in \mathcal{P}^*} \left[\int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right]$$

is equivalent to cross-validating $\lambda \in \{\lambda_1 \bar{\psi}, \ldots, \lambda_M \bar{\psi}\}$ among distributions in

$$\arg\min_{\rho \in \mathcal{P}^*} \left[\int_\Theta \bar{R}_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right], \tag{44}$$

where $\bar{R}_n(\theta) = R_n(\theta)/\bar{\psi}$, and $\bar{\psi} = n^{-1} \sum_{i=1}^n \psi(X_i, Y_i)$. We work with the adjusted minimization problem in (44). In the general setting where the Gibbs posterior $\hat{\rho}_\lambda$ is considered, $\mathcal{P}^*$ is $\mathcal{P}_\pi(\Theta)$ whereas in the linear decision setting associated with the optimization problem in (35), $\mathcal{P}^*$ is the set of normal distributions over Θ with an identity covariance matrix (and π is the standard normal distribution).

To compute $\hat{\mu}_\rho$ in the setting of Section 3.2 discussed above, we follow a similar strategy to that of Germain et al. (2009) who analyze the 0/1-loss version of this problem. Incorporating the adjustment in (44) into the objective in (35), the optimization problem is now

$$\hat{\mu}_\rho = \arg\min_{\mu_\rho} \frac{\lambda}{n} \sum_{i=1}^n \frac{\psi(X_i, Y_i)}{\bar{\psi}} \Phi\left(-\frac{V(X_i, Y_i, \mu_\rho)}{||\phi(X_i)||} \right) + \frac{1}{2} ||\mu_\rho||^2.$$

The gradient of the objective function above with respect to μ_ρ is given by

$$-\frac{\lambda}{n} \sum_{i=1}^n \frac{\psi(X_i, Y_i)}{\bar{\psi}} \dot{\Phi}\left(\frac{Y_i [\phi(X_i)\mu_\rho - c(X_i)]}{||\phi(X_i)||} \right) \frac{Y_i \phi(X_i)}{||\phi(X_i)||} + \mu_\rho,$$

where $\dot{\Phi}$ denotes the standard normal probability density function. For a given value of λ, $\hat{\mu}_\rho$ is calculated by gradient descent. As there can be multiple local minima, we tried 15 random starting points when $\lambda/n \leq 10$ and 100 random starting points when $\lambda/n > 10$. We performed 5-fold cross validation to select $\lambda \in \{2^0, 2^1, \ldots, 2^{18}\}$. The discussion and references in Alquier et al. (2016) suggest alternative implementation methods. These can be useful for the more general settings in Section 3.2, for example when the covariance matrix Σ_ρ is not set to the identity matrix.

To implement the majority vote rule based on $\hat{\rho}_\lambda$ in Definition 4, now with $R_n(\theta)$ replaced by $\bar{R}_n(\theta)$, we utilize the tempering SMC procedure of Del Moral et al. (2006). While MCMC is a typical choice for simulating from $\hat{\rho}_\lambda$, recently Ridgway et al. (2014) and Alquier et al. (2016) have highlighted the usefulness of the SMC procedure in various PAC-Bayesian settings.

One benefit is that each run of the procedure produces a sample from each member of a set of Gibbs posterior distributions corresponding increasing λ values. This can ease the computational burden of cross-validation.

To touch on a few elements of the tempering SMC algorithm in our setting, assume $\Theta = \mathbb{R}^q$ for some $q \in \mathbb{N}$ and that we are able to sample from a prior probability distribution π over Θ. It is assumed that there is an increasing temperature ladder

$$0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T, \ T \in \mathbb{N}.$$

$\{\lambda_t\}_{t=0}^T$ here is not generally the same set that was considered for cross-validation in the earlier procedure for the linear decision rule. The temperature ladder is intended to be such that as λ_t increases, the corresponding distributions $\hat{\rho}_{\lambda_t}$ progress gradually from $\pi = \hat{\rho}_{\lambda_0}$ to distributions $\hat{\rho}_{\lambda_t}$ with higher values of λ_t that are of greater interest. For each $t = 0, \ldots, T$, the SMC algorithm produces a set of weighted samples, $\{W_t^{(i)}, \theta_t^{(i)}\}_{i=1}^N$ with $W_t^{(i)} > 0$ and $\sum_{i=1}^N W_t^{(i)} = 1$, of size N and a scaling factor estimate $\hat{Z}_t$. The set of parameter draws $\{\theta_t^{(i)}\}_{i=1}^N$ are referred to as particles (there are N weighted particles for each t). SMC combines MCMC moves with sequential importance sampling. This produces weighted particles that emulate, in terms of computing expectations, samples from the probability distributions $\hat{\rho}_{\lambda_t}$ associated with the densities

$$\frac{d\hat{\rho}_{\lambda_t}}{d\pi}(\theta) = \frac{\exp\left[-\lambda_t \bar{R}_n(\theta)\right]}{Z_t}, \quad Z_t = \int_\Theta \exp\left[-\lambda_t \bar{R}_n(\theta)\right] d\pi(\theta), \quad t = 0, 1, \ldots, T.$$

Under general conditions, for a $\hat{\rho}_{\lambda_T}$-integrable function $\varphi : \Theta \to \mathbb{R}$,

$$\sum_{i=1}^N W_T^{(i)} \varphi\left(\theta_T^{(i)}\right) \overset{a.s.}{\to} E_{\theta \sim \hat{\rho}_{\lambda_T}} \varphi(\theta),$$

as $N \to \infty$ while $\hat{Z}_T$ is consistent for Z_T. In our setting we are interested in $\varphi(\theta) = a(x, \theta)$ where $a(x, \theta) \in \mathcal{R}_\Theta$, enabling us to compute the key ingredient to the majority vote decision rule. For additional details regarding the SMC procedure and its applications, we refer to Del Moral et al. (2006) and Jasra et al. (2007).

The SMC algorithm we apply in Section 5 is detailed below. We set the input parameters τ_{ESS} and N there equal to 1/2 and 1000, respectively. For the $\{\lambda_t\}_{t=1}^T$ input, we adopt the piece-wise linear structure utilized in the simulations of Del Moral et al. (2006) and Jasra et al. (2007) with $T = 320$ and $\lambda_T = 1600$. In particular, the first 20% of steps increase uniformly from 0 to 0.15×1600 (i.e. $\lambda_j = (j/64) \times 240$ for $j = 1, \ldots, 64$), the next 40% of steps increase uniformly from 240 to 0.4×1600 (i.e. $\lambda_j = 240 + (j/128) \times 400$ for $j = 65, \ldots, 192$), and the last 40% of steps increase uniformly from 640 to 1600 (i.e. $\lambda_j = 640 + (j/128) \times 960$ for $j = 193, \ldots, 320$). In practice, it may be beneficial to consider higher (or lower) values of λ_T and or include a greater number of steps (higher T value). Depending on the data generating process, higher values of λ_T can push some components such as $\hat{Z}_t$ close to machine epsilon for t near T. One can experiment a little to check that the temperature range doesn't appear to be limited unnecessarily and if increasing the number of steps improves performance in cross-validation samples. Alternatives to the piece-wise linear ladder design are discussed in Del Moral et al. (2006) and Jasra et al. (2007). Additionally, the SMC algorithm requires a resampling step. We utilize systematic resampling, which is also outlined below. Additional algorithm choices and cross-validation points are detailed below the algorithm descriptions.

Tempering SMC Algorithm

Input N (number of particles), $\tau_{\text{ESS}} \in (0,1)$ (ESS threshold), $\{\lambda_t\}_{t=1}^T$ (temperature ladder with $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_T$).

Output $\{W_t^{(i)}, \theta_t^{(i)}\}_{i=1}^N$ for $t = 0, \ldots, T$, $\{\hat{Z}_t\}_{t=1}^T$.

Step 1: initialization

- Set $t \leftarrow 0$, $\hat{Z}_0 \leftarrow 1$. For $i = 1, \ldots, N$, draw $\theta_0^{(i)} \sim \pi$ and set $W_0^{(i)} \leftarrow 1/N$.

Iterate steps 2 and 3
Step 2: Resampling

- If
$$\left\{\sum_{i=1}^N \left(W_t^{(i)}\right)^2\right\}^{-1} < \tau_{\text{ESS}} N,$$
resample $\left\{W_t^{(i)}, \theta_t^{(i)}\right\}_{i=1}^N$ yielding equally weighted resampled particles $\left\{\frac{1}{N}, \overline{\theta}_t^{(i)}\right\}_{i=1}^N$ and set $\left\{W_t^{(i)}, \theta_t^{(i)}\right\}_{i=1}^N \leftarrow \left\{\frac{1}{N}, \overline{\theta}_t^{(i)}\right\}_{i=1}^N$. Otherwise, leave $\left\{W_t^{(i)}, \theta_t^{(i)}\right\}_{i=1}^N$ unaltered.

Step 3: Sampling

- Set $t \leftarrow t+1$; if $t = T+1$, stop.
- For $i = 1, \ldots, N$, draw $\theta_t^{(i)} \sim K_t(\theta_{t-1}^{(i)}, \cdot)$, where K_t is an MCMC kernel with invariant distribution ρ_{λ_t}, and evaluate the unnormalized importance weights
$$\omega_t^{(i)}\left(\theta_{t-1}^{(i)}\right) = \exp\left[-(\lambda_t - \lambda_{t-1})\bar{R}_n\left(\theta_{t-1}^{(i)}\right)\right].$$
- For $i = 1, \ldots, N$, set
$$W_t^{(i)} \leftarrow \frac{W_{t-1}^{(i)}\omega_t\left(\theta_{t-1}^{(i)}\right)}{\sum_{j=1}^N W_{t-1}^{(j)}\omega_t\left(\theta_{t-1}^{(j)}\right)}, \quad \hat{Z}_t \leftarrow \hat{Z}_{t-1} \times \left\{\sum_{i=1}^N W_{t-1}^{(i)}\omega_t\left(\theta_{t-1}^{(i)}\right)\right\}.$$

Resampling Algorithm (systematic resampling):

Input A set of (normalized) weights and associated particles, $\left\{W_t^{(i)}, \theta_t^{(i)}\right\}_{i=1}^N$ for some $t \in \{0, \ldots, T\}$.

Output Resampled particles for equal weighting, $\left\{\overline{\theta}_t^{(i)}\right\}_{i=1}^N$

- Draw $u \sim U\left[0, \frac{1}{N}\right]$.
- Compute cumulative weights $C^{(i)} = \sum_{m=1}^i W_t^{(m)}$ for $i = 1, ..., N$.

- Set $m \leftarrow 1$.
- **For** $i = 1 : N$

  While $u < C^{(i)}$ **do** $\overline{\theta}_t^{(m)} \leftarrow \theta_t^{(i)}$.

  $m \leftarrow m + 1$, and $u \leftarrow u + 1/N$.

  End For

For the MCMC kernel in the sampling step of the SMC algorithm, we use a Gaussian random-walk Metropolis kernel with covariance matrix proportional to the empirical covariance matrix of the current set of particles. We scale the empirical covariance of the step t particles by $1/t$ which produced produced reasonable acceptance rates in the first simulated training set across the various simulation setups. The priors utilized for the majority vote associated with the Gibbs posterior in our simulations are described in Section 5 below. We use 5-fold cross-validation to select λ from λ_t values for which $t > 25$.

5 Simulation Study

To investigate the performance of the utility-based PAC-Bayesian decision rules, we consider two data generating processes and two sets of preferences, one set with each DGP. We utilize the same simulation design as Elliott and Lieli (2013) and Su (2020). The DGPs and the associated sets of preferences are as follows.

DGP 1: $\mathcal{X} = [-2.5, 2.5]$, $X \sim 5 \times \text{Beta}(1, 1.3) - 2.5$, and $P(x) = \Lambda(-0.5X + 0.2X^3)$ where $\Lambda(\cdot)$ is the logistic function described in Example 2 and recall $P(x)$ is defined in (2).

- Preference 1: $b(x) = 20$ and $c(x) = 0.5$.
- Preference 2: $b(x) = 20$ and $c(x) = 0.5 + 0.025X$.

DGP 2: $\mathcal{X} = [-3.5, 3.5]^2$, covariates X_1 and X_2 are each uniformly distributed on $[-3.5, 3.5]$ and are independent of one another, and $P(x_1, x_2) = \Lambda(Q(1.5x_1 + 1.5x_2))$ where $Q(v) = (1.5 - 0.1v)\exp\{-(0.25v + 0.1v^2 - 0.04v^3)\}$.

- Preference 3: $b((x_1, x_2)) = 20$ and $c((x_1, x_2)) = 0.75$.
- Preference 4: $b((x_1, x_2)) = 20 + 40 \cdot 1\{|x_1 + x_2| < 1.5\}$ and $c((x_1, x_2)) = 0.75$.

To evaluate the performance of a decision rule, we compute, by Monte Carlo simulation, the ratio of its expected utility to the expected utility of the optimal decision in (9) if $P(x)$ were known. This metric is intuitive as utility has no natural unit, however the ratio changes when a constant is added to the utility function. In Elliott and Lieli (2013) and Su (2020), this is dealt with by choosing some normalization of the utility function. We follow the same normalization and Monte Carlo setup as Su (2020), so that our simulation results can be compared directly to theirs. Noting that

$$U(a, y, x) = \frac{1}{4}b(x)\left[y + 1 - 2c(x)\right]a + \frac{1}{4}b(x)\left[y + 1 - 2c(x)\right] + U(-1, y, x),$$

Su (2020) normalizes the utility function by setting $U(-1, y, x) = -0.25b(x)[y + 1 - 2c(x)]$ for all $x \in \mathcal{X}$ and multiplying the utility function by 4. For any decision rule $a_n(x) : \mathcal{X} \to \{-1, 1\}$, this results in the following measurement that he calls the generalized expected utility,

$$S(a_n) = E\{b(X)[Y + 1 - 2c(X)]a_n(X)\}.$$

With this normalization, denote

$$a^*(x) = \text{sign}[P(x) - c(x)],\ x \in \mathcal{X},$$

i.e., a^* is the optimal forecast rule. Then define the relative generalized expected utility (RGEU) of any decision rule a_n by

$$\text{RGEU}(a_n) \equiv \frac{E[S(a_n(X))]}{S(a^*(X))}.$$

As noted in Su (2020), the RGEU of the decision rule a_n can be approximated by simulation as

$$\text{RGEU}(a_n) = E\left[\frac{S(a)}{S(a^*)}\right] \simeq \frac{1}{\mathcal{S}}\sum_{j=1}^{\mathcal{S}} \frac{S_{\ell,j}(a|\mathcal{D}_{n,j})}{S_{\ell,j}(a^*)}.$$

Here, $S_{\ell,j}(a_n|\mathcal{D}_{n,j})$ is the jth out of sample empirical utility with training sample size ℓ of the decision rule a_n, which is estimated on the jth training sample $\mathcal{D}_{n,j}$ with training sample size n. $S_{\ell,j}(a^*)$ is the jth out-of-sample empirical utility with training sample size ℓ of a^*, and $\mathcal{S}$ is the number of simulation replications. Still following Su (2020), we take $n \in \{500, 1000\}$, $\ell = 5000$, and $\mathcal{S} = 500$.

We compare the following models. Firstly, we consider maximum likelihood estimators, which are denoted by ML in Tables 1 and 2. For $k = 1, 2, 3$, the maximum likelihood estimator presumes a logistic model linear in the polynomial transformations of the $\mathcal{X}$ up to order k. Secondly, we consider the maximum utility estimator of Elliott and Lieli (2013) (denoted MU); it is presumed that $m(x, \theta)$ belongs to the class of polynomial transformations of $\mathcal{X}$ for $k = 1, 2, 3$ for these decision rules. Hence the ML estimator is correctly specified for $P(X)$ when $k = 3$ for DGP 1. Thirdly, we consider one of the best performing (in this simulation design) model selection procedures from Su (2020), based on the simulated maximal discrepancy penalty. This is a penalized version of the MU models here (selecting the best k among $k = 1, 2, 3$). This model is denoted MU-SMD. Fourthly, we consider the linear PAC-Bayesian model associated with (35) from Section 3.2 when the posterior is also constrained to be normal with identity covariance matrix. Here we take $\{\phi_1, \ldots, \phi_{q_3}\}$ to consist of the polynomial transformations of $\mathcal{X}$ up to order 3. We normalize the data (using training sample mean and standard deviation) as is common with SVM. This model is denoted PB-NP (NP for normal posterior). Lastly, we consider the non-constrained PAC-Bayesian method whereby the decision rule is the majority vote associated with the Gibbs posterior $\hat{\rho}_\lambda$ in Definition 4. In this case, we consider the multiple model class setting of Section 3.3. For the model classes, we use consider 3 classes of polynomial transformations on $\mathcal{X}$ of orders $k \in \{1, 2, 3\} = \mathcal{K}$ as specified in Example 1. We cross validate λ according to the temperature ladder described in Section 4.2 and take $\pi(\theta_{(k)}|k)$ to be $N(0, 4I_{q_k})$ for each k. These decision rules are denoted PB-GP (GP for Gibbs posterior). To specify $\pi(k)$ for $k \in \mathcal{K}$, we evaluate three choices. First, we take $\pi(k) = 1/3$ for $k = 1, 2, 3$; this is denoted EQ. Second, we take $\pi(k) = q_k/(\sum_{j=1}^3 q_j)$ where q_k is defined in Example 1 and denotes the number of parameters associated with model class k; this is denoted NP. Third, we utilize the weights in (43) and cross-validate $\tau \in \{2^{-2}, 2^{-1}, \ldots, 2^3\}$; this prior choice is denoted CV in the tables.

The simulation results are presented in Tables 1 and 2 after the Conclusion. The utility-based PAC-Bayesian decision models PB-GP and PB-NP perform very well, achieving higher RGEU than the MU and MU-SMD decision rules across all preferences and DGPs. The margin of the improvements is often sizable. Only the ML rule with a correctly specified DGP (ML with $k = 3$ for DGP 1) outperforms the BP- models. However, whenever the ML procedure is misspecified, it mostly performs quite poorly relative to all the utility-based methods. This performance further deteriorates when the preferences vary with the covariates as they do for Preferences 2 and 4. As shown in Elliott and Lieli (2013), the cubic MU ($k = 3$) is correctly specified in both the DGP 1 and DGP 2 settings. However, it is also observed there that MU can be prone to overfitting and aided by model selection procedures. Nonetheless, the PB- models outperform against the MU-SMD procedure in this simulation setting as well.

The restricted PAC-Bayesian decision model, PB-NP, performs slightly worse than the general version associated with the Gibbs posterior, PB-GP in most settings. However, the margin between the PB-NP and PB-GP models is not always very sizable. This may suggest that the restricted model can stand on its own, particularly when the sample size is larger or when there is a set $\{\phi_j(x)\}_{j=1}^q$ of interest that could be difficult to work into a more general Gibbs posterior setting. For example, when $\{\phi_j(x)\}_{j=1}^q$ is a larger set of weak learners as in Germain et al. (2009), the setting of Section 3.2 may be easier to implement. Lastly, we did not observe much of an impact on the RGEU from cross validating the choice of τ in the prior $\pi(k)$.

6 Conclusion

An asymmetric payoff structure is often a salient feature of economic decision making problems. For the binary decision/forecast problem where the decision maker faces asymmetric payoffs that vary with observable variables, we propose a PAC-Bayesian approach. We show that many key elements of the PAC-Bayesian classification literature can be extended to accommodate this setting, deriving high probability training sample bounds and oracle inequalities that suggest decision rules of interest. The decision rules perform very well against alternatives methods in Monte Carlo experiments, allow for flexible functional decision rule forms, allow for valid training-sample risk bounds and confidence interval computation, and can take advantage of Bayesian estimation machinery.

Table 1: Relative generalized expected utility, $n = 500$

DGP 1	$P(x) = \Lambda(-0.5x + 0.2x^3)$					
Preference	$b(x) = 20$, $c(x) = 0.5$			$b(x) = 20$, $c(x) = 0.5 + 0.025x$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	81.74	81.28	80.72	81.84	81.82	80.41
PB-NP	74.21			77.13		
MU-SMD	65.54			58.87		
Poly. order:	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ML	34.16	29.57	93.09	9.13	10.92	94.37
MU	51.02	52.85	65.74	32.40	43.80	53.26
DGP 2	$P(x) = \Lambda(Q(1.5x_1 + 1.5x_2))$, $Q(v) = \frac{(1.5-0.1v)}{\exp(0.25v+0.1v^2-0.04v^3)}$					
Preference	$b(x) = 20$, $c(x) = 0.75$			$b(x) = 20 + 1\lvert x_1 + x_2\rvert < 1.5$, $c(x) = 0.75$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	72.81	72.62	72.49	61.77	61.65	61.40
PB-NP	69.75			56.45		
MU-SMD	68.81			52.84		
Poly. order:	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ML	60.17	58.75	59.48	29.52	27.87	33.81
MU	66.71	51.87	67.69	48.35	33.14	51.47

Note: The MATLAB packages *glmfit* and *simulannealbnd* with default settings for each algorithm were used to compute the ML and MU models. The code implementing the ML, MU and MU-SMD models was provided by the author of Su (2020).

Table 2: Relative generalized expected utility, $n = 1000$

DGP 1	$P(x) = \Lambda(-0.5x + 0.2x^3)$					
Preference	$b(x) = 20$, $c(x) = 0.5$			$b(x) = 20$, $c(x) = 0.5 + 0.025x$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	87.73	87.86	87.47	90.81	90.65	90.43
PB-NP	81.52			88.83		
MU-SMD	70.75			67.30		
Poly. order:	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ML	30.92	30.03	96.97	7.12	6.26	97.42
MU	53.24	58.19	69.50	36.91	49.13	60.41
DGP 2	$P(x) = \Lambda(Q(1.5x_1 + 1.5x_2))$, $Q(v) = \frac{(1.5-0.1v)}{\exp(0.25v+0.1v^2-0.04v^3)}$					
Preference	$b(x) = 20$, $c(x) = 0.75$			$b(x) = 20 + 1\lvert x_1 + x_2\rvert < 1.5$, $c(x) = 0.75$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	78.61	78.46	78.25	70.09	70.18	69.86
PB-NP	73.72			63.75		
MU-SMD	71.94			59.72		
Poly. order:	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ML	58.71	57.36	59.40	27.16	24.15	31.14
MU	69.97	58.14	70.81	54.92	39.24	55.97

Note: The MATLAB packages *glmfit* and *simulannealbnd* with default settings for each algorithm were used to compute the ML and MU models. The code implementing the ML, MU and MU-SMD models was provided by the author of Su (2020).

7 Appendix of Proofs

7.1 Proofs for Section 2

Proof of Lemma 1. First we will show that for any $(x, y) \in \mathcal{X} \times \{-1, 1\}$,

$$\psi(x, y) 1\{y \neq a_{B,\rho}(x)\} \leq 2E_{\theta\sim\rho}\psi(x, y)1\{y \neq a(x, \theta)\}. \tag{45}$$

To show this, note that $\psi(x, y) > 0$ by Assumption 1 (i) and the fact that $\psi(x, y) = U(1, 1, x) - U(-1, 1, x)$ when $y = 1$ and $\psi(x, y) = U(-1, -1, x) - U(1, -1, x)$ when $y = -1$. Therefore, (45) holds when $y = a_{B,\rho}(x) = \text{sign}\{E_{\theta\sim\rho}a(x, \theta)\}$ as then the left hand side is zero. When $y \neq a_{B,\rho}(x)$, this implies that $y \cdot E_{\theta\sim\rho}a(x, \theta) \leq 0$. Therefore in the alternative case when $y \neq a_{B,\rho}(x)$,

$$\begin{aligned}
\psi(x, y)1\{y \neq a_{B,\rho}\} &= \psi(x, y) \\
&\leq \psi(x, y)\{1 - yE_{\theta\sim\rho}a(x, \theta)\} \\
&= 2E_{\theta\sim\rho}\psi(x, y)\frac{1}{2}\{1 - y \cdot a(x, \theta)\} \\
&= 2E_{\theta\sim\rho}\psi(x, y)1\{y \neq a(x, \theta)\}.
\end{aligned}$$

This shows that (45) holds. By (45) and the monotonicity of expectation,

$$E_{X,Y\sim P(X,Y)}\psi(X, Y)1\{Y \neq a_{B,\rho}(X)\} \leq 2E_{X,Y\sim P(X,Y)}E_{\theta\sim\rho}\psi(X, Y)1\{Y \neq a(X, \theta)\}.$$

The statement of Lemma 1 then follows from an application of Fubini's theorem. ■

Proof of Lemma 2. By definition, we have

$$\begin{aligned}
&D_{\text{KL}}(\rho, \rho_{A,\pi}) \\
&= \int_\Theta \log\left[\frac{d\rho}{d\rho_{A,\pi}}(\theta)\right] d\rho(\theta) \\
&= \int_\Theta \log\left\{\frac{d\rho}{d\pi}(\theta)\left[\frac{d\rho_{A,\pi}}{d\pi}(\theta)\right]^{-1}\right\} d\rho(\theta) \\
&= \int_\Theta \left[\log\frac{d\rho}{d\pi}(\theta) - \log\frac{\exp(-A(\theta))}{\int_\Theta \exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)}\right] d\rho(\theta) \\
&= \int_\Theta A(\theta) d\rho(\theta) + \int_\Theta \log\left[\int_\Theta \exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)\right] d\rho(\theta) + \int_\Theta \left[\log\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta) \\
&= \int_\Theta A(\theta) d\rho(\theta) + \log\left[\int_\Theta \exp(-A(\theta)) d\pi(\theta)\right] + \int_\Theta \left[\log\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta) \\
&= \int_\Theta A(\theta) d\rho(\theta) + \log\left[\int_\Theta \exp(-A(\theta)) d\pi(\theta)\right] + D_{\text{KL}}(\rho, \pi).
\end{aligned}$$

Hence,

$$\log\left[\int_\Theta \exp(-A(\theta)) d\pi(\theta)\right] = -\left[\int_\Theta A(\theta) d\rho(\theta) + D_{\text{KL}}(\rho, \pi)\right] + D_{\text{KL}}(\rho, \rho_{A,\pi}).$$

■

Proof of Corollary 3.

Part (a). Since $\rho_{A,\pi} = \arg\min_{\rho\in\mathcal{P}_\pi(\Theta)} D_{\mathrm{KL}}(\rho, \rho_{A,\pi})$ and the left hand side of (21) does not depend on ρ, we have

$$\begin{aligned}\rho_{A,\pi} &= \arg\max_{\rho\in\mathcal{P}_\pi(\Theta)} -\left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] \\ &= \arg\min_{\rho\in\mathcal{P}_\pi(\Theta)} \left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right].\end{aligned}$$

By (21), we then have

$$\begin{aligned}&\min_{\rho\in\mathcal{P}_\pi(\Theta)} \left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] \\ &= \int_\Theta A(\theta)\, d\rho_{A,\pi}(\theta) + D_{\mathrm{KL}}(\rho_{A,\pi},\pi) \\ &= -\log\left[\int_\Theta \exp(-A(\theta))\, d\pi(\theta)\right].\end{aligned}$$

Part (b). Taking $A = -\mathcal{A}$ in Lemma 2, we obtain that for any probability measure $\rho \in \mathcal{P}_\pi(\Theta)$,

$$\log\left[\int_\Theta \exp(\mathcal{A}(\theta))\, d\pi(\theta)\right] = \left[\int_\Theta \mathcal{A}(\theta)\, d\rho(\theta) - D_{\mathrm{KL}}(\rho,\pi)\right] + D_{\mathrm{KL}}(\rho,\rho_{-A,\pi}). \tag{46}$$

Note that $D_{\mathrm{KL}}(\rho,\rho_{-A,\pi}) \geq 0$. It follows from (46) that

$$\begin{aligned}\log\left[\int_\Theta \exp(\mathcal{A}(\theta))\, d\pi(\theta)\right] &= \left[\int_\Theta \mathcal{A}(\theta)\, d\rho(\theta) - D_{\mathrm{KL}}(\rho,\pi)\right] + D_{\mathrm{KL}}(\rho,\rho_{-A,\pi}) \\ &\geq \left[\int_\Theta \mathcal{A}(\theta)\, d\rho(\theta) - D_{\mathrm{KL}}(\rho,\pi)\right].\end{aligned}$$

This implies that

$$\int_\Theta \mathcal{A}(\theta)\, d\rho(\theta) \leq D_{\mathrm{KL}}(\rho,\pi) + \log\left[\int_\Theta \exp(\mathcal{A}(\theta))\, d\pi(\theta)\right].$$

■

7.2 Proofs for Section 3.1

Proof of Theorem 5. Let $A(\theta) = \lambda D[R(\theta), R_n(\theta)]$ and $\lambda \in I$. (25) and Fubini's theorem imply that

$$\int_\Theta \exp(\lambda D[R(\theta), R_n(\theta)])\, d\pi(\theta) < \infty$$

holds almost surely. Therefore, by Corollary 3 (b), the event

$$\begin{aligned}&\left\{\int_\Theta \lambda D[R(\theta), R_n(\theta)]\, d\rho(\theta)\right. \\ &\left.\leq \log\left[\int_\Theta \exp(\lambda D[R(\theta), R_n(\theta)])\, d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho,\pi) \text{ for all } \rho\in\mathcal{P}_\pi(\Theta) \text{ simultaneously}\right\}\end{aligned} \tag{47}$$

occurs with probability one. Applying Jensen's inequality to the object on the left-hand side of the inequality in this event, we obtain that

$$\begin{aligned}
&\Pr\Big\{\lambda D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] \\
&\leq \log\left[\int_\Theta \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho,\pi) \text{ for all } \rho \in \mathcal{P}_\pi(\Theta) \text{ simultaneously}\Big\} \\
&= 1. \qquad (48)
\end{aligned}$$

Next, we establish a high-probability bound for $\log\left[\int_\Theta \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi\left(\theta\right)\right]$ using the Markov inequality: for any constant C,

$$\begin{aligned}
&\Pr\left\{\log\left[\int_\Theta \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi\left(\theta\right)\right] > C\right\} \\
&\leq \Pr\left\{\left[\int_\Theta \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi\left(\theta\right)\right] > \exp C\right\} \\
&\leq \frac{E\left[\int_\Theta \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi\left(\theta\right)\right]}{\exp C} \\
&= \frac{\int_\Theta E\exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi\left(\theta\right)}{\exp C} \leq \exp\left(f\left(\lambda, n\right) - C\right).
\end{aligned}$$

where the equality follows from Fubini's theorem and the last inequality follows from (25).

Solving the equation $\exp\left(f\left(\lambda, n\right) - C\right) = \epsilon$ for C, we find

$$C = f\left(\lambda, n\right) + \log\frac{1}{\epsilon}.$$

So

$$\Pr\left\{\log\left[\int_\Theta \exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi\left(\theta\right)\right] \leq f\left(\lambda, n\right) + \log\frac{1}{\epsilon}\right\} \geq 1 - \epsilon.$$

Note that the above high probability bound does not involve ρ. Combining this with (48), we have

$$\begin{aligned}
&\Pr\left\{D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] \leq \frac{f\left(\lambda, n\right) + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}\left(\rho,\pi\right)}{\lambda} \text{ for all } \rho \in \mathcal{P}_\pi(\Theta) \text{ simultaneously}\right\} \\
&\geq 1 - \epsilon. \qquad (49)
\end{aligned}$$

■

The proof of Lemma 6 below will utilize the following two lemmas.

Lemma 16 *Let X be a random variable with $EX = 0$ such that for some constant $K > 0$, the MGF of X^2 satisfies*

$$E\exp\left(\lambda^2 X^2\right) \leq \exp\left(K^2\lambda^2\right) \text{ for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K}. \qquad (50)$$

Then

$$E\exp\left(\lambda X\right) \leq \exp\left(K^2\lambda^2\right) \text{ for all } \lambda \in \mathbb{R}.$$

Proof of Lemma 16. This follows from the proof of Proposition 2.5.2 in Vershynin (2018), pages 22-23. ■

Lemma 17 *Let X be any random variable taking values in $[0,1]$ with $EX = \mu$. Denote $\mathbf{X} = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are iid realizations of X. Let $\mathbf{X}' = (X_1', \ldots, X_n')$ where $X_1', \ldots, X_n'$ are iid realizations of a Bernoulli random variable X' with probability of success μ. If $f : [0,1]^n \to \mathbb{R}$ is convex, then*

$$E\left[f\left(\mathbf{X}\right)\right] \leq E\left[f\left(\mathbf{X}'\right)\right]$$

Proof of Lemma 17. This lemma is due to Maurer (2004). Another proof with more details is given in Germain et al. (2015); see Lemmas 51 and 52 there. For intuition, we can regard $\mathbf{X}'$ as a mean-preserving spread of $\mathbf{X}$ and $-f$ as the utility function. Then the lemma says that $\mathbf{X}$ is preferred by an expected utility maximizer having concave utility $-f\left(\cdot\right)$. ■

Proof of Lemma 6.

Part (a). Let $(X,Y) \sim P(X,Y)$ and let $\mu_\psi = E\psi(X,Y) < \infty$ where finiteness follows from Assumption 2 (iv). Recall that under Assumption 2 (iv), there exists a constant $K_\psi > 0$ such that

$$E\exp\left\{\lambda^2\psi(X,Y)^2\right\} \leq \exp\left(K_\psi^2\lambda^2\right) \text{ for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_\psi}. \tag{51}$$

Now for either $s \in \{-1,1\}$ and any $\theta \in \Theta$, consider

$$s\left[E\ell\left(\theta,Y,X\right) - \ell\left(\theta,Y,X\right)\right] = s\left[E\left(\psi(X,Y)1\left\{Y \neq a(X,\theta)\right\}\right) - \psi(X,Y)1\left\{Y \neq a(X,\theta)\right\}\right]. \tag{52}$$

Recall that $\psi(X,Y) > 0$. Using $(a-b)^2 \leq a^2 + b^2$ for $a > 0$ and $b > 0$, we have

$$\begin{aligned}
&E\exp\left\{\lambda^2\left(s\left[E\left\{\psi(X,Y)1\left\{Y \neq a(X,\theta)\right\}\right\} - \psi(X,Y)1\left\{Y \neq a(X,\theta)\right\}\right]\right)^2\right\} \\
&\leq E\exp\left\{\lambda^2\psi(X,Y)^2 + \lambda^2\left(E\left\{\psi(X,Y)\right\}\right)^2\right\}.
\end{aligned}$$

Additionally,

$$E\exp\left\{\lambda^2\psi(X,Y)^2 + \lambda^2\mu_\psi^2\right\} \leq \exp\left(\lambda^2\left[K_\psi^2 + \mu_\psi^2\right]\right)$$

for any λ such that $|\lambda| \leq 1/K_\psi$, which follows from (51). Seeing as $1/(K_\psi^2 + \mu_\psi^2)^{1/2} < 1/K_\psi$, the following condition holds

$$E\exp\left\{\lambda^2\left(s\left[E\ell\left(\theta,Y,X\right) - \ell\left(\theta,Y,X\right)\right]\right)^2\right\} \leq \exp\left(\lambda^2\left[K_\psi^2 + \mu_\psi^2\right]\right),$$

for all λ such that $|\lambda| \leq 1/(K_\psi^2 + \mu_\psi^2)^{1/2}$. As the expression in (52) has mean zero, Lemma 16 yields that

$$E\exp\left\{\lambda\left(s\left[E\ell\left(\theta,Y,X\right) - \ell\left(\theta,Y,X\right)\right]\right)\right\} \leq \exp\left(\left[K_\psi^2 + \mu_\psi^2\right]\lambda^2\right) \text{ for all } \lambda \in \mathbb{R}. \tag{53}$$

Applying (53),

$$\begin{aligned} E\exp\left\{\lambda D\left[R\left(\theta\right),R_n\left(\theta\right)\right]\right\} &= E\exp\left\{\lambda s\left[R\left(\theta\right)-R_n\left(\theta\right)\right]\right\} \\ &= E\exp\left\{\sum_{i=1}^{n}\left[\frac{\lambda}{n}\left(s\left[E\ell\left(\theta,Y_i,X_i\right)-\ell\left(\theta,Y_i,X_i\right)\right]\right)\right]\right\} \\ &= \prod_{i=1}^{n} E\exp\left\{\frac{\lambda}{n}\left(s\left[E\ell\left(\theta,Y_i,X_i\right)-\ell\left(\theta,Y_i,X_i\right)\right]\right)\right\} \\ &\leq \prod_{i=1}^{n}\exp\left(\frac{\lambda^2\left[K_\psi^2+\mu_\psi^2\right]}{n^2}\right)=\exp\left(\frac{\lambda^2\left[K_\psi^2+\mu_\psi^2\right]}{n}\right). \end{aligned}$$

Taking an integral with respect to π yields

$$\int_\Theta E\exp\left\{\lambda D\left[R\left(\theta\right),R_n\left(\theta\right)\right]\right\}d\pi(\theta)\leq\exp\left(\frac{\lambda^2\left[K_\psi^2+\mu_\psi^2\right]}{n}\right),$$

implying the first expression for $f(\lambda,n)$.

To derive the second expression for $f(\lambda,n)$ in the case that the utility function is bounded, note that $\psi(x,y)=U(1,1,x)-U(-1,1,x)$ when $y=1$ and $\psi(x,y)=U(-1,-1,x)-U(1,-1,x)$ when $y=-1$. When

$$U_{\max}=\sup_{a,y,x}|U(a,y,x)|<\infty,$$

it follows that $0\leq\ell(\theta,y,x)=\psi(x,y)1\{y\neq\text{sign}[m(x,\theta)-c(x)]\}<2U_{\max}$ under Assumption 1. By Hoeffding's lemma (see, for example, Massart and Picard (2007), page 21), with $s=-1$, for any $\theta\in\Theta$ we have

$$\begin{aligned} E\exp\left(\lambda\left[R_n\left(\theta\right)-R\left(\theta\right)\right]\right) &= E\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left[\ell(\theta,Y_i,X_i)-E\ell(\theta,Y_i,X_i)\right]\right) \\ &= \prod_{i=1}^{n}E\exp\left\{\frac{\lambda}{n}\left[\ell(\theta,Y_i,X_i)-E\ell(\theta,Y_i,X_i)\right]\right\} \\ &\leq \prod_{i=1}^{n}\exp\left(\frac{\lambda^2U_{\max}^2}{2n^2}\right)=\exp\left(\frac{\lambda^2U_{\max}^2}{2n}\right). \end{aligned} \tag{54}$$

Nearly identical steps in the $s=1$ case, now with Hoeffding's lemma applied to $-\ell(\theta,Y_i,X_i)$, $i=1,\ldots,n$, produce that

$$E\exp\left(\lambda\left[R\left(\theta\right)-R_n\left(\theta\right)\right]\right)\leq\exp\left(\frac{\lambda^2U_{\max}^2}{2n}\right) \tag{55}$$

Integrating with respect to π, (54) and (55) yield that

$$\int_\Theta E\exp\left(\lambda s\left[R\left(\theta\right)-R_n\left(\theta\right)\right]\right)d\pi\left(\theta\right)\leq\exp\left(\frac{\lambda^2U_{\max}^2}{2n}\right),\ s\in\{-1,1\}.$$

This demonstrates that (25) holds with $f\left(\lambda,n\right)=\frac{\lambda^2U_{\max}^2}{2n}$ in the bounded utility setting.

Part (b). Again note that when the utility function is bounded by $U_{\max}$ we have $0 \leq \ell(\theta, y, x) < 2U_{\max}$ under Assumption 1. Therefore $\ell(\theta, y, x)/(2U_{\max}) \in [0, 1]$. Set

$$\mathbf{X} = \left(\frac{\ell(\theta, Y_1, X_1)}{2U_{\max}}, \ldots, \frac{\ell(\theta, Y_n, X_n)}{2U_{\max}} \right),$$

and note that for any $\theta \in \Theta$,

$$\begin{aligned} \exp\{\lambda D(R(\theta), R_n(\theta))\} &= \exp\left[\lambda \mathcal{F}(R(\theta)) - 2U_{\max}\lambda \cdot \frac{R_n(\theta)}{2U_{\max}} \right] \\ &= \exp\left\{ \lambda \mathcal{F}(R(\theta)) - \frac{2U_{\max}\lambda}{n} \sum_{i=1}^{n} \frac{\ell(\theta, Y_i, X_i)}{2U_{\max}} \right\} \end{aligned} \tag{56}$$

is a convex mapping of $\mathbf{X}$. By (Maurer's) Lemma 17,

$$E\exp\left\{ \lambda \mathcal{F}(R(\theta)) - \frac{2U_{\max}\lambda}{n} \sum_{i=1}^{n} \frac{\ell(\theta, Y_i, X_i)}{2U_{\max}} \right\} \leq E\exp\left\{ \lambda \mathcal{F}(R(\theta)) - \frac{2U_{\max}\lambda}{n} \sum_{i=1}^{n} X_i' \right\}, \tag{57}$$

where $X_1', \ldots, X_n'$ are iid Bernoulli random variables with success probability $R(\theta)/(2U_{\max}) \in [0, 1]$. From here we can continue as in the proof of Corollary 2.2 in Germain et al. (2009). We have for any $\theta \in \Theta$,

$$\begin{aligned} &E\exp\left\{ \lambda \mathcal{F}(R(\theta)) - \frac{2U_{\max}\lambda}{n} \sum_{i=1}^{n} X_i' \right\} \\ &= \exp\{\lambda \mathcal{F}(R(\theta))\} E\exp\left\{ -\frac{2U_{\max}\lambda}{n} \sum_{i=1}^{n} X_i' \right\} \\ &= \exp\{\lambda \mathcal{F}(R(\theta))\} \sum_{k=1}^{n} \Pr\left(\sum_{i=1}^{n} X_i' = k \right) \exp\left(-\frac{2U_{\max}\lambda}{n} k \right) \\ &= \exp\{\lambda \mathcal{F}(R(\theta))\} \sum_{k=1}^{n} \binom{n}{k} \left(\frac{R(\theta)}{2U_{\max}} \right)^k \left(1 - \frac{R(\theta)}{2U_{\max}} \right)^{n-k} \left[\exp\left(-\frac{2U_{\max}\lambda}{n} \right) \right]^k \\ &= \exp\{\lambda \mathcal{F}(R(\theta))\} \left[\left(\frac{R(\theta)}{2U_{\max}} \right) \exp\left(-\frac{2U_{\max}\lambda}{n} \right) + \left(1 - \frac{R(\theta)}{2U_{\max}} \right) \right]^n \\ &= \exp\{\lambda \mathcal{F}(R(\theta))\} \left\{ 1 - \left(\frac{R(\theta)}{2U_{\max}} \right) \left[1 - \exp\left(-\frac{2U_{\max}\lambda}{n} \right) \right] \right\}^n, \end{aligned}$$

where the second to last equality is from the binomial theorem. Now, noting that

$$\exp\{\lambda \mathcal{F}(R(\theta))\} = \left\{ 1 - \left(\frac{R(\theta)}{2U_{\max}} \right) \left[1 - \exp\left(-\frac{2U_{\max}\lambda}{n} \right) \right] \right\}^{-n},$$

we have

$$E\exp\left\{ \lambda \mathcal{F}(R(\theta)) - \frac{2U_{\max}\lambda}{n} \sum_{i=1}^{n} X_i' \right\} = 1. \tag{58}$$

Combining equations (56), (57), and (58), we have

$$E\exp\{\lambda [\mathcal{F}(R(\theta)) - R_n(\theta)]\} \leq 1, \tag{59}$$

and so equation (25) holds with $f(\lambda, n) = 0$.

Part (c). Let $\theta \in \Theta$. If $R(\theta) - \lambda U_{\max}^2/(2n) \geq \mathcal{F}(R(\theta))$, which is not random, then clearly

$$D(R(\theta), R_n(\theta)) = R(\theta) - \lambda U_{\max}^2/(2n) - R_n(\theta).$$

In this case,

$$\begin{aligned} E\exp\left(\lambda D[R(\theta), R_n(\theta)]\right) &= E\exp\left(\lambda\left[R(\theta) - \frac{\lambda U_{\max}^2}{2n} - R_n(\theta)\right]\right) \\ &= \exp\left(-\frac{\lambda^2 U_{\max}^2}{2n}\right) E\exp\left(\lambda\left[R(\theta) - R_n(\theta)\right]\right) \\ &\leq \exp\left(-\frac{\lambda^2 U_{\max}^2}{2n}\right)\exp\left(\frac{\lambda^2 U_{\max}^2}{2n}\right) = 1 \end{aligned} \tag{60}$$

where the inequality follows from (55). Alternatively, in the case that $R(\theta) - \lambda U_{\max}^2/(2n) < \mathcal{F}(R(\theta))$, we have

$$D[R(\theta), R_n(\theta)] = \mathcal{F}(R(\theta)) - R_n(\theta).$$

Then, by (59),

$$E\exp\left(\lambda D[R(\theta), R_n(\theta)]\right) = E\exp\left(\lambda\left[\mathcal{F}\left(R\left(\theta\right)\right) - R_n\left(\theta\right)\right]\right) \leq 1. \tag{61}$$

Integrating over Θ with respect to π, it follows from (60) and (61) that

$$\int_\Theta E\exp\left(\lambda D\left[R\left(\theta\right), R_n\left(\theta\right)\right]\right) d\pi(\theta) \leq 1,$$

so condition (25) holds with $f(\lambda, n) = 0$. ■

Proof of Theorem 7.

Part (a) follows directly from Theorem 5 and Lemma 6 (a) with D as in Lemma 6 (a).

Part (b). Let D be as specified in Lemma 6 (b). It is straightforward to verifty that D is convex. Note that

$$D(r_1, r_2) = \mathcal{F}_{\lambda,n}(r_1) - r_2 \leq \mathfrak{d}$$

for any $\mathfrak{d} \in \mathbb{R}$ if and only if

$$1 - \left(\frac{r_1}{2U_{\max}}\right)\left[1 - \exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right] \geq \exp\left[-\frac{\lambda}{n}\left(r_2 + \mathfrak{d}\right)\right].$$

The latter is equivalent to

$$\begin{aligned} r_1 &\leq \frac{2U_{\max}}{1 - \exp\left(-2U_{\max}\lambda/n\right)}\left\{1 - \exp\left[-\frac{\lambda}{n}\left(r_2 + \mathfrak{d}\right)\right]\right\} \\ &:= \mathcal{F}_{\lambda,n}^{-1}(r_2 + \mathfrak{d}). \end{aligned}$$

Setting $r_1 = \int_\Theta R\left(\theta\right) d\rho(\theta)$, $r_2 = \int_\Theta R_n\left(\theta\right) d\rho(\theta)$ and $\mathfrak{d} = \frac{1}{\lambda}\left[\log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\rho, \pi)\right]$ and using Theorem 5 and Lemma 6 (b) yields the desired result.

Part (c). Now let D be as specified in Lemma 6 (c). That D is convex follows from the convexity of D specified in part (a) plus a constant, the convexity of D specified in part (b), and

the fact that the maximum of two convex functions is convex. Theorem 5 combined with Lemma 6 (c) yields that

$$\Pr\left\{\max\left[\int_{\Theta} R\left(\theta\right) d\rho(\theta) - \frac{\lambda U_{\max}^2}{2n} - \int_{\Theta} R_n\left(\theta\right) d\rho(\theta),\ \mathcal{F}\left(\int_{\Theta} R\left(\theta\right) d\rho(\theta)\right) - \int_{\Theta} R_n\left(\theta\right) d\rho(\theta)\right]\right.$$
$$\left. \leq \frac{D_{\mathrm{KL}}\left(\rho,\pi\right) + \log\frac{1}{\epsilon}}{\lambda} \ \textit{for all } \rho \in \mathcal{P}_{\pi}\left(\Theta\right) \ \textit{simultaneously}\right\} \geq 1-\epsilon \tag{62}$$

Now, observe that

$$\max\left[\int_{\Theta} R\left(\theta\right) d\rho(\theta) - \frac{\lambda U_{\max}^2}{2n} - \int_{\Theta} R_n\left(\theta\right) d\rho(\theta),\ \mathcal{F}\left(\int_{\Theta} R\left(\theta\right) d\rho(\theta)\right) - \int_{\Theta} R_n\left(\theta\right) d\rho(\theta)\right]$$
$$\leq \frac{D_{\mathrm{KL}}\left(\rho,\pi\right) + \log\frac{1}{\epsilon}}{\lambda}$$

holds if and only if

$$\int_{\Theta} R\left(\theta\right) d\rho(\theta) \leq \int_{\Theta} R_n(\theta) d\rho(\theta) + \frac{1}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n} + D_{\mathrm{KL}}\left(\rho,\pi\right) + \log\frac{1}{\epsilon}\right] = U_{\lambda,\pi,\rho}(\epsilon),$$

and

$$\int_{\Theta} R\left(\theta\right) d\rho(\theta) \leq \mathcal{F}_{n,\lambda}^{-1}\left(\int_{\Theta} R_n\left(\theta\right) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\log\frac{1}{\epsilon}\right) = U_{\lambda,\pi,\rho}^{\mathcal{F}}(\epsilon)$$

hold simultaneously. Additionally, the two inequalities directly above hold simultaneously if and only if

$$\int_{\Theta} R\left(\theta\right) d\rho(\theta) \leq \min\left\{U_{\lambda,\pi,\rho}(\epsilon), U_{\lambda,\pi,\rho}^{\mathcal{F}}(\epsilon)\right\}.$$

Therefore,

$$\left\{\max\left[\int_{\Theta} R\left(\theta\right) d\rho(\theta) - \frac{\lambda U_{\max}^2}{2n} - \int_{\Theta} R_n\left(\theta\right) d\rho(\theta),\ \mathcal{F}\left(\int_{\Theta} R\left(\theta\right) d\rho(\theta)\right) - \int_{\Theta} R_n\left(\theta\right) d\rho(\theta)\right]\right.$$
$$\left. \leq \frac{D_{\mathrm{KL}}\left(\rho,\pi\right) + \log\frac{1}{\epsilon}}{\lambda} \ \textit{for all } \rho \in \mathcal{P}_{\pi}\left(\Theta\right) \ \textit{simultaneously}\right\}$$
$$= \left\{\int_{\Theta} R\left(\theta\right) d\rho(\theta) \leq \min\left\{U_{\lambda,\pi,\rho}(\epsilon), U_{\lambda,\pi,\rho}^{\mathcal{F}}(\epsilon)\right\} \ \textit{for all } \rho \in \mathcal{P}_{\pi}\left(\Theta\right) \ \textit{simultaneously}\right\}. \tag{63}$$

Combined, (62) and (63) imply the result of Theorem 7 (c). ■

Proof of Theorem 8. Part (a). This part follows directly from Theorem 7 with $s = 1$ and $\rho = \hat{\rho}_{\lambda}$.

Part (b). Define the events $\mathcal{E}_1$ and $\mathcal{E}_2$:

$$\mathcal{E}_1 = \left\{\int_{\Theta} R\left(\theta\right) d\hat{\rho}_{\lambda}\left(\theta\right) \leq \int_{\Theta} R_n\left(\theta\right) d\hat{\rho}_{\lambda}\left(\theta\right) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda},\pi\right) + \frac{\lambda^2\left(K_{\psi}^2 + \mu_{\psi}^2\right)}{n} + \log\frac{2}{\epsilon}\right]\right\},$$

$$\mathcal{E}_2 = \left\{\int_{\Theta} R_n\left(\theta\right) d\hat{\rho}_{\lambda}\left(\theta\right) \leq \int_{\Theta} R\left(\theta\right) d\hat{\rho}_{\lambda}\left(\theta\right) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda},\pi\right) + \frac{\lambda^2\left(K_{\psi}^2 + \mu_{\psi}^2\right)}{n} + \log\frac{2}{\epsilon}\right]\right\}.$$

Then

$$\Pr(\mathcal{E}_1) \geq 1 - \frac{\epsilon}{2} \text{ and } \Pr(\mathcal{E}_2) \geq 1 - \frac{\epsilon}{2}.$$

So

$$\begin{aligned}\Pr(\mathcal{E}_1 \cap \mathcal{E}_2) &= 1 - \Pr(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \geq 1 - \Pr(\mathcal{E}_1^c) - \Pr(\mathcal{E}_2^c) \\ &\geq 1 - \frac{\epsilon}{2} - \frac{\epsilon}{2} = 1 - \epsilon.\end{aligned}$$

But, the event given in Part (b) is just $\mathcal{E}_1 \cap \mathcal{E}_2$. Hence, the inequality in Part (b) holds with probability at least $1 - \epsilon$.

Part (c). By the definition of $\hat{\rho}_\lambda$, we have

$$\int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_\lambda, \pi) \leq \int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi)$$

for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously. Hence, by part (a), with probability at least $1 - \epsilon/2$:

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho, \pi) + \frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right]$$

for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously. Using Theorem 7 (a) now with $s = -1$, we have, with probability at least $1 - \epsilon/2$,

$$\int_\Theta R_n(\theta)\, d\rho(\theta) \leq \int_\Theta R(\theta)\, d\rho(\theta) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho, \pi) + \frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right].$$

Therefore, with probability at least $1 - \epsilon$,

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R(\theta)\, d\rho(\theta) + \frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \frac{2}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right] \tag{64}$$

for all $\rho \in \mathcal{P}_\pi(\Theta)$ simultaneously. Hence, with probability at least $1 - \epsilon$,

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \sup_{\rho \in \mathcal{P}_\pi(\Theta)} \left[\int_\Theta R(\theta)\, d\rho(\theta) + \frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi)\right] + \frac{2}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\epsilon}\right].$$

In the case where $U_{\max} < \infty$, we can follow the same steps based off Theorem 7 but with $(K_\psi^2 + \mu_\psi^2)$ replaced by $U_{\max}^2/2$.

Part (d). This follows directly from Theorem 7(c) with $\rho = \hat{\rho}_\lambda$. ■

Proof of Theorem 9.

Part (a). Let $s \in \{0, 1\}$. For any $\rho \in \mathcal{P}(\Theta)_\pi$, including sample dependent ρ, let

$$B_n(\lambda_1, \lambda_2, z; \rho, \pi) = \frac{1}{\lambda_1}\left[\frac{\lambda_2^2 U_{\max}^2}{2n} + \log z + D_{\mathrm{KL}}(\rho, \pi)\right],$$

and define the event

$$\mathcal{E}_n\left(\lambda_1, \lambda_2, z; \rho, \pi\right) = \left\{ \int_\Theta s\left[R(\theta) - R_n(\theta)\right] d\rho(\theta) > B_n\left(\lambda_1, \lambda_2, z; \rho, \pi\right) \right\}.$$

Note that by Theorem 7(a), $\Pr(\mathcal{E}_n\left(\lambda, \lambda, 1/\epsilon; \rho, \pi\right)) \leq \epsilon$ for any $\lambda > 0$. Additionally, holding the other arguments constant, $B_n\left(\lambda_1, \lambda_2, z; \rho, \pi\right)$ is decreasing in λ_1, increasing in λ_2, and increasing in z. Hence

$$\begin{aligned}
\mathcal{E}_n\left(\lambda_1, \lambda_2, z; \rho, \pi\right) &\subseteq \mathcal{E}_n\left(\tilde{\lambda}_1, \lambda_2, z; \rho, \pi\right) \text{ for } \tilde{\lambda}_1 \geq \lambda_1, \\
\mathcal{E}_n\left(\lambda_1, \lambda_2, z; \rho, \pi\right) &\subseteq \mathcal{E}_n\left(\lambda_1, \tilde{\lambda}_2, z; \rho, \pi\right) \text{ for } \tilde{\lambda}_2 \leq \lambda_2, \\
\mathcal{E}_n\left(\lambda_1, \lambda_2, z; \rho, \pi\right) &\subseteq \mathcal{E}_n\left(\lambda_1, \lambda_2, \tilde{z}; \rho, \pi\right) \text{ for } \tilde{z} \leq z.
\end{aligned}$$

Now, fix $\alpha > 1$. With some abuse of notation, the event of interest in Part (a) is the complement of the event $\mathcal{E}_n\left(\alpha; \rho, \pi\right)$ defined by

$$\mathcal{E}_n\left(\alpha; \rho, \pi\right) := \left\{ \int_\Theta s\left[R(\theta) - R_n(\theta)\right] d\rho(\theta) > \inf_{\lambda > 1} \left\{ B_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right) \right\} \right\}.$$

Note that

$$\mathcal{E}_n\left(\alpha; \rho, \pi\right) = \bigcup_{\lambda > 1} \mathcal{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right).$$

But

$$\bigcup_{\lambda > 1} \mathcal{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right) \subseteq \bigcup_{k=0}^{\infty} \bigcup_{\lambda \in (\alpha^k, \alpha^{k+1}]} \mathcal{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right),$$

and for all $\lambda \in (\alpha^k, \alpha^{k+1}]$ it holds that

$$\mathcal{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right) \subseteq \mathcal{E}_n\left(\frac{\alpha^{k+1}}{\alpha}, \alpha^k, \frac{1}{\epsilon}\left(\frac{\log\left(\alpha^2 \cdot \alpha^k\right)}{\log \alpha}\right)^2; \rho, \pi\right).$$

Hence

$$\bigcup_{\lambda \in (\alpha^k, \alpha^{k+1}]} \mathcal{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right) \subseteq \mathcal{E}_n\left(\frac{\alpha^{k+1}}{\alpha}, \alpha^k, \frac{1}{\epsilon}\left(\frac{\log\left(\alpha^2 \cdot \alpha^k\right)}{\log \alpha}\right)^2; \rho, \pi\right),$$

and

$$\begin{aligned}
\Pr\left(\mathcal{E}_n\left(\alpha; \rho, \pi\right)\right) &\leq \sum_{k=0}^{\infty} \Pr\left[\bigcup_{\lambda \in (\alpha^k, \alpha^{k+1}]} \mathcal{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\epsilon}\left(\frac{\log \alpha^2 \lambda}{\log \alpha}\right)^2; \rho, \pi\right) \right] \\
&\leq \sum_{k=0}^{\infty} \Pr\left[\mathcal{E}_n\left(\frac{\alpha^{k+1}}{\alpha}, \alpha^k, \frac{1}{\epsilon}\left(\frac{\log\left(\alpha^2 \cdot \alpha^k\right)}{\log \alpha}\right)^2; \rho, \pi\right) \right] \\
&= \sum_{k=0}^{\infty} \Pr\left[\mathcal{E}_n\left(\alpha^k, \alpha^k, \frac{(k+2)^2}{\epsilon}; \rho, \pi\right) \right] \\
&\leq \sum_{k=0}^{\infty} \frac{\epsilon}{(k+2)^2} = \left(\frac{1}{6}\pi^2 - 1\right)\epsilon < \epsilon,
\end{aligned}$$

where last inequality follows from Theorem 7(a). Therefore,

$$\Pr\left(\mathcal{E}_n\left(\alpha;\rho,\pi\right)^c\right) \geq 1-\epsilon$$

This is the statement for Part (a).

Part (b). Applying Part (a) with $\rho = \hat{\rho}_{\tilde{\lambda}}$, the following event holds with probability probability $1-\epsilon$

$$\int_\Theta s\left[R(\theta) - R_n(\theta)\right] d\hat{\rho}_{\tilde{\lambda}}(\theta) \leq \inf_{\lambda>1}\left\{\frac{\alpha}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi) + 2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right]\right\}.$$

Then, Part (b) follows from the above and the observation that, for $\tilde{\lambda} > 1$, it holds that

$$\begin{aligned}
&\inf_{\lambda>1}\left\{\frac{\alpha}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi) + 2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right]\right\} \\
&\qquad \leq \frac{\alpha}{\tilde{\lambda}}\left[\frac{\tilde{\lambda}^2 U_{\max}^2}{2n} + \log\frac{1}{\epsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi) + 2\log\frac{\log\left(\alpha^2\tilde{\lambda}\right)}{\log\alpha}\right].
\end{aligned}$$

Part (c). We proceed similarly to part (a). For any $\rho \in \mathcal{P}_\pi(\Theta)$ that may be sample dependent, define

$$\overline{B}_n\left(\lambda_1,\lambda_2,z;\rho,\pi\right) = \int_\Theta R_n(\theta)d\rho(\theta) + \frac{1}{\lambda_1}\left[\frac{\lambda_2^2 U_{\max}^2}{2n} + \log z + D_{\mathrm{KL}}(\rho,\pi)\right],$$

and

$$\overline{B}_n^{\mathcal{F}}\left(\lambda_1,\lambda_2,z;\rho,\pi\right) = \frac{2U_{\max}}{1-\exp\left(-\frac{2\lambda_1 U_{\max}}{n}\right)}\left\{1-\exp\left[-\frac{\lambda_2}{n}\int_\Theta R_n d\rho(\theta) - \frac{1}{n}\log z - \frac{1}{n}D_{\mathrm{KL}}\left(\rho,\pi\right)\right]\right\}.$$

Note that

$$\overline{B}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right);\rho,\pi\right) = \overline{U}_{\lambda,\pi,\rho,\alpha}(\epsilon),$$

and

$$\begin{aligned}
&\overline{B}_n^{\mathcal{F}}\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right) \\
&= \frac{2U_{\max}}{1-\exp\left(-\frac{2\lambda U_{\max}}{\alpha n}\right)}\left\{1-\exp\left[-\frac{\lambda}{n}\int_\Theta R_n d\rho(\theta) - \frac{1}{n}\log\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2 - \frac{1}{n}D_{\mathrm{KL}}\left(\rho,\pi\right)\right]\right\} \\
&= \frac{2U_{\max}}{1-\exp\left(-\frac{2\lambda U_{\max}}{\alpha n}\right)}\left\{1-\exp\left[-\frac{\lambda}{n}\left(\int_\Theta R_n d\rho(\theta) + \frac{1}{\lambda}\log\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2 + \frac{1}{\lambda}D_{\mathrm{KL}}\left(\rho,\pi\right)\right)\right]\right\} \\
&= \mathcal{F}_{n,\lambda,\alpha}^{-1}\left(\int_\Theta R_n d\rho(\theta) + \frac{1}{\lambda}\log\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2 + \frac{1}{\lambda}D_{\mathrm{KL}}\left(\rho,\pi\right)\right) = \overline{U}_{\lambda,\pi,\rho,\alpha}^{\mathcal{F}}(\epsilon).
\end{aligned}$$

Now, holding the other arguments constant, notice that $\min[\overline{B}_n\left(\lambda_1,\lambda_2,z;\rho,\pi\right), \overline{B}_n^{\mathcal{F}}\left(\lambda_1,\lambda_2,z;\rho,\pi\right)]$ is decreasing in λ_1, increasing in λ_2, and increasing in z as both $\overline{B}_n\left(\lambda_1,\lambda_2,z;\rho,\pi\right)$ and $\overline{B}_n^{\mathcal{F}}\left(\lambda_1,\lambda_2,z;\rho,\pi\right)$ have these properties.

With some abuse of notation, define the two events:

$$\overline{\mathcal{E}}_n(\lambda_1,\lambda_2,z;\rho,\pi)=\left\{\int_\Theta R(\theta)d\rho(\theta)>\min\left[\overline{B}_n\left(\lambda_1,\lambda_2,z;\rho,\pi\right),\overline{B}_n^{\mathcal{F}}\left(\lambda_1,\lambda_2,z;\rho,\pi\right)\right]\right\},$$

$$\overline{\mathcal{E}}_n(\alpha;\rho,\pi)=\left\{\int_\Theta R(\theta)d\rho(\theta)>\inf_{\lambda>1}\left\{\min\left[\overline{B}_n\left(\lambda_1,\lambda_2,z;\rho,\pi\right),\overline{B}_n^{\mathcal{F}}\left(\lambda_1,\lambda_2,z;\rho,\pi\right)\right]\right\}\right\}$$

and notice that by Theorem 7 (c), $\Pr(\overline{\mathcal{E}}_n(\lambda,\lambda,1/\epsilon;\rho,\pi))\leq\epsilon$ for any $\lambda>0$.

The event of interest in Part (c) is the complement of $\overline{\mathcal{E}}(\alpha;\rho,\pi)$. Now, we have

$$\begin{aligned}\overline{\mathcal{E}}_n(\alpha;\rho,\pi)&=\bigcup_{\lambda>1}\overline{\mathcal{E}}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right)\\&\subseteq\bigcup_{k=0}^{\infty}\bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]}\overline{\mathcal{E}}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right).\end{aligned}$$

Hence, following arguments similar to those in part (a),

$$\begin{aligned}\Pr\left(\overline{\mathcal{E}}_n\left(\alpha;\rho,\pi\right)\right)&\leq\sum_{k=0}^{\infty}\Pr\left[\bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]}\overline{\mathcal{E}}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\epsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right)\right]\\&\leq\sum_{k=0}^{\infty}\Pr\left[\overline{\mathcal{E}}_n\left(\frac{\alpha^{k+1}}{\alpha},\alpha^k,\frac{1}{\epsilon}\left(\frac{\log\left(\alpha^2\cdot\alpha^k\right)}{\log\alpha}\right)^2;\rho,\pi\right)\right]\\&=\sum_{k=0}^{\infty}\Pr\left[\overline{\mathcal{E}}_n\left(\alpha^k,\alpha^k,\frac{(k+2)^2}{\epsilon};\rho,\pi\right)\right]\\&<\epsilon,\end{aligned}$$

It follows that $\Pr\left(\tilde{\mathcal{E}}_n\left(\alpha;\rho,\pi\right)^c\right)\geq1-\epsilon$, which is the statement of interest for part (c).

Part (d) follows from Part (c) via steps parallel to those in the proof of Part (b). ■

Proof of Lemma 10. We will show that for all $\theta\in\Theta$,

$$E\exp\left\{\lambda D\left(\frac{R(\theta)}{2U_{\max}},\frac{R_n(\theta)}{2U_{\max}}\right)\right\}=E\exp\left\{n\cdot\mathrm{kl}\left(\frac{R_n(\theta)}{2U_{\max}},\frac{R(\theta)}{2U_{\max}}\right)\right\}\leq\xi(n).\tag{65}$$

Then the result follows from integrating over Θ with respect to π.

First consider any θ such that $R(\theta)=0$ or $R(\theta)=2U_{\max}$. Recall

$$R(\theta)=E\psi(X,Y)1\{a_\theta(X)\neq Y\},$$

$\psi(X,Y)$ can be written $\psi(X,Y)=U(Y,Y,X)-U(-Y,Y,X)\leq2U_{\max}$, and $\psi(X,Y)>0$ by Assumption 1. If $R(\theta)=0$ then it follows that we must have $\Pr(a_\theta(X)=Y)=1$ and hence $1\{a_\theta(X_i)\neq Y_i\}=0$ for $i=1,\ldots,n$ (a.s.). Hence $R_n(\theta)=0$ in this case (a.s.), so that (65) holds. If $R(\theta)=2U_{\max}$, it follows that we must have $\Pr(\psi(X,Y)=2U_{\max})=1$ and $\Pr(a_\theta(X)=Y)=0$, so that now $R_n(\theta)=2U_{\max}$ (a.s.) and again (65) holds.

When θ is such that $R(\theta) \notin \{0, 2U_{\max}\}$, the proof follows that in Theorem 1 of Maurer (2004) or Lemma 19 in Germain et al. (2015) with minor adjustments. Note that

$$\exp\left\{\lambda D\left(\frac{R(\theta)}{2U_{\max}}, \frac{R_n(\theta)}{2U_{\max}}\right)\right\} = \exp\left\{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\ell(\theta, Y_i, X_i)}{2U_{\max}}, \frac{R(\theta)}{2U_{\max}}\right)\right\}$$

is a convex function of $\mathbf{X} = (\ell(\theta, Y_1, X_1)/2U_{\max}, \ldots, \ell(\theta, Y_n, X_n)/2U_{\max})$ and $\ell(\theta, x, y)/2U_{\max} \in [0, 1]$. Then, by Lemma 17,

$$E\exp\left\{\lambda D\left(\frac{R(\theta)}{2U_{\max}}, \frac{R_n(\theta)}{2U_{\max}}\right)\right\} \leq E\exp\left\{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n}X_i', \frac{R(\theta)}{2U_{\max}}\right)\right\} \tag{66}$$

where $X_1', \ldots, X_n'$ are iid Bernoulli random variables with success probability $R(\theta)/(2U_{\max})$. Denoting $X' = \sum_{i=1}^{n} X_i'$,

$$\begin{aligned}
&E\exp\left\{n \cdot \mathrm{kl}\left(\frac{1}{n}X', \frac{R(\theta)}{2U_{\max}}\right)\right\} \\
&\quad= E\left(\frac{\frac{1}{n}X'}{\frac{R(\theta)}{2U_{\max}}}\right)^{X'}\left(\frac{1-\frac{1}{n}X'}{1-\frac{R(\theta)}{2U_{\max}}}\right)^{n-X'} \\
&\quad= \sum_{k=0}^{n}\Pr\left(X'=k\right)\left(\frac{\frac{k}{n}}{\frac{R(\theta)}{2U_{\max}}}\right)^{k}\left(\frac{1-\frac{k}{n}}{1-\frac{R(\theta)}{2U_{\max}}}\right)^{n-k} \\
&\quad= \sum_{k=0}^{n}\binom{n}{k}\left(\frac{R(\theta)}{2U_{\max}}\right)^{k}\left(1-\frac{R(\theta)}{2U_{\max}}\right)^{n-k}\left(\frac{\frac{k}{n}}{\frac{R(\theta)}{2U_{\max}}}\right)^{k}\left(\frac{1-\frac{k}{n}}{1-\frac{R(\theta)}{2U_{\max}}}\right)^{n-k} \\
&\quad= \sum_{k=0}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^{k}\left(1-\frac{k}{n}\right)^{n-k} = \xi(n)
\end{aligned} \tag{67}$$

Therefore (65) holds for any $\theta \in \Theta$, completing the proof. ■

Proof of Corollary 12. We have

$$\begin{aligned}
&\int_{\Theta}\mathbb{U}_n(\theta)d\hat{\rho}\left(\theta\right) - \int_{\Theta}\mathbb{U}(\theta)d\hat{\rho}\left(\theta\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[U(Y_i, Y_i, X_i) - EU(Y_i, Y_i, X_i)\right] + \int_{\Theta}\left[R\left(\theta\right) - R_n\left(\theta\right)\right]d\hat{\rho}\left(\theta\right).
\end{aligned}$$

Using Hoeffding's inequality, we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left[U(Y_i, Y_i, X_i) - EU(Y_i, Y_i, X_i)\right] > U_{\max}\sqrt{\frac{2\log\frac{2}{\epsilon}}{n}}\right) \leq \frac{\epsilon}{2}.$$

Therefore,

$$\begin{aligned}
&\Pr\left(\int_{\Theta}\mathbb{U}_n(\theta)d\hat{\rho}\left(\theta\right)-\int_{\Theta}\mathbb{U}(\theta)d\hat{\rho}\left(\theta\right)>B_U+B_R\left(\hat{\rho}\right)\right)\\
&\leq\Pr\left\{\frac{1}{n}\sum_{i=1}^{n}\left[U(Y_i,Y_i,X_i)-EU(Y_i,Y_i,X_i)\right]>B_U\right\}\\
&+\Pr\left\{\int_{\Theta}\left[R\left(\theta\right)-R_n\left(\theta\right)\right]d\hat{\rho}\left(\theta\right)>B_R\left(\hat{\rho}\right)\right\}\\
&\leq\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon,
\end{aligned}$$

and the result follows. ■

7.3 Proofs for Section 3.2

Proof of Lemma 13. By definition and via simple calculations, we have

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\rho,\pi\right)\\
&=-\frac{1}{2}E_{\theta\sim\rho}\left[\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}+\left(\theta-\mu_\rho\right)'\Sigma_\rho^{-1}\left(\theta-\mu_\rho\right)-\left(\theta-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\theta-\mu_\pi\right)\right]\\
&=-\frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}-\frac{1}{2}\left[M-E_{\theta\sim\rho}\left(\theta-\mu_\rho+\mu_\rho-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\theta-\mu_\rho+\mu_\rho-\mu_\pi\right)\right]\\
&=-\frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}-\frac{1}{2}\left[M-tr\left(\Sigma_\rho\Sigma_\pi^{-1}\right)-\left(\mu_\rho-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\mu_\rho-\mu_\pi\right)\right]\\
&=\frac{1}{2}\left(\mu_\rho-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\mu_\rho-\mu_\pi\right)+\frac{1}{2}\left[tr\left(\Sigma_\rho\Sigma_\pi^{-1}\right)-M\right]-\frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}.
\end{aligned}$$

■

Proof of Lemma 14. We have

$$\begin{aligned}
&\int_{\Theta}R_n\left(\theta\right)d\rho\left(\theta\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{\theta\sim\rho}1\left\{Y_i\neq\mathrm{sign}\left[\phi(X_i)'\theta-c(X_i)\right]\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{\theta\sim\rho}1\left\{Y_i\left[\phi(X_i)'\theta-c\left(X_i\right)\right]\leq0\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{\theta\sim\rho}1\left\{\left[Y_i\phi(X_i)'\theta-Y_ic\left(X_i\right)\right]\leq0\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{Z\sim N(0,I_d)}1\left\{\left[Y_i\phi(X_i)'\left(\mu_\rho+\Sigma_\rho^{1/2}Z\right)-Y_ic\left(X_i\right)\right]\leq0\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)\mathrm{Pr}_{Z\sim N(0,I_d)}\left\{Y_i\phi(X_i)'\Sigma_\rho^{1/2}Z\leq Y_i\left[c\left(X_i\right)-X_i'\mu_\rho\right]\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)\Phi\left(\frac{Y_i\left[c\left(X_i\right)-\phi(X_i)_i'\mu_\rho\right]}{\sqrt{\phi(X_i)'\Sigma_\rho\phi(X_i)}}\right).
\end{aligned}$$

■

7.4 Proofs for Section 3.3

Proof of Lemma 15. The proof is essentially the same as that for Lemma 2, but we can be more explicit. By definition, we have

$$\begin{aligned}\frac{\rho_{A,\pi}(k)}{\pi(k)}\cdot\frac{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)} &= \frac{\nu_A(k)}{\sum_{j=1}^K \pi(j)\,\nu_A(j)}\cdot\frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\nu_A(k)}\\ &= \frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\sum_{j=1}^K \pi(j)\,\nu_A(j)}.\end{aligned}$$

Now, using the definition of the KL divergence, we have, for any $\rho\in\mathcal{P}_\pi(\Theta)$:

$$\begin{aligned}&D_{\mathrm{KL}}(\rho,\rho_{A,\pi})\\ &=\sum_{k=1}^K\left\{\int_{\Theta_{(k)}}\log\left[\frac{\rho(k)}{\rho_{A,\pi}(k)}\cdot\frac{d\rho\left(\theta_{(k)}|k\right)}{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}\right]d\rho\left(\theta_{(k)}|k\right)\right\}\rho(k)\\ &=\sum_{k=1}^K\left\{\int_{\Theta_{(k)}}\log\left\{\frac{\rho(k)}{\pi(k)}\cdot\frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\left[\frac{\rho_{A,\pi}(k)}{\pi(k)}\cdot\frac{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right]^{-1}\right\}d\rho\left(\theta_{(k)}|k\right)\right\}\rho(k)\\ &=\sum_{k=1}^K\left\{\int_{\Theta_{(k)}}\left[\log\left(\frac{\rho(k)}{\pi(k)}\cdot\frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right)-\log\left(\frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\sum_{j=1}^K\pi(j)\,\nu_A(j)}\right)\right]d\rho\left(\theta_{(k)}|k\right)\right\}\rho(k)\\ &=\sum_{k=1}^K\left\{\int_{\Theta_{(k)}}\log\left[\frac{\rho(k)}{\pi(k)}\cdot\frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right]d\rho\left(\theta_{(k)}|k\right)\right\}\rho(k)\\ &+\sum_{k=1}^K\left[\int_{\Theta_{(k)}}A\left(k,\theta_{(k)}\right)d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)+\log\left[\sum_{j=1}^K\pi(j)\,\nu_A(j)\right]\\ &=D_{\mathrm{KL}}(\rho,\pi)+\sum_{k=1}^K\left[\int_{\Theta_{(k)}}A\left(k,\theta_{(k)}\right)d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)+\log\left[\sum_{j=1}^K\pi(j)\,\nu_A(j)\right].\end{aligned}$$

Hence

$$\begin{aligned}&\log\left[\sum_{j=1}^K\pi(j)\,\nu_A(j)\right]\\ &=-\left\{D_{\mathrm{KL}}(\rho,\pi)+\sum_{k=1}^K\left[\int_{\Theta_{(k)}}A\left(k,\theta_{(k)}\right)d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)\right\}+D_{\mathrm{KL}}(\rho,\rho_{A,\pi}).\end{aligned}$$

■

References

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Babii, A., Chen, X., Ghysels, E., and Kumar, R. (2020). Binary choice with asymmetric loss in a data-rich environment: Theory and an application to racial justice.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1-3):85–113.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851. Springer Science & Business Media.

Catoni, O. (2007). Pac-bayesian supervised classification: The thermodynamics of statistical learning. institute of mathematical statistics lecture notes—monograph series 56. *IMS, Beachwood, OH. MR2483528*.

Cover, T. M. and Thomas, J. A. (2006). Elements of information theory second edition solutions to problems. *Internet Access*, pages 19–20.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Elliott, G. and Lieli, R. P. (2013). Predicting binary outcomes. *Journal of Econometrics*, 174(1):15–26.

Freund, Y., Mansour, Y., Schapire, R. E., et al. (2004). Generalization bounds for averaged classifiers. *Annals of Statistics*, 32(4):1698–1722.

Germain, P., Lacasse, A., Laviolette, F., March, M., and Roy, J.-F. (2015). Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). Pac-bayesian learning of linear classifiers. *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360.

Granger, C. W. and Machina, M. J. (2006). Forecasting and decision theory. volume 1 of *Handbook of Economic Forecasting*, pages 81–98. Elsevier.

Guedj, B. (2013). *Aggregation of estimators and classifiers: theory and methods*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207 – 2231.

Kaplan, D. M. and Sun, Y. (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory*, 33(1):105–157.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. (2006). PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. *Advances in Neural Information Processing Systems*, pages 769–776.

Langford, J. and Shawe-Taylor, J. (2003). Pac-bayes & margins. *Advances in neural information processing systems*, pages 439–446.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.

Massart, P. and Picard, J. (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer Berlin Heidelberg.

Maurer, A. (2004). A note on the pac bayesian theorem. *arXiv preprint cs/0411099*.

McAllester, D. (2003a). Simplified pac-bayesian margin bounds. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, pages 203–215, Berlin, Heidelberg. Springer Berlin Heidelberg.

McAllester, D. A. (1999). Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363.

McAllester, D. A. (2003b). PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. (2017). Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 30:5947–5956.

Ridgway, J., Alquier, P., Chopin, N., and Liang, F. (2014). Pac-bayesian auc classification and scoring. *arXiv preprint arXiv:1410.1771*.

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

Su, J.-H. (2020). Model selection in utility-maximizing binary prediction. *Journal of Econometrics*.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.