Bayesian posterior distributions are widely used for inference, but their dependence on a statistical model creates some challenges. In particular, there may be lots of nuisance parameters that require prior distributions and posterior computations, plus a potentially serious risk of model misspecification bias. Gibbs posterior distributions, on the other hand, offer direct, principled, probabilistic inference on quantities of interest through a loss function, not a model-based likelihood. Here we provide simple sufficient conditions for establishing Gibbs posterior concentration rates when the loss function is of a sub-exponential type. We apply these general results in a range of practically relevant examples, including mean regression, quantile regression, and sparse high-dimensional classification. We also apply these techniques in an important problem in medical statistics, namely, estimation of a personalized minimum clinically important difference.

Keywords: classification; generalized Bayes; high-dimensional problem; M-estimation; model misspecification

1. Introduction

A major selling point of the Bayesian framework is that it is normative: to solve a new problem, one only needs a statistical model/likelihood, a prior distribution for the model parameters, and the means to compute the corresponding posterior distribution. Bayesians’ obligation to specify a prior attracts criticism, but their need to specify a likelihood has a number of potentially negative consequences too, especially when the quantity of interest has meaning independent of a statistical model, like a quantile. On the one hand, even if the posited model is “correct,” it is rare that all the parameters of that model are relevant to the problem at hand. For example, if we are interested in a quantile of a distribution, which we model as skew-normal, then the focus is only on a specific real-valued function of the model parameters. In such a case, the non-trivial effort invested in dealing with these nuisance parameters, e.g., specifying prior distributions and designing computational algorithms, is effectively wasted. On the other hand, in the far more likely case where the posited model is “wrong,” that model misspecification can negatively impact one’s conclusions about the quantity of interest. For example, both skew-normal and Pareto models have a \( \tau \) quantile, but the quality of inferences drawn about that quantile will vary dramatically depending on which of these two models is chosen.

The non-negligible dependence on the posited statistical model puts a burden on the data analyst, and those reluctant to take that risk tend to opt for a non-Bayesian approach. After all, if one can get a solution without specifying a statistical model, then it is impossible to incur model misspecification bias. But in taking such an approach, they give up the normative advantage of Bayesian analysis. Can they get the best of both worlds? That is, can one construct a posterior distribution for the quantity of interest directly, incorporating available prior information, without specifying a statistical model and incurring the associated model misspecification risks, and without the need for marginalization over nuisance parameters? Fortunately, the answer is Yes, and this is the present paper’s focus.

The so-called Gibbs posterior distribution is the proper prior-to-posterior update when data and the interest parameter are linked by a loss function rather than a likelihood [8, 50, 5]. Intuitively, the
Gibbs and Bayesian posterior distributions coincide when the loss function linking data and parameter is a (negative) log-likelihood. In that case the properties of the Gibbs posterior can be inferred from the literature on Bayesian asymptotics in both the well-specified and misspecified contexts. For cases where the link is not through a likelihood, the large-sample behavior of the Gibbs posterior is less clear and elucidating this behavior under some simple and fairly general conditions is our goal here.

As a practical example, medical investigators are often interested to determine if a treatment whose effect has been judged to be statistically significant is also clinically significant in the sense that the patients actually feel better post-treatment. Therefore, they are interested in inference about the effect size cutoff beyond which patients feel better; this is called the minimum clinically important difference or MCID, e.g., [22]. Estimation of the MCID boils down to a classification problem, and standard Bayesian approaches to binary regression do not perform well in this setting; misspecifying the link function leads to bias, and nonparametric modeling of the link function is inefficient [11]. Instead, we found that a Gibbs posterior distribution, as described above, provided a very reasonable and robust solution to the MCID problem [40]. In some applications, one seeks a “personalized” or subject-specific cutoff that depends on a set of additional covariates. This personalized MCID could be high- or even infinite-dimensional, and our previous Gibbs posterior analysis is not equipped to handle such situations. But the framework developed here is; see Section 5.5.

In the following sections we lay out and apply conditions under which a Gibbs posterior distribution concentrates, asymptotically, on a neighborhood of the true value of the inferential target as the sample sizes increases. Our focus is not on the most general set of sufficient conditions for concentration; rather, we aim for conditions that are both widely applicable and easily verified. To this end, we consider loss functions of a sub-exponential type, ones that satisfy an inequality similar to the moment-generating function bound for sub-exponential random variables [6]. We can apply this condition in a variety of problems, from regression to classification, and in both fixed- and high-dimensional settings. An added advantage is that our conditions lead to straightforward proofs of concentration.

Section 2 provides some background and formally defines the Gibbs posterior distribution. In Section 3, we state our theoretical objectives and present our main results, namely, sets of sufficient conditions under which the Gibbs posterior achieves a specified asymptotic concentration rate. A unique attribute of the Gibbs posterior distribution is its dependence on a tuning parameter called the learning rate, and our results cover a constant, vanishing sequence, and even data-dependent learning rates. Section 4 further discusses verifying our conditions and extends our conditions and main results to handle certain unbounded loss functions. Section 5 applies our general theorems to establish Gibbs posterior concentration rates in a number of practically relevant examples, including nonparametric curve estimation, high-dimensional sparse classification, and our motivating example, estimation of the MCID function. Further examples, proofs, and technical details are included in supplementary material.

2. Background on Gibbs posteriors

2.1. Notation and definitions

Consider a measurable space \((\mathcal{U}, \mathcal{U})\), with \(\mathcal{U}\) a \(\sigma\)-algebra of subsets of \(\mathcal{U}\), on which a probability measure \(P\) is defined. A random element \(U \sim P\) need not be a scalar, and many of the applications we have in mind involve \(U = (X, Y)\) or \(U = (X, Y, Z)\), where \(Y\) denotes a “response” variable and \(X\) or \((X, Z)\) denotes a “predictor” variable, and \(P\) encodes the dependence between the entries in \(U\). Then the real-world phenomenon under investigation is determined by \(P\) and our goal is to make inference on a relevant feature of \(P\), which we define as a given functional \(\theta = \theta(P)\), taking values in \(\Theta\). Note that \(\theta\) could be finite-, high-, or even infinite-dimensional.
Gibbs posterior concentration rates

The specific way \( \theta \) relates to \( P \) guides our posterior construction. Suppose there is a loss function, \( \ell_\theta(u) \), that measures how closely a generic value of \( \theta \) agrees with a data point \( u \). (As is customary, “\( \theta \)” will denote both the quantity of interest and a generic value of that quantity; when we need to distinguish the true from a generic value, we will write “\( \theta^\star \).”) For example, if \( u = (x, y) \) is a predictor–response pair, and \( \theta \) is a function, then the loss might be

\[
\ell_\theta(u) = |y - \theta(x)| \quad \text{or} \quad \ell_\theta(u) = 1\{y \neq \theta(x)\},
\]

(1)
depending on whether \( y \) is continuous or discrete/binary, where \( 1(A) \) denotes the indicator function at the event \( A \). Another common situation is when one specifies a statistical model, say, \( \{P_\theta : \theta \in \Theta\} \), indexed by a parameter \( \theta \), and sets \( \ell_\theta(u) = -\log p_\theta(u) \), where \( p_\theta \) is the density of \( P_\theta \) with respect to some fixed dominating measure. In all of these cases, the idea is that a loss is incurred when there is a certain discrepancy between \( \theta \) and the data point \( u \). Then our inferential target is the value of \( \theta \) that minimizes the risk or average loss/discrepancy.

**Definition 1.** Consider a real-valued loss function \( \ell_\theta(u) \) defined on \( U \times \Theta \), and define the risk function \( R(\theta) = P\ell_\theta \), the expected loss with respect to \( P \); throughout, we write \( Pf \) to denote expected value of \( f(U) \) with respect to \( U \sim P \). Then the inferential target is

\[
\theta^\star \in \arg\min_{\theta \in \Theta} R(\theta).
\]

(2)

Given that estimation/inference is our goal, our focus will be on case where the risk minimizer, \( \theta^\star \), is unique, so that the “\( \in \)” in (2) is an equality. But this is not absolutely necessary for our theory. Indeed, the main results in Section 3 remain valid even if the risk minimizer is not unique, and we make a few brief comments about this extension in the discussion following Theorem 1.

The risk function itself is unavailable—it depends on \( P \)—and, therefore, so is \( \theta^\star \). However, suppose that we have an independent and identically distributed (iid) sample \( U^n = (U_1, \ldots, U_n) \) of size \( n \), with each \( U_i \) having marginal distribution \( P \) on \( U \). The iid assumption is not crucial, but it makes the notation and discussion easier; an extension to independent but not identically distributed (inid) cases is discussed in the context of an example in Section 5.2. In general, we have data \( U^n \) taking values in the measurable space \((U^n, \mathcal{U}^n)\), with joint distribution denoted by \( P^n \). From here, we can naturally replace the unknown risk in Definition 1 with an empirical version and proceed accordingly.

**Definition 2.** For a loss function \( \ell_\theta \) as described above, define the empirical risk as

\[
R_n(\theta) = \mathbb{P}_n \ell_\theta = \frac{1}{n} \sum_{i=1}^n \ell_\theta(U_i),
\]

(3)

where \( \mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{U_i} \), with \( \delta_u \) the Dirac point-mass measure at \( u \), is the empirical distribution.

Naturally, if the inferential target is the risk minimizer, then it makes sense to estimate that quantity based on data \( U^n \) by minimizing the empirical risk, i.e.,

\[
\hat{\theta}_n \in \arg\min_{\theta \in \Theta} R_n(\theta).
\]

(4)

This is the \( M \)-estimator based on an objective function determined by the loss \( \ell_\theta \); when \( R_n \) is differentiable, the root of \( R_n \), the derivative of \( R_n \), is a \( Z \)-estimator and “\( R_n(\theta) = 0 \)” is often called an estimating equation [14, 46]. Since \( \theta \mapsto \ell_\theta \) need not be smooth or convex, and \( R_n \) is an average over
a finite set of data, it is possible that its minimizer is not unique, even if $\theta^*$ is. These computational challenges are, in fact, part of what motivates the Gibbs posterior, as we discuss below.

There is a rich literature on the asymptotic distribution properties of M-estimators, which can be used for developing hypothesis tests and confidence intervals [32, 21]. As an alternative, one might consider a Bayesian approach to quantify uncertainty, but there is an immediate obstacle, namely, no statistical model/likelihood connecting the data to the quantity of interest. If we did have a statistical model, with a density function $p_{\theta}$, then the most natural loss is $\ell_{\theta}(u) = -\log p_{\theta}(u)$ and the likelihood is $\exp\{-nR_n(\theta)\}$. It is, therefore, tempting to follow that same strategy for general losses, resulting in a sort of generalized posterior distribution for $\theta$.

**Definition 3.** Given a loss function $\ell_{\theta}$ and the corresponding empirical risk $R_n$ in Definition 2, define the Gibbs posterior distribution as

$$
\Pi_n(d\theta) \propto e^{-\omega nR_n(\theta)} \Pi(d\theta), \quad \theta \in \Theta,
$$

where $\Pi$ is a prior distribution on $\Theta$ and $\omega > 0$ is a so-called learning rate [20, 41, 15, 47]. The dependence of $\Pi_n$ on $\omega$ will generally be omitted from the notation, but see Sections 2.2 and 3.3.

We will assume that the right-hand side of (5) is integrable in $\theta$, so that the proportionality constant is well-defined. Integrability holds whenever the loss function is bounded from below, like for those in (1), but this could fail in some cases where the loss is not bounded away from $-\infty$, e.g., when $\ell_{\theta}(u)$ is a negative log-density. In such cases, extra conditions on the prior distribution would be required to ensure the Gibbs posterior is well-defined.

An immediate advantage of this approach, compared to the M-estimation strategy described above, is that the user is able to incorporate available prior information about $\theta$ directly into the analysis. This is especially important in cases where the quantity of interest has a real-world meaning, as opposed to being just a model parameter, where having genuine prior information is the norm rather than the exception. Additionally, even though there is no likelihood, the same computational methods, such as Markov chain Monte Carlo [9] and variational approximations [2], common in Bayesian analysis, can be employed to numerically approximate the Gibbs posterior.

We have opted here to define the Gibbs posterior directly as an object to be used and studied, but there is a more formal, more principled way in which Gibbs posteriors emerge. In the PAC-Bayes literature, the goal is to construct a randomized estimator that concentrates in regions of $\Theta$ where the risk, $R(\theta)$, or its empirical version, $R_n(\theta)$, is small [45, 36, 1, 18]. That is, the Gibbs posterior can be viewed as a solution to an optimization problem rather than a solution to the updating-prior-beliefs problem. More formally, for a given prior $\Pi$ on $\Theta$, suppose the goal is to find

$$
\inf_{\mu} \left\{ \int R_n(\theta) \mu(d\theta) + (\omega n)^{-1} K(\mu, \Pi) \right\},
$$

where the infimum is over all probability measures $\mu$ that are absolutely continuous with respect to $\Pi$, and $K$ denotes the Kullback–Leibler divergence. Then it can be shown [e.g., 50, 5] that the unique solution is $\Pi_n$, the Gibbs posterior defined in (5). Therefore, the Gibbs posterior distribution is the measure minimizing a penalized risk, averaged with respect to a given prior, $\Pi$.

### 2.2. Learning rate

Readers familiar with M-estimation may not recognize the learning rate, $\omega$. This does not appear in the M-estimation context because all that influences the optimization problem—and the corresponding
asymptotic distribution theory—is the shape of the loss/risk function, not its magnitude or scale. On the other hand, the learning rate is an essential component of the Gibbs posterior distribution in (5) since the distribution depends on both the shape and scale of the loss function. Data-driven strategies for tuning the learning rate are available [20, 41, 29, 48, 49].

Here we focus on how the learning rate affects posterior concentration. In typical examples, our results require the learning rate to be a sufficiently small constant. That constant depends on features of $P$, which are generally unknown, so, in practice, the learning rate can be taken to be a slowly vanishing sequence, which has a negligible effect on the concentration rate. In more challenging examples, we require the learning rate to vanish sufficiently fast in $n$; this is also the case in [16].

2.3. Relation to other generalized posterior distributions

A generalized posterior is any data-dependent distribution other than a well-specified Bayesian posterior. Examples include Gibbs and misspecified Bayesian posteriors, which we compare first.

A key characteristic of the misspecified Bayesian posterior is that it is accidentally misspecified. That is, the data analyst does his/her best to posit a sound model, $Q_\gamma$, for the data-generating process and obtains the corresponding posterior for the model parameter $\gamma$. That model will typically be misspecified, i.e., $P \notin \{Q_\gamma : \gamma \in \Gamma\}$, so the aforementioned posterior will, under certain conditions, concentrate around the point $\gamma^\dagger$ that minimizes the Kullback–Leibler divergence of $Q_\gamma$ from $P$ [25, 12, 37]. For a feature $\theta(\cdot)$ of interest, typically $\theta(Q_\gamma^\dagger) \neq \theta(P)$, so there is generally a bias that the data analyst can do nothing about. A Gibbs posterior, on the other hand, is purposely misspecified—no attempt is made to model $P$. Rather, it directly targets the feature of interest via the loss function that defines it. This strategy avoids model misspecification bias, but its point of view also sheds light on the importance of the choice of learning rate. Since the data analyst knows the Gibbs posterior is not a correctly specified Bayesian posterior, they know the learning rate must be handled with care.

A number of authors have studied generalized posterior distributions formed using a likelihood raised to a fractional power $\eta \in (0, 1]$; see, e.g., [34] and [17] among others. These $\eta$-generalized posteriors tend to be robust to misspecification of the probability model, and data-driven methods to tune $\eta$ are developed in, e.g., [17]. These posteriors coincide with Gibbs posteriors based on a log-loss and with the learning rate $\omega$ corresponding to $\eta$.

A common non-Bayesian approach to statistics bases inferences on moment conditions of the form $Pf_j = 0$ for functions $f_1, \ldots, f_J$, rather than on a fully-specified likelihood. A number of authors have extended this framework to posterior inference. Kim [24] uses the moment conditions to construct a so-called limited-information likelihood to use in place of a fully-specified likelihood in a Bayesian formulation. Similarly, Chernozhukov and Hong [9] construct a posterior by taking a (pseudo) log-likelihood equal to a quadratic form determined by the set of moment conditions. Chib et al. [10] studies Bayesian exponentially-tilted empirical likelihood posterior distributions, also defined by moment conditions. In some cases the Gibbs posterior distribution coincides with the above moment-based methods. For instance, in the special case that the risk $R$ is differentiable at $\theta^*$ with derivative $\dot{R}(\theta^*)$, risk-minimization corresponds to the single moment condition $\dot{R}(\theta^*) = 0$.

3. Asymptotic concentration rates

3.1. Objective

A large part of the Bayesian literature concerns the asymptotic concentration properties of their posterior distributions. Roughly, if data are generated from a distribution for which the quantity of interest
takes value \( \theta^* \), then, as \( n \to \infty \), the posterior distribution ought to concentrate its mass around that same \( \theta^* \). As we will show, optimal concentration rates are possible with Gibbs posteriors, so the robustness achieved by not specifying a statistical model has no cost in terms of (asymptotic) efficiency.

Towards this, for a fixed \( \theta^* \in \Theta \), let \( d(\theta; \theta^*) \) denote a divergence measure on \( \Theta \) in the sense that \( d(\theta; \theta^*) \geq 0 \) for all \( \theta \), with equality if and only if \( \theta = \theta^* \). The divergence measure could depend on the sample size \( n \) or other deterministic features of the problem at hand, especially in the independent but not iid setting; see Section 5.2. Our objective is to provide conditions under which the Gibbs posterior will concentrate asymptotically, at a certain rate, around \( \theta^* \). Throughout this paper, \( (\varepsilon_n) \) denotes a deterministic sequence of positive numbers with \( \varepsilon_n \to 0 \), which will be referred to as the Gibbs posterior concentration rate.

**Definition 4.** The Gibbs posterior \( \Pi_n \) in (5) asymptotically concentrates around \( \theta^* \) at rate (at least) \( \varepsilon_n \), with respect to \( d \), if

\[
P^n\Pi_n(\{\theta : d(\theta; \theta^*) > M_n\varepsilon_n\}) \to 0, \quad \text{as } n \to \infty,
\]

where \( M_n > 0 \) is either a (deterministic) sequence satisfying \( M_n \to \infty \) arbitrarily slowly or is a sufficiently large constant, \( M_n \equiv M \).

In the PAC-Bayes literature, the Gibbs posterior distribution is interpreted as a “randomized estimator,” a generator of random \( \theta \) values that tend to make the risk difference small. For iid data and with risk divergence \( d(\theta; \theta^*) = \{R(\theta) - R(\theta^*)\}^{1/2} \), a concentration result like that in Definition 4 makes this strategy clear, since the \( \Pi_n \)-probability of the event \( \{R(\theta) - R(\theta^*) \leq \varepsilon_n^2\} \) would be \( \to 1 \).

If the Gibbs posterior concentrates around \( \theta^* \) in the sense of Definition 4, then any reasonable estimator derived from that distribution, such as the mean, should inherit the \( \varepsilon_n \) rate at \( \theta^* \) relative to the divergence measure \( d \). This can be made formal under certain conditions on \( d \); see, e.g., Corollary 1 in [3] and the discussion following the proof of Theorem 2.5 in [13].

Besides concentration rates, in certain cases it is possible to establish distributional approximations to Gibbs posteriors, i.e., Bernstein–von Mises theorems. Results for finite-dimensional problems and with sufficiently smooth loss functions can be found in, e.g., [4] and [9].

### 3.2. Conditions

Here we discuss a general strategy for proving Gibbs posterior concentration and the kinds of sufficient conditions needed for the strategy to be successful. To start, set \( A_n = \{\theta : d(\theta, \theta^*) > M_n\varepsilon_n\} \subset \Theta \).

Then our first step towards proving concentration is to express \( \Pi_n(A_n) \) as the ratio

\[
\Pi_n(A_n) = \frac{N_n(A_n)}{D_n} = \frac{\int_{A_n} \exp[-\omega n \{R_n(\theta) - R_n(\theta^*)\}] \Pi(d\theta)}{\int_{\Theta} \exp[-\omega n \{R_n(\theta) - R_n(\theta^*)\}] \Pi(d\theta)}.
\]

The goal is to suitably upper and lower bound \( N_n(A_n) \) and \( D_n \), respectively, in such a way that the ratios of these bounds vanish. Two sufficient conditions are discussed below, the first primarily dealing with the loss function and aiding in bounding \( N_n(A_n) \) and the second primarily concerning the prior distribution and aiding in bounding \( D_n \). Both conditions concern the excess loss \( \ell_\theta(U) - \ell_{\theta^*}(U) \) and its mean and variance:

\[
m(\theta, \theta^*) := P(\ell_\theta - \ell_{\theta^*}) \quad \text{and} \quad v(\theta, \theta^*) := P(\ell_\theta - \ell_{\theta^*})^2 - m(\theta, \theta^*)^2.
\]
3.2.1. Sub-exponential type losses

Our method for proving posterior concentration requires a vanishing upper bound on the expectation of the numerator term $N_n(A_n)$ in (7). Since the integrand, $\exp[-\omega n (R_n(\theta) - R_n(\theta^*))]$, in $N_n(A_n)$ is non-negative, Fubini’s theorem says it suffices to bound its expectation. Further, by independence

$$p^ne^{-\omega n (R_n(\theta) - R_n(\theta^*))} = \{p e^{\omega (\ell_\theta - \ell_{\theta^*})}\}^n,$$

which reveals that the key to bounding $p^n N_n(A_n)$ is to bound $p e^{-\omega (\ell_\theta - \ell_{\theta^*})}$, the expected exponentiated excess loss. In order for the bound on $p^n N_n(A_n)$ to vanish it must be that $p e^{-\omega (\ell_\theta - \ell_{\theta^*})} < 1$, but this is not enough to identify the concentration rate in (6). Rather, the speed at which $p e^{-\omega (\ell_\theta - \ell_{\theta^*})}$ vanishes must be a function of $d(\theta, \theta^*)$, and we take this relationship as our key condition for Gibbs posterior concentration. When this holds we say the loss function is of sub-exponential type.

**Condition 1.** There exists an interval $(0, \omega)$ and constants $K, r > 0$, such that for all $\omega \in (0, \omega)$ and for all sufficiently small $\delta > 0$, for $\theta \in \Theta$

$$d(\theta; \theta^*) > \delta \implies p e^{-\omega (\ell_\theta - \ell_{\theta^*})} < e^{-K \omega^r}.$$  

(8)

(The constant $r > 0$ that appears here and below can take on different values depending on the context. The case is $r = 2$ common, but some “non-regular” problems require $r \neq 2$; see Section 5.3.)

An immediate consequence of Condition 1 and the definition of $A_n$ is the key finite-sample, exponential bound on the Gibbs posterior numerator

$$p^n N_n(A_n) \leq e^{-K \omega n M_n^e e_n^e}.$$  

(9)

For some intuition behind Condition 1, consider the following. Let $f$ be a real-valued function such that the random variable $f(U)$ has a distribution with sufficiently thin tails. This is automatic when $f$ is bounded, but suppose $f(U)$ has a moment-generating function, i.e., is sub-exponential. For bounding the moment-generating function, the dream case would be if $p e^{-\omega f} \leq e^{-\omega Pf}$. Unfortunately, Jensen’s inequality implies the dream is an impossibility. It is possible, however, to show

$$p e^{-\omega f} \leq e^{-\omega Pf + \omega^2 G(f)},$$

for suitable $G(f) > 0$ depending on certain features of $f$ (and of $P$). If it could also be shown, e.g., that $G(f) \leq Pf$, then we are in a “near-dream” case where

$$p e^{-\omega f} \leq e^{-c \omega Pf},$$

for a constant $c \in (0, 1)$ and sufficiently small $\omega$.

The little extra needed beyond sub-exponential to achieve the “near-dream” case bound above is why we refer to such $f$ as being sub-exponential type.

Towards verifying Condition 1 we briefly review some developments in Grünwald and Mehta [16]; further comments can be found in Section 3.4. These authors focus on an annealed expectation which, for a real-valued function $f$ of the random element $U \sim P$ and a fixed constant $\omega > 0$, is defined as $P^{\text{ANN}}(\omega) f = -\omega^{-1} \log P e^{-\omega f}$. With this, we find that $P e^{-\omega f} = \exp\{-\omega P^{\text{ANN}}(\omega) f\}$. So an upper bound as we require in Condition 1 is equivalent to a corresponding lower bound on $P^{\text{ANN}}(\omega) (\ell_\theta - \ell_{\theta^*})$. The “strong central condition” in [16] states that

$$P^{\text{ANN}}(\omega) (\ell_\theta - \ell_{\theta^*}) \geq 0,$$

for all $\theta \in \Theta$. 

The above inequality implies \( P e^{-\omega(\ell_\theta - \ell_{\theta^*})} \leq 1 \) and, in turn, that \( P^n N_n(A_n) = O(1) \) as \( n \to \infty \). The other conditions in [16] aim to lower-bound the annealed expected excess loss by a suitable function of the excess risk. For example, Lemma 13 in [16] shows that a “witness condition” implies

\[
P^{\text{ANN}}(\omega)(\ell_\theta - \ell_{\theta^*}) \geq R(\theta) - R(\theta^*), \quad \text{for all } \theta \in \Theta,
\]

so, if \( R(\theta) - R(\theta^*) \) is of the order \( d(\theta, \theta^*)^r \) for some \( r > 0 \), then we recover Condition 1. Therefore, our Condition 1 is exactly what is needed to control the numerator of the Gibbs posterior, and the strong central and witness conditions developed in [16] and elsewhere constitute a set of sufficient conditions for our Condition 1.

A subtle difference between our approach and [16]’s is that the bounds in the above two displays are required for all \( \theta \in \Theta \); Condition 1 only deals with \( \theta \) bounded away from \( \theta^* \). This difference arises because we directly target a bound on the Gibbs posterior probability, \( \Pi_n(A_n) \), an integration over \( A_n \neq \emptyset \), whereas [16] targets a bound on the Gibbs posterior mean of \( \theta \to R(\theta) - R(\theta^*) \), an integration over all of \( \Theta \). In Example 3.8 of [47] the global requirement in [16]’s condition is a disadvantage as [47] illustrate it creates some challenges in checking the bounds in the above two displays.

Besides Condition 1, other strategies for bounding \( N_n(A_n) \) in (7) have been employed in the literature. For example, empirical process techniques are used in [40, 4] to prove Gibbs posterior concentration in finite-dimensional applications. Generally, such proofs hinge on a uniform law of large numbers, which can be challenging to verify in non-parametric problems. Chernozhukov and Hong [9] require the stronger condition that \( \sup_{d(\theta, \theta^*) \geq \delta} \{ R_n(\theta) - R_n(\theta^*) \} \to 0 \) in \( P^n \)-probability. When it holds they immediately recover an in-probability bound on \( N_n(A_n) \), irrespective of \( \omega \), but they do not obtain a finite-sample bound on \( P^n \Pi_n(A_n) \). Their analysis is limited to finite-dimensional parameters and empirical risk functions \( R_n(\theta) \) that are smooth in a neighborhood of \( \theta^* \). This latter condition excludes important examples like the misclassification-error loss function; see Section 5.3.

There are situations in which Condition 1 does not hold, and we address this formally in Section 4. As an example, note that both Condition 1 and the witness condition used in [16] are closely related to a Bernstein-type condition relating the first two moments of the excess loss: \( P(\ell_\theta - \ell_{\theta^*})^2 \leq c(\theta - \theta^*)^2 \), for constants \( (c, \alpha) > 0 \). For the “check” loss used in quantile estimation, the Bernstein condition is generally not satisfied and, consequently, neither [16]’s witness condition nor our Condition 1—at least not in their original forms—can be verified. Similarly, if the excess loss, \( \ell_\theta(U) - \ell_{\theta^*}(U) \), a function of data \( U \), is heavy-tailed, then the moment-generating function bound in Condition 1 does not hold. For these cases, some modifications to the basic setup are required, which we present in Section 4.

3.2.2. Prior distribution

Generally, the prior must place enough mass on certain “risk-function” neighborhoods \( G_n := \{ \theta : m(\theta, \theta^*) \leq \epsilon_n^\theta \} \). This is analogous to the requirement in the Bayesian literature that the prior place sufficient mass on Kullback–Leibler neighborhoods of \( \theta^* \); see, e.g., [39] and [13]. Some version of the following prior bound is needed

\[
\log \Pi(G_n) \geq -n \epsilon_n^\theta, \quad (10)
\]

for \( r \) as in Condition 1. Lemma 1 in Section 1 of the supplementary material along with (10) provides a lower probability bound on the denominator term \( D_n \) defined in (7). That is,

\[
P^n(D_n \leq \frac{1}{2} \Pi(G_n) e^{-2\omega n \epsilon_n^\theta}) \leq P(D_n \leq \frac{1}{2} e^{-(C+2\omega)n \epsilon_n^\theta}) \leq (n \epsilon_n^\theta)^{-1}, \quad (11)
\]
Gibbs posterior concentration rates

where \( b_n = \frac{1}{4}\Pi(G_n)e^{-2\omega n\varepsilon_n^r} \), which, according to (10), is of the order \( e^{-C\omega n\varepsilon_n^r} \), for an appropriate constant \( C > 0 \). Both the form of the risk-function neighborhoods and the precise lower bound in (10) depend on the concentration rate and the learning rate, so the results in Section 3.3 all require their own version of the above prior bound.

Grünwald and Mehta [16] only require bounds on the prior mass of the larger neighborhoods \( \{ \theta : m(\theta, \theta^*) \leq \varepsilon_n^r \} \). Under their condition, we can derive a lower bound on \( P^n D_n \) similar to Lemma 1 in [33]. However, our proofs require an in-probability lower bound on \( G \), which in turn requires stronger prior concentration like that in (10). While there are important examples where the lower bounds on \( m(\theta, \theta^*) \) and \( v(\theta, \theta^*) \) are of different orders, none of the applications we consider here are of that type. Therefore, the stronger prior concentration condition in (10) does not affect the rates we derive for the examples in Section 5 and in the Supplementary Materials. Moreover, as discussed following the statement of Theorem 1 below, our finite-sample bounds are better than those in [16], a direct consequence of our method of proof that uses a smaller neighborhood \( G_n \).

### 3.3. Main results

In this section we present general results on Gibbs posterior concentration. Proofs can be found in Section 1 of the supplementary material. Our first result establishes Gibbs posterior concentration, under Condition 1 and a local prior condition, for sufficiently small constant learning rates.

**Theorem 1.** Let \( \varepsilon_n \) be a vanishing sequence satisfying \( n\varepsilon_n^r \to \infty \) for a constant \( r > 0 \). Suppose the prior satisfies

\[
\log \Pi(\{ \theta : m(\theta, \theta^*) \vee v(\theta, \theta^*) \leq \varepsilon_n^r \}) \geq -C_1 n\varepsilon_n^r,
\]

for \( C_1 > 0 \) and for divergence measure \( d \), the same \( r > 0 \) as above, and learning rate \( \omega \in (0, \bar{\omega}) \) for some \( \bar{\omega} > 0 \). If the loss function satisfies Condition 1, then the Gibbs posterior distribution in (5) has asymptotic concentration rate \( \varepsilon_n \), with \( M_n \equiv M \) a large constant as in Definition 4.

For a brief sketch of the proof bound the posterior probability of \( A_n \) by

\[
\Pi_n(A_n) \leq \frac{N_n(A_n)}{D_n} \frac{1(D_n > b_n) + 1(D_n \leq b_n)}{b_n^{-1}N_n(A_n) + 1(D_n \leq b_n)},
\]

where \( b_n \) is as above, given explicitly in the proof Lemma 1. Taking expectation of both sides, and applying (11) and the consequence (9) of Condition 1, we get

\[
P^n\Pi_n(A_n) \leq b_n^{-1}e^{-K\omega nM_n^0\varepsilon_n^r} + (n\varepsilon_n^r)^{-1}.
\]

Then the right-hand side is generally of order \( (n\varepsilon_n^r)^{-1} \to 0 \). To compare with the results in [16, Example 2], their upper bound on \( P^n\Pi_n(A_n) \) is \( O(M_n^{-1}) \), which vanishes arbitrarily slowly.

For the case where the risk minimizer in (2) is not unique, certain modifications of the above argument can be made, similar to those in [25, Theorem 2.4]. Roughly, to our Theorem 1 above, we would add the requirements that (12) and Condition 1 hold uniformly in \( \theta^* \in \Theta^* \), where \( \Theta^* \) is the set of risk minimizers. Then virtually the same proof strategy could be used to show that

\[
E\Pi_n(\{ \theta : d(\theta, \Theta^*) \geq \varepsilon_n \}) \to 0,
\]

where \( d(\theta, \Theta^*) = \inf_{\theta^* \in \Theta^*} d(\theta, \theta^*) \).

Theorem 1 is quite flexible and can be applied in a range of settings; see Section 5 and the Supplementary Materials. However, one case in which it cannot be directly applied is when \( n\varepsilon_n^r \) is bounded.
For example, in sufficiently smooth finite-dimensional problems, we have $r = 2$ and the target rate is $\varepsilon_n = n^{-1/2}$. The difficulty is caused by the prior bound in (12), since it is impossible—at least with a fixed prior—to assign mass bounded away from 0 to a shrinking neighborhood of $\theta^*$. One option is to add a logarithmic factor to the rate, i.e., take $e^{-Cn\varepsilon_n^2}$ is a power of $n^{-1/2}$.

Alternatively, a refinement of the proof of Theorem 1 lets us avoid slowing down the rate.

**Theorem 2.** Consider a finite-dimensional $\theta$, taking values in $\Theta \subseteq \mathbb{R}^q$ for some $q \geq 1$. Suppose that the target rate $\varepsilon_n$ is such that $n\varepsilon_n^r$ is bounded for some constant $r > 0$. If the prior $\Pi$ satisfies

$$\Pi(\{\theta : m(\theta, \theta^*) \lor v(\theta, \theta^*) \leq \varepsilon_n^2\}) \gtrsim \varepsilon_n^q,$$

and if Condition 1 holds, then the Gibbs posterior distribution in (5), with any learning rate $\omega \in (0, \bar{\omega})$, has asymptotic concentration rate $\varepsilon_n$ at $\theta^*$ with respect to any divergence $d(\theta, \theta^*)$ satisfying $\|\theta - \theta^*\| \lesssim d(\theta, \theta^*) \lesssim \|\theta - \theta^*\|$ and for any diverging, positive sequence $M_n$ in Definition 4.

The learning rate is critical to the Gibbs posterior’s performance, but in applications it may be challenging to determine the upper bound $\bar{\omega}$ for which Condition 1 holds. For a simple illustration, suppose the excess loss $\ell_0 - \ell_0^*$ is normally distributed with variance $\sigma^2(\theta)$ satisfying $\sigma^2(\theta) \leq c(R(\theta) - R(\theta^*))$, a kind of Bernstein condition. In this case, Condition 1 holds for $d(\theta, \theta^*) = \{R(\theta) - R(\theta^*)\}^{1/2}$, with $r = 2$, if $\omega < 2c^{-1}$. Bernstein conditions can be verified in many practical examples, but the factor $c$ would rarely be known. Consequently, we need $\omega$ to be sufficiently small, but the meaning of “sufficiently small” depends on unknowns involving $P$. However, any positive, vanishing learning rate sequence $\omega = \omega_n$ satisfies $\omega_n \in (0, \bar{\omega})$ for all sufficiently large $n$. And if $\omega_n$ vanishes arbitrarily slowly, then it has no effect on the Gibbs posterior concentration rate; see Section 5.4. All we require to accommodate a vanishing learning rate is a slightly stronger $\omega_n$-dependent version of the prior concentration bound in (12).

**Theorem 3.** Let $\varepsilon_n$ be a vanishing sequence and $\omega_n$ be a learning rate sequence satisfying $n\omega_n\varepsilon_n^r \to \infty$ for a constant $r > 0$. Consider a Gibbs posterior distribution $\Pi_n = \Pi_n^{\omega_n}$ in (5) based on this sequence of learning rates. If the prior satisfies

$$\log \Pi(\{\theta : m(\theta, \theta^*) \lor v(\theta, \theta^*) \leq \varepsilon_n^r\}) \geq -Cn\omega_n\varepsilon_n^r,$$

and if Condition 1 holds for $\omega_n$, then the Gibbs posterior distribution in (5), with learning rate sequence $\omega_n$, has concentration rate $\varepsilon_n$ at $\theta^*$ for a sufficiently large constant $M > 0$ in Definition 4.

The proof of Theorem 3 is almost identical to that of Theorem 1, hence omitted. But to see the basic idea, we mention two key observations. First, since Condition 1 holds for $\omega_n$ for all sufficiently large $n$, and since $\omega_n$ is deterministic, by the same argument producing the bound in (9), we get

$$P^n N_n(A_n) \leq e^{-Kn\omega_n M_n e^{r_n}} \text{ for all sufficiently large } n.$$ 

Second, the same argument producing the bound in (11) shows that

$$P^n (D_n \leq \frac{1}{2}\Pi(G_n) e^{-2\omega_n\varepsilon_n^r}) \leq P(D_n \leq \frac{1}{2} e^{-(C+2)\omega_n\varepsilon_n^r}) \leq (n\omega_n\varepsilon_n^r)^{-1}.$$ 

Then the only difference between this situation and that in Theorem 1 is that, here, the numerator bound only holds for “sufficiently large $n$,“ instead of for all $n$. This makes Theorem 3 is slightly weaker than Theorem 1 since there are no finite-sample bounds.
When $\omega_n \equiv \omega$ the constant learning rate is absorbed by $C$ and there is no difference between the prior bounds in (12) and (14). But, the prior probability assigned to the $(m \lor v)$-neighborhood of $\theta^*$ does not depend on $\omega_n$, so if it satisfies (12), then the only way it could also satisfy (14) is if $\varepsilon_n$ is bigger than it would have been without a vanishing learning rate. In other words, Theorem 1 requires $n\varepsilon_n^r \to \infty$ whereas Theorem 3 requires $n\omega_n\varepsilon_n^r \to \infty$, which implies that for a given vanishing $\omega_n$ Theorem 1 potentially can accommodate a faster rate $\varepsilon_n$. Therefore, we see that a vanishing learning rate can slow down the Gibbs posterior concentration rate if it does not vanish arbitrarily slowly. There are applications that require the learning rate to vanish at a particular $n$-dependent rate, and these tend to be those where adjustments like in Section 4 are needed; see Sections 5.1.2 and 5.3.2.

If we can use the data to estimate $\bar{\omega}$ consistently, then it makes sense to choose a learning rate sequence depending on this estimator. Suppose our data-dependent learning rate $\hat{\omega}_n$ satisfies $\hat{\omega}_n < \bar{\omega}$ with probability converging to 1 as $n \to \infty$. For $\omega \equiv \hat{\omega}_n$ the conclusion of Theorem 1 holds for all sufficiently large $n$; and see Section 4.2 in the Supplementary Material. One advantage of this strategy is that it avoids using a vanishing learning rate, which may slow concentration.

**Theorem 4.** Fix a positive deterministic learning rate sequence $\omega_n$ such that the conditions of Theorem 3 hold and as a result $\Pi_{\omega_n}^\omega$ has asymptotic concentration rate $\varepsilon_n$. Consider a random learning rate sequence $\hat{\omega}_n$ satisfying 

$$P^n(\omega_n/2 < \hat{\omega}_n < \omega_n) \to 1, \quad n \to \infty. \quad (15)$$

Then $\Pi_{\hat{\omega}_n}^\omega$, the Gibbs posterior distribution in (5) scaled by the random learning rate sequence $\hat{\omega}_n$, also has concentration rate $\varepsilon_n$ at $\theta^*$ for a sufficiently large constant $M > 0$ in Definition 4.

### 3.4. Checking Condition 1

Of course, Condition 1 is only useful if it can be checked in practically relevant examples. As mentioned in Section 3.2, Grünwald and Mehta’s strong central and witness conditions are sufficient for Condition 1. A pair of slightly stronger, but practically verifiable conditions require the excess loss is sub-exponential with first and second moments that obey a Bernstein condition. For the examples we consider in Section 5 where Condition 1 can be used, these two conditions are convenient.

#### 3.4.1. Bounded excess losses

For bounded excess losses, i.e., $\ell_{\theta}(u) - \ell_{\theta^*}(u) < C$ for all $(\theta, u)$, Lemma 7.26 in [28] gives:

$$Pe^{-\omega(\ell_{\theta} - \ell_{\theta^*})} \leq \exp\left[-\omega m(\theta, \theta^*) + \omega^2 v(\theta, \theta^*) \left\{ \frac{\exp(C\omega) - 1 - C\omega}{C^2\omega^2} \right\} \right]. \quad (16)$$

Whether Condition 1 holds depends on the choice of $\omega$ and the relationship between $m(\theta, \theta^*)$ and $v(\theta, \theta^*)$ as defined by the Bernstein condition:

$$v(\theta, \theta^*) \leq C_1 m(\theta, \theta^*)^\alpha, \quad \alpha \in (0, 1], \quad C_1 > 0. \quad (17)$$

Denote the bracketed expression in the exponent of (16) by $C(\omega) = (C^2\omega^2)^{-1}\{\exp(C\omega) - 1 - C\omega\}$. When $\alpha = 1$, the $m$ and $v$ functions are of the same order and, if $d(\theta, \theta^*) = m(\theta, \theta^*)$, and $\omega C(\omega) \leq (2C_1)^{-1}$ then Condition 1 holds with $K = 1/2$. Since $C(\omega) \approx C\omega$ for small $\omega$, it suffices to take the learning rate a sufficiently small constant.
On the other hand, when \( v \) is larger than \( m \), i.e., the Bernstein condition holds with \( \alpha < 1 \), Condition 1 requires that \( \omega \) depend on \( \alpha \) and \( m(\theta, \theta^*) \). Suppose \( m(\theta, \theta^*) \geq d(\theta, \theta^*) > \varepsilon_n \). For all small enough \( \varepsilon_n \), we can set \( \omega_n = (4C)^{-1/2} \varepsilon_n^{(1-\alpha)/2} \) and Condition 1 is again satisfied with \( K = 1/2 \).

The above strategies are implemented in the classification examples in Section 5.3 where we also discuss connections to the Tsybakov [44] and Massart [35] conditions. For the former, see Theorem 22 and the subsequent discussion in [16], where the learning rate ends up being a vanishing sequence depending on the concentration rate and the Bernstein exponent \( \alpha \).

3.4.2. Sub-exponential excess losses

Unbounded but light-tailed loss differences \( \ell_\theta - \ell_{\theta^*} \) may also satisfy Condition 1. Both sub-exponential and sub-Gaussian random variables [6, Sec. 2.3–4] admit an upper bound on their moment-generating functions. When the loss difference \( \ell_\theta - \ell_{\theta^*} \) is sub-Gaussian

\[
P(e^{-\omega(\ell_\theta - \ell_{\theta^*})}) \leq \exp\left\{ \frac{\omega^2}{2} \sigma^2(\theta, \theta^*) - \omega m(\theta, \theta^*) \right\},
\]

for all \( \omega \), where the variance proxy \( \sigma^2(\theta, \theta^*) \) may depend on \( (\theta, \theta^*) \). If \( \ell_\theta - \ell_{\theta^*} \) is sub-exponential, then the above bound holds for all \( \omega \leq (2b)^{-1} \) for some \( b < \infty \) indexing the tail behavior of \( P \).

If \( \sigma^2(\theta, \theta^*) \) is upper-bounded by \( L m(\theta, \theta^*) \) for a constant \( L > 0 \), then the above bound can be rewritten as

\[
P(e^{-\omega(\ell_\theta - \ell_{\theta^*})}) \leq \exp\left\{ -\omega m(\theta, \theta^*) (1 - \frac{\omega L}{2}) \right\}.
\]

Then Condition 1 holds if the loss difference is sub-Gaussian and sub-exponential if \( \omega < 2L^{-1} \) and \( \omega < 2L^{-1} \) and \( \omega \) respectively.

In practice it may be awkward to assume \( \ell_\theta(U) - \ell_{\theta^*}(U) \) is sub-exponential, but in certain problems it is sufficient to make such an assumption about features of \( U \), which may be more reasonable. See Section 5.2 for an application of this idea to a fixed-design regression problem where the fact the response variable is sub-Gaussian implies the excess loss is itself sub-Gaussian.

4. Extensions

4.1. Locally sub-exponential type loss functions

In some cases the moment generating function bound in Condition 1 can be verified in a neighborhood of \( \theta^* \) but not for all \( \theta \in \Theta \). For example, suppose \( \theta \mapsto \ell_\theta(u) \) is Lipschitz with respect to a metric \( \| \cdot \| \) with uniformly bounded Lipschitz constant \( L = L(u) \), and that, for some \( \delta > 0 \),

\[
\| \theta - \theta^* \| < \delta \implies v(\theta, \theta^*) \lesssim m(\theta, \theta^*) \quad \text{and} \quad \| \theta - \theta^* \| > \delta \implies v(\theta, \theta^*) \lesssim m(\theta, \theta^*)^2.
\]

That is, the Bernstein condition (17) holds but for different values of \( \alpha \) depending on \( \theta \). This is the case for quantile regression; see Section 5.4. This class of problems, where the Bernstein exponent varies across \( \Theta \), apparently has not been considered previously. For example, [16] only consider cases where the Bernstein exponent \( \alpha \) is constant across the entire parameter space; consequently, their results assume \( \Theta \) is bounded [e.g., 16, Example 10], so (17) holds trivially with exponent \( \alpha = 0 \).

To address this problem, our idea is a simple one. We propose to introduce a sieve which is large enough that we can safely assume it contains \( \theta^* \) but small enough that, on which, the \( m \) and \( v \) functions can be appropriately controlled. Towards this, let \( \Theta_n \) be an increasing sequence of subsets of \( \Theta \), indexed by the sample size \( n \). The “size” of \( \Theta_n \) will play an important role in the result below. While
more general sieves are possible, to keep the notion of size concrete, let \( \Theta_n = \{ \theta \in \Theta : \| \theta \| \leq \Delta_n \} \), so that size is controlled by the non-decreasing sequence \( \Delta_n > 0 \), which would typically satisfy \( \Delta_n \to \infty \) as \( n \to \infty \). The metric \( \| \cdot \| \) in the definition of \( \Theta_n \) is at the user’s discretion, it would be chosen so that \( \theta \in \Theta_n \) provides information that can be used to control the excess loss \( \ell_\theta - \ell_{\theta^*} \). This leads to the following straightforward modification of Condition 1.

**Condition 2.** For \( \Theta_n \) with size controlled by \( \Delta_n \), there exists an interval \((0, \bar{\omega})\), a constant \( r > 0 \), and a sequence \( K_n = K(\Delta_n) > 0 \) such that, for all \( \omega \in (0, \bar{\omega}) \) and for all small \( \delta > 0 \),

\[
\theta \in \Theta_n \quad \text{and} \quad d(\theta; \theta^*) > \delta \implies P e^{-\omega(\ell_\theta - \ell_{\theta^*})} < e^{-\omega K_n^r \delta^r}.
\]

(20)

Aside from the restriction \( \Theta_n \), the key difference between the bounds here and in Condition 1 is in the exponent. Instead of there being a constant \( K \), there is a sequence \( K_n \) which is determined by the sieve’s size, controlled by \( \Delta_n \). If the sequence \( K_n \) is increasing, as we expect it will be, then we can anticipate that the overall concentration rate would be adversely affected—unless the learning rate is vanishing suitably fast. The following theorem explains this more precisely.

Condition 2 can be used exactly as Condition 1 to suitably upper bound \( P^n N_n(A_n \cap \Theta_n) \). However, the Gibbs posterior probability assigned to \( A_n \cap \Theta_n^c \) must be handled separately.

**Theorem 5.** Let \( \Theta_n \) be a sequence of subsets for which the loss function satisfies Condition 2 for a sequence \( K_n > 0 \), a constant \( r > 0 \), and a learning rate \( \omega_n \in (0, \bar{\omega}) \) for all sufficiently large \( n \). Let \( \varepsilon_n \) be a vanishing sequence satisfying \( n \omega_n K_n^r \varepsilon_n^r \to \infty \). Suppose the prior satisfies

\[
\log \Pi(\{ \theta : m(\theta, \theta^*) \vee v(\theta, \theta^*) \leq (K_n \varepsilon_n)^r \}) \geq -C n \omega_n K_n^r \varepsilon_n^r,
\]

(21)

for some \( C > 0 \) and the same \( K_n \), \( r \), and \( \omega_n \) as above. Then the Gibbs posterior in (5) satisfies

\[
\limsup_{n \to \infty} P^n \Pi_n(A_n) \leq \limsup_{n \to \infty} P^n \Pi_n(A_n \cap \Theta_n^c).
\]

Consequently, if

\[
P^n \Pi_n(\Theta_n^c) \to 0 \quad \text{as} \quad n \to \infty,
\]

(22)

then the Gibbs posterior has concentration rate \( \varepsilon_n \) for all large constants \( M > 0 \) in Definition 4.

There are two aspects of Theorem 5 that deserve further explanation. We start with the point about separate handling of \( \Theta_n^c \). Condition (22) is easy to check for well-specified Bayesian posteriors, but these results do not carry over even to a misspecified Bayesian model. Of course, (22) can always be arranged by restricting the support of the prior distribution to \( \Theta_n \), which is the suggestion in [25, Theorem 2.3] for the Bayesian case and how we handle (22) for our infinite-dimensional Gibbs example in Section 5.4. However, as (22) suggests, this is not entirely necessary. Indeed, [25] describe a trade-off between model complexity and prior support, and offer a more complicated form of their sufficient condition, which they do not explore in that paper. The fact is, without a well-specified likelihood, checking a condition like our (22) or Equation (2.13) in [25] is a challenge, at least in infinite-dimensional problems. For finite-dimensional problems, it may be possible to verify (22) directly using properties of the loss functions. For example, we use convexity of the check loss function for quantile estimation to verify (22) in Sections 2 and 4.1 of the Supplementary Material.
Next, how might one proceed to check Condition 2? Go back to the Lipschitz loss case at the start of this subsection. For a sieve $\Theta_n$ as described above, if $\theta$ and $\theta^*$ are in $\Theta_n$, then $||\theta - \theta^*||$ is bounded by a multiple of $\Delta_n$. This, together with the Lipschitz property and (16), implies

$$P e^{-\omega_n (t_0 - t_{0*})} \leq \exp \{ C_n \omega_n^2 \theta (\theta, \theta^*) - \omega_n m (\theta, \theta^*) \},$$

where $C_n = O(1 + \Delta_n \omega_n)$. Suppose the user chooses $\omega_n$ and $\Delta_n$ such that $\omega_n \Delta_n = O(1)$; then we can replace $C_n$ by a constant $C$. If there exist functions $g$ and $G$ such that

$$m (\theta, \theta^*) \geq g (\Delta_n) ||\theta - \theta^*||^2 \quad \text{and} \quad v (\theta, \theta^*) \leq G (\Delta_n) ||\theta - \theta^*||^2, \quad \theta \in \Theta_n,$$

then the above upper bound simplifies to

$$\exp \{-\omega_n ||\theta - \theta^*||^2 \{ g (\Delta_n) - C \omega_n G (\Delta_n) \} \}, \quad \theta \in \Theta_n.$$

Then Condition 2 holds with $K_n = g(\Delta_n) - C \omega_n G(\Delta_n)$, and it remains to balance the choices of $\omega_n$ and $\Delta_n$ to achieve the best possible concentration rate $\varepsilon_n$; see Section 5.4.

### 4.2. Clipping the loss function

When the excess loss is heavy-tailed, i.e., not sub-exponential, like in Section 5.1, its moment-generating function does not exist and, therefore, Condition 1 cannot be satisfied. In this section, we assume the loss function $\ell_\theta (u)$ is non-negative or lower-bounded by a negative constant. In the latter case we can work with the shifted loss—$\ell_\theta$ minus its lower bound. Many practically useful loss functions are non-negative, including squared-error loss, which we cover in Section 5.1.

In such cases, it may be reasonable to replace the heavy-tailed loss with a clipped version $\ell_\theta^\alpha (u) = \ell_\theta (u) \wedge t_n,$

where $t_n > 0$ is a diverging clipping sequence. Since $\ell_\theta^\alpha (u)$ is bounded in $(u, \theta)$ for each fixed $n$, the strategy for checking Condition 1 described in Section 3.4.1 suggests that, for certain choices of $(t_n, \varepsilon_n, \omega_n)$, $\Pi_n$ places vanishing mass on the sequence of sets $\{ \theta : P \ell_\theta^\alpha - P \ell_\theta^\alpha > \varepsilon_n \}$, where $\theta_n^\alpha = \arg \min \{ \ell_\theta : P \ell_\theta - P \ell_{\theta_n^\alpha} > \varepsilon_n \}$. This makes the $\theta_n^\alpha$ the (moving) target of the Gibbs posterior, instead of the fixed $\theta^*$. On the other hand, if the loss function admits more than one finite moment, then for a corresponding, increasing clipping sequence $t_n$, the clipped risk neighborhoods of $\theta_n^\alpha$ contain the risk neighborhoods of $\theta^*$ for all large $n$, that is,

$$\{ \theta : R (\theta) - R (\theta^*) > C \varepsilon_n \} \subset \{ \theta : P \ell_\theta^\alpha - P \ell_{\theta_n^\alpha} > \varepsilon_n \}, \quad \text{for some } C > 0.$$  

Then we further expect concentration of the clipped loss-based Gibbs posterior at $\theta^*$ with respect to the excess risk divergence at rate $\varepsilon_n$. Condition 3 and Theorem 6 below provide a set of sufficient conditions under which these expectations are realized and a concentration rate can be established.

**Condition 3.** Let $\ell_\theta$ be the loss. Define $\ell_\theta^\alpha = \ell_\theta \wedge t_n$ as the clipped loss and $\Theta_n = \{ \theta : ||\theta|| \leq \Delta_n \}$ as a sieve, depending on constants $t_n, \Delta_n \to \infty$.

1. For some $s > 1$, the sequence $B_n = \sup_{\theta \in \Theta_n} P [|\ell_\theta|^s]$ is finite for all $n$.
2. There exists a sequence $\omega_n > 0$, and a sequence $K_n > 0$, such that for all sequences $0 < \omega_n \leq \omega_n$ and for all sufficiently small $\delta > 0$,

$$\theta \in \Theta_n \quad \text{and} \quad P \ell_\theta^\alpha - P \ell_{\theta_n^\alpha} > \delta \implies P e^{-\omega_n (t_0 - t_{0*})} < e^{-K_n \omega_n \delta}.$$  

(25)
Theorem 6. For a given loss $\ell_\theta$ and sieve $\Theta_n$, suppose that Condition 3 holds for $(\omega_n, \Delta_n, t_n)$; and let $B_n = B_n(s)$ be as defined in Condition 3.1, for $s > 1$. Let $\varepsilon_n$ be a vanishing sequence such that $n\omega_n K_n \varepsilon_n \to \infty$, and suppose the prior satisfies
\[
\log \Pi(\{\theta : m_n(\theta, \theta^*_n) \lor v_n(\theta, \theta^*_n) \leq K_n \varepsilon_n\}) \geq -C n\omega_n K_n \varepsilon_n. \tag{26}
\]
Then the Gibbs posterior in (5) based on the clipped loss $\ell_\theta$ satisfies
\[
\limsup_{n \to \infty} P^n \Pi_n(A_n) \leq \limsup_{n \to \infty} P^n \Pi_n(A_n \cap \Theta_n^C)
\]
where $A_n := \{\theta : R(\theta) - R(\theta^*) > \varepsilon_n \lor B_n t_n^{1-s}\}$. Consequently, if
\[
P^n \Pi_n(\Theta_n^C) \to 0 \quad \text{as } n \to \infty,
\]
then the Gibbs posterior has asymptotic concentration rate $\varepsilon_n \lor B_n t_n^{1-s}$ at $\theta^*$ with respect to the excess risk $d(\theta, \theta^*) = R(\theta) - R(\theta^*)$.

The setup here is more complicated than in previous sections, so some further explanation is warranted. First, we sketch out how Condition 3.1 leads to the critical property (24). A well-known bound on the expectation of a non-negative random variable, plus the moment bound in Condition 3.1, for $s > 1$, and Markov’s inequality leads to
\[
P_{\ell_\theta}(\ell_\theta > t_n) = \int_{t_n}^{\infty} P(\ell_\theta > x) \, dx \leq B_n \int_{t_n}^{\infty} x^{-s} \, dx = B_n t_n^{1-s}. \tag{27}
\]
This, in turn, implies $R(\theta) = P_{\ell_\theta} + O(B_n t_n^{1-s})$. If $B_n t_n^{1-s} \to 0$, then the difference between the risk and the clipped risk is vanishing, so we can bound $P_{\ell_\theta} - P_{\ell_\theta}^{t_n}$ by a multiple of the excess risk $R(\theta) - R(\theta^*)$ for all sufficiently large $n$. In the application we explore in Section 5.1, $B_n$ is related to the radius of the sieve $\Theta_n$, which grows logarithmically in $n$, while $t_n$ is related to the polynomial tail behavior of $\ell_\theta$, so that $B_n t_n^{1-s} \to 0$ happens naturally if $s > 1$. Additional details are given in the proof of Theorem 6, which can be found in the Supplementary Material.

Second, how might Condition 3.2 be checked? Start by defining the excess clipped risk $m_n(\theta, \theta^*_n) = P_{\ell_\theta}^{t_n} - P_{\ell_\theta}^{t_n}$, and the corresponding variance $v_n(\theta, \theta^*_n) = (P_{\ell_\theta}^{t_n} - P_{\ell_\theta}^{t_n})^2 - m_n(\theta, \theta^*_n)^2$. Now suppose it can be shown that there exists a function $G$ such that
\[
v_n(\theta, \theta^*_n) \leq G(\Delta_n) m_n(\theta, \theta^*_n), \quad \text{for all } \theta \in \Theta_n. \tag{17}
\]
$\Delta_n$ is the size index of the sieve. This amounts to the excess clipped loss satisfying a Bernstein condition (17), with exponent $\alpha = 1$. The excess clipped loss itself is $\leq t_n$ (and maybe substantially smaller, depending on form of $\ell_\theta$). So, we can apply the moment-generating function bound in (16) for bounded excess losses to get
\[
P e^{-\omega_n(\ell_\theta^{t_n} - \ell_\theta^{t_n})} < e^{-K_n \omega_n m_n(\theta, \theta^*_n)},
\]
where $K_n = \{1 - C \omega_n G(\Delta_n)\}$, for a constant $C > 0$, provided that $\omega_n t_n = O(1)$. Now it is easy to see that the above display implies Condition 3.2.

For the rate calculation with respect to the excess risk, the decomposition of $R(\theta)$ based on (27) implies we need $\varepsilon_n \geq B_n t_n^{1-s}$, subject to the constraint $n\omega_n K_n \varepsilon_n \to \infty$, for $K_n$ as above. As we see in Section 5.1, $B_n, \Delta_n,$ and $K_n$ can often be taken as powers of $\log n$, so the critical components determining the rate are $s$ and $t_n$. The optimal rate depends on the upper bound of the excess
clipped loss, which, in the worst case, equals $t_n$. To apply (16) we need the learning rate to vanish like the reciprocal of this bound, so we take $\omega_n = t_n^{-1}$. Then, we determine $\varepsilon_n$ to satisfy $nt_n^{-1} \varepsilon_n \to \infty$ and $\varepsilon_n \geq t_n^{1-s}$, up to a log term. The clipping sequence $t_n \approx n^{1/s}$ is sufficient, and yields the rate $\varepsilon_n \approx n^{1/s-1}$, modulo log terms.

5. Examples

This section presents several illustrations of the general theory presented in Sections 3 and 4. The strategies laid out in Sections 3.4, 4.1, and 4.2 are put to use in the following examples to verify our sufficient conditions for Gibbs posterior concentration. Sections 3 and 4 of the Supplementary Material include proofs of concentration results in this section as well as two additional illustrations.

5.1. Finite-dimensional regression with squared-error loss

Consider predicting a response $y \in \mathbb{R}$ using a linear function $x^T \theta$ by minimizing the sum of squared-error losses $\ell_\theta(u) = (y - x^T \theta)^2$, with $u = (x, y) \in \mathbb{R}^{J+1}$. Over a parameter space $\theta \in \Theta \subseteq \mathbb{R}^J$. Suppose the covariate-response variable pairs $(X_i, Y_i)$ are iid with $X$ taking values in a compact subset $X \subseteq \mathbb{R}^J$. To complement this example we present a more flexible, non-parametric regression problem in Section 5.2 below. For the current example, we focus on how the tail behavior of the response variable affects posterior concentration; see Assumptions 1a and 1b below.

5.1.1. Light-tailed response

When the response is sub-exponential so is the excess loss, and by the argument outlined in Section 3.4 we can verify Condition 1 for $d(\theta, \theta^*) = \|\theta - \theta^*\|_2$ and with $r = 2$. Then, the Gibbs posterior distribution concentrates at rate $n^{-1/2}$ as a consequence of Theorem 2.

Assumption 1a.

1. The response $Y$, given $x$, is sub-exponential with parameters $(\sigma^2, b)$ for all $x$;
2. $X$ is bounded and its marginal distribution is such that $PX X^T$ exists and is positive definite with eigenvalues bounded away from 0; and
3. the prior $\Pi$ has a density bounded away from 0 in a neighborhood of $\theta^*$.

Proposition 1a. If Assumption 1a holds, and the learning rate $\omega$ is a sufficiently small constant, then the Gibbs posterior concentrates at rate $\varepsilon_n = n^{-1/2}$ with respect to $d(\theta, \theta^*) = \|\theta - \theta^*\|_2$.

5.1.2. Heavy-tailed response

As discussed in Section 3.4, Condition 1 can be expected to hold when the response is light-tailed, but not when it is heavy-tailed. However, for a capped loss, $\ell_\theta^u = \ell_\theta \wedge t_n$, with increasing $t_n$, and for a suitable sieve on the parameter space, we can show concentration of the Gibbs posterior at the risk minimizer $\theta^*$ via the argument given in Section 4.2.

Assumption 1b. The marginal distribution of $Y$ admits $s > 2$ finite moments, i.e., $P|Y|^s < \infty$.

Define a sieve by $\Theta_n = \{\theta \in \mathbb{R}^J : \|\theta\|_2 < \Delta_n\}$ for an increasing sequence $\Delta_n$, e.g., $\Delta_n = \log n$. The moment condition in Assumption 1b implies three important properties of $\ell_\theta^u$ for $\theta \in \Theta_n$:...
The excess clipped loss satisfies $\sup_{y,x} |\ell_{\theta_n}(y,x) - \ell_{\theta_n^*}(y,x)| < \Delta_n t_n^{1/2}$.

- The risk and clipped risk are equivalent up to an error of order $\Delta_n^2 t_n^{1-s/2}$.

- The clipped loss satisfies $m_n(\theta, \theta_n^*) \vee v_n(\theta, \theta_n^*) \lesssim \Delta_n^2 \|\theta - \theta^*\|_2^2 + \Delta_n t_n^{1-s/2}$.

These three properties can be used to verify Condition 3 using the argument sketched out in Section 4.2 and to verify the prior bound in (26) required to apply Theorem 6.

**Proposition 1b.** Suppose Assumption 1b holds for some $s > 2$. Let $t_n = n^{2/(s-1)}$, $\Delta_n = \log n$, $\omega_n = \Delta_n^{-1} t_n^{-1/2}$, and $\varepsilon_n = \Delta_n^{s/2} n^{-(s-2)/(2s-2)}$. If Assumption 1a.2–3 also holds, then the Gibbs posterior with learning rate $\omega_n$ concentrates at rate $\varepsilon_n$ with respect to $d(\theta, \theta^*) = \|\theta - \theta^*\|_2$.

Proposition 1b continues to hold if we replace $s$ in the definitions of $t_n$ and $\varepsilon_n$ by $s'$ satisfying $2 < s' < s$. That means we only need an accurate lower bound for $s$ to construct a consistent Gibbs posterior for $\theta^*$, albeit with a slower concentration rate.

It is clear the clipping sequence may bias the clipped risk minimizer. For a given clip, the bias is less for light-tailed (large $s$) losses compared to heavy-tailed losses because a loss in excess of the clip is rarer for the former. This explains why the clipping sequence $t_n = n^{2/(s-1)}$ decreases in $s$.

Note that our $\Delta_n^2$ can be compared to the rate derived in [16, Example 11]. Indeed, up to log factors, their rate, $n^{-s/(s+2)}$, is smaller than ours for $s < 4$ but larger for $s > 4$. That is, their rate is slightly better when the response has between 2 and 4 moments, while ours is better when the response has 4 or more moments. Also, their result assumes that the parameter space is fixed and bounded, whereas we avoid this assumption with a suitably chosen sieve.

### 5.2. Mean regression curve

Let $Y_1, \ldots, Y_n$ be independent, where the marginal distribution of $Y_i$ depends on a fixed covariate $x_i \in [0, 1]$ through the mean, i.e., the expected value of $Y_i$ is $\theta^*(x_i)$, $i = 1, \ldots, n$. For simplicity, set $x_i = i/n$, corresponding to an equally-spaced design. Then the goal is estimation of the mean function $\theta^*: [0, 1] \to \mathbb{R}$, which resides in a specified function class $\Theta$ defined below.

A natural starting point is to define an empirical risk based on squared error loss, i.e.,

\[
R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \theta(x_i)\}^2.
\]  

However, any function $\theta$ that passes through the observations would be an empirical risk minimizer, so some additional structure is needed to make the solution to the empirical risk minimization problem meaningful. Towards this, as is customary in the literature, we parametrize the mean function as a linear combination of a fixed set of basis functions, $f(x) = (f_1(x), \ldots, f_J(x))^\top$. That is, we consider only functions $\theta = \theta_\beta$, where

\[
\theta_\beta(x) = \beta^\top f(x), \quad \beta \in \mathbb{R}^J.
\]  

Note that we do not assume that $\theta^*$ is of the specified form; more specifically, we do not assume existence of a vector $\beta^*$ such that $\theta^* = \theta_{\beta^*}$. The idea is that the structure imposed via the basis functions will force certain smoothness, etc., so that minimization of the risk over this restricted class of functions would identify a suitable estimate.
This structure changes the focus of our investigation from the mean function \( \theta \) to the \( J \)-vector of coefficients \( \beta \). We now proceed by first constructing a Gibbs posterior for \( \beta \) and then obtain the corresponding Gibbs posterior for \( \theta \) by pushing the former through the mapping \( \beta \to \theta_\beta \). In particular, define the empirical risk function in terms of \( \beta \):

\[
 r_n(\beta) = R_n(\theta_\beta) = \frac{1}{n} \sum_{i=1}^n \{Y_i - \theta_\beta(x_i)\}^2 = \frac{1}{n} (Y - F_n \beta)\trans (Y - F_n \beta),
\]

(30)

where \( \beta \in \mathbb{R}^J \) and where \( F_n \) is the \( n \times J \) matrix whose \((i,j)\) entry is \( f_j(x_i) \), assumed to be positive definite; see below. Given a prior distribution \( \Pi \) for \( \beta \)—which determines a prior \( \Pi \) for \( \theta \) through the aforementioned mapping—we can first construct the Gibbs posterior for \( \beta \) as in (5) with the pseudo-likelihood \( \beta \to \exp[-\omega r_n(\beta)] \). If we write \( \Pi_n \) for this Gibbs posterior for \( \beta \), then the corresponding Gibbs posterior for \( \theta \) is given by

\[
 \Pi_n(A) = \Pi_n(\{\beta : \theta_\beta \in A\}), \quad A \subseteq \Theta.
\]

(31)

Therefore, the concentration properties of \( \Pi_n \) are determined by those of \( \Pi_n \).

We can now proceed very much like we did before, but the details are slightly more complicated in the present iid case. Taking expectation with respect to the joint distribution of \((Y_1, \ldots, Y_n)\) is, as usual, the average of marginal expectations; however, since the data are not iid, these marginal expectations are not all the same. Therefore, the expected empirical risk function is

\[
 \bar{r}_n(\beta) = P^m r_n(\beta) = \frac{1}{n} \sum_{i=1}^n P_i \{Y_i - \theta_\beta(x_i)\}^2
\]

(32)

where \( P_i = P_{x_i} \) is the marginal distribution of \( Y_i \) and where \( F_{n,i} \) is the \( i \)th row of \( F_n \). Since the expected empirical risk function depends on \( n \), through \((x_1, \ldots, x_n)\), so too does the risk minimizer

\[
 \beta_n^\star = \arg \min_{\beta} \bar{r}_n(\beta).
\]

(33)

If \( P_i \) has finite variance, then \( \bar{r}_n(\beta) \) differs from \( \{\theta^\star(x_{1:n}) - F_n \beta\}^2 \) by only an additive constant not depending on \( \beta \), and this becomes a least-squares problem, with solution

\[
 \beta_n^\star = (F_n\trans F_n)^{-1} F_n\trans \theta^\star(x_{1:n}),
\]

(34)

where \( \theta^\star(x_{1:n}) \) is the \( n \)-vector \((\theta^\star(x_1), \ldots, \theta^\star(x_n))\)\trans. Our expectation is that the Gibbs posterior \( \Pi_n \) for \( \beta \) will suitably concentrate around \( \beta_n^\star \), which implies that the Gibbs posterior \( \Pi_n \) for \( \theta \) will suitably concentrate around \( \theta_\beta^\star \). Finally, if the above holds and the basis representation is suitably flexible, then \( \theta_\beta^\star \) will be close to \( \theta^\star \) in some sense and, hence, we achieve the desired concentration.

The flexibility of the basis representation depends on the dimension \( J \). Since \( \theta^\star \) need not be of the form \( \theta_\beta \), a good approximation will require that \( J = J_n \) be increasing with \( n \). How fast \( J = J_n \) must increase depends on the smoothness of \( \theta^\star \). Indeed, if \( \theta^\star \) has smoothness index \( \alpha > 0 \) (made precise below), then many systems of basis functions—including Fourier series and B-splines—have the following approximation property: there exists an \( H > 0 \) such that for every \( J \)

\[
 \text{there exists } \beta \in \mathbb{R}^J \text{ such that } \|\beta\|_\infty < H \text{ and } \|\theta_\beta - \theta^\star\|_\infty \lesssim J^{-\alpha}.
\]

(35)
Gibbs posterior concentration rates

Then the idea is to set the approximation error in (35) equal to the target rate of convergence, which depends on \(n\) and on \(\alpha\), and then solve for \(J = J_n\).

For Gibbs posterior concentration at or near the optimal rate, we need the prior distribution for \(\beta\) to be sufficiently concentrated in a bounded region of the \(J\)-dimensional space in the sense that

\[
\tilde{\Pi}(\{\beta : \|\beta - \beta'\|_2 \leq \varepsilon\}) \gtrsim (C\varepsilon)^J, \quad \forall \beta' \in \mathbb{R}^J \quad \text{with} \quad \|\beta'\|_\infty \leq H,
\]

for the same \(H\) as in (35), for a small constant \(C > 0\), and for all small \(\varepsilon > 0\).

**Assumption 2.**

1. The function \(\theta^* : [0, 1] \to \mathbb{R}\) belongs to a class \(\Theta = \Theta(\alpha, L)\) of Hölder smooth functions parametrized by \(\alpha \geq 1/2\) and \(L > 0\). That is, \(\theta^*\) satisfies

\[
|\theta^*[\alpha](x) - \theta^*[\alpha](x')| \leq L|x - x'|^{\alpha - [\alpha]}, \quad \forall x, x' \in [0, 1],
\]

where the superscript "(k)" means \(k\)th derivative and \([\alpha]\) is the integer part of \(\alpha\);

2. for a given \(x\), the response \(Y\) is sub-Gaussian and with variance and variance proxy—both of which can depend on \(x\)—uniformly upper bounded by \(\sigma^2\);

3. the eigenvalues of \(F_n^\top F_n\) are bounded away from zero and \(\infty\);

4. the approximation property (35) holds; and

5. the prior for \(\beta\) satisfies (36) and has a bounded density on the \(J_n\)-dimensional parameter space.

The bounded variance assumption is implied, for example, if the variance of \(Y\) is a smooth function of \(x\) in \([0, 1]\), which is rather mild. And assuming the eigenvalues of \(F_n^\top F_n\) are bounded is not especially strong since, in many cases, the basis functions would be orthonormal. In that case, the diagonal and off-diagonal entries of \(F_n^\top F_n\) would be approximately 1 and 0, respectively, and the bounds are almost trivial. The conditions on the prior distribution are weak; as we argue in the proof, it can be satisfied by taking the joint prior density to be the product of \(J\) independent prior densities on the components of \(\beta\), and where each component density is strictly positive.

**Proposition 2.** If Assumption 2 holds, and the learning rate \(\omega\) is a sufficiently small constant, then the Gibbs posterior \(\Pi_n\) for \(\theta\) concentrates at \(\theta^*\) with rate \(\varepsilon_n = n^{-\alpha/(1+2\alpha)}\) with respect to the empirical \(L_2\) norm \(\|\theta - \theta^*\|_{n, 2}\), where \(\|f\|_{n, 2}^2 = n^{-1}\sum_{i=1}^n f^2(x_i)\).

We should emphasize that the quantity of interest, \(\theta\), is high-dimensional, and the rate \(\varepsilon_n\) given in Proposition 2 is optimal for the given smoothness \(\alpha\); there are not even any nuisance logarithmic terms.

The simpler fixed-dimensional setting with constant \(J\) can be analyzed similarly as above. In that case, we can simultaneously weaken the requirement on the response \(Y\) in Assumption 2.2 from sub-Gaussian to sub-exponential, and strengthen the conclusion to a root-\(n\) concentration rate.

### 5.3. Binary classification

Let \(Y \in \{0, 1\}\) be a binary response variable and \(X = (X_0, X_1, \ldots, X_q)\) a \((q + 1)\)-dimensional predictor variable. We consider classification rules of the form

\[
\phi_q(X) = 1\{X^\top \theta > 0\} = 1\{\alpha X_0 + (X_1, \ldots, X_q)^\top \beta > 0\}, \quad \theta = (\alpha, \beta) \in \mathbb{R}^{q+1},
\]

and the goal is to learn the optimal \(\theta\) vector, i.e., \(\theta^* = \arg\min_{\theta} R(\theta)\), where \(R(\theta) = P\{Y \neq \phi_q(X)\}\) is the misclassification error probability, and \(P\) is the joint distribution of \((X, Y)\). This optimal \(\theta^*\) is
such that $\eta(x) > \frac{1}{2}$ if $x^\top \theta^* > 0$ and $\eta(x) < \frac{1}{2}$ if $x^\top \theta^* \leq 0$, where $\eta(x) = P(Y = 1 \mid X = x)$ is the conditional probability function. Below we construct a Gibbs posterior distribution that concentrates around this optimal $\theta^*$ at rate that depends on certain local features of that $\eta$ function.

Suppose our data consists of iid copies $(X_i, Y_i)$, $i = 1, \ldots, n$, of $(X,Y)$ from $P^*$, and define the empirical risk function

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \neq \phi_\theta(X_i)\}. \quad (38)$$

In addition to the empirical risk function we need specify a prior II and here the prior plays a significant role in the Gibbs posterior concentration results.

A unique feature of this problem, which makes the prior specification a little different than in a linear regression problem, is that the scale of $\theta$ does not affect classification performance, e.g., replacing $\theta$ with $1000\theta$ gives exactly the same classification performance. To fix a scale, we follow [23], and

- assume that the $x_0$ component of $x$ is of known importance and always included in the classifier,
- and constrain the corresponding coefficient, $\alpha$, to take values $\pm 1$.

This implies that the $\alpha$ and $\beta$ components of the $\theta$ vector should be handled very differently in terms of prior specification. In particular, $\alpha$ is a scalar with a discrete prior—which we take here to be uniform on $\pm 1$—and $\beta$, being potentially high-dimensional, will require setting-specific prior considerations.

The characteristic that determines the difficulty of a classification problem is the distribution of $\eta(X)$ or, more specifically, how concentrated $\eta(X)$ is near the value $\frac{1}{2}$, where one could do virtually no worse by classifying according to a coin flip. The set $\{x : \eta(x) = \frac{1}{2}\}$ is called the margin, and conditions that control the concentration of the distribution of $\eta(X)$ around $\frac{1}{2}$ are generally called margin conditions. Roughly, if $\eta$ has a “jump” or discontinuity at the margin, then classification is easier and $\eta(X)$ does not have to be so tightly concentrated around $\frac{1}{2}$. On the other hand, if $\eta$ is smooth at the margin, then the classification problem is more challenging in the sense that more data near the margin is needed to learn the optimal classifier, hence, tighter concentration of $\eta(X)$ near $\frac{1}{2}$ is required.

In Sections 5.3.1 and 5.3.2 that follow, we consider two such margin conditions, namely, the so-called Massart and Tsybakov conditions. The first is relatively strong, corresponding to a jump in $\eta$ at the margin, and the result we we establish in Proposition 3a is accordingly strong. In particular, we show that the Gibbs posterior achieves the optimal and adaptive concentration rate in a class of high-dimensional problems ($q \gg n$) under a certain sparsity assumption on $\theta^*$. The Tsybakov margin condition we consider below is weaker than the first, in the sense that $\eta$ can be smooth near the “$\eta = \frac{1}{2}$” boundary and, as expected, the Gibbs posterior concentration rate result is not as strong as the first.

### 5.3.1. Massart’s noise condition

Here we allow the dimension $q + 1$ of the coefficient vector $\theta = (\alpha, \beta)$ to exceed the sample size $n$, i.e., we consider the so-called high-dimensional problem with $q \gg n$. Accurate estimation and inference is not possible in high-dimensional settings without imposing some low-dimensional structure on the inferential target, $\theta^*$. Here, as is typical in the literature on high-dimensional inference, we assume that $\theta^*$ is sparse in the sense that most of its entries are exactly zero, which corresponds to most of the predictor variables being irrelevant to classification. Below we construct a Gibbs posterior distribution for $\theta$ that concentrates around the unknown sparse $\theta^*$ at a (near) optimal rate.

Since the sparsity in $\theta^*$ is crucial to the success of any statistical method, the prior needs to be chosen carefully so that sparsity is encouraged in the posterior. The prior II for $\theta$ will treat $\alpha$ and $\beta$ independent, and the prior for $\beta$ will be defined hierarchically. Start with the reparametrization $\beta \to (S, \beta_S)$, where
Assumption 3a.

1. The marginal distribution of $X$ is compactly supported, say, on $[-1, 1]^{q+1}$.
2. The conditional distribution of $X_0$, given $X = (X_1, \ldots, X_q)$, has density with respect to Lebesgue measure that is uniformly bounded.
3. The rate parameter $\lambda$ in the Laplace prior satisfies $\lambda \lesssim (\log q)^{1/2}$.
4. The optimal $\theta^* = (\alpha^*, \beta^*)$ is sparse in the sense that $|S^*| \log q = o(n)$, where $S^*$ is the configuration of non-zero entries in $\beta^*$, and $\|\beta^*\|_\infty = O(1)$.
5. There exists $h \in (0, \frac{1}{2})$ such that $P(|\eta(X) - \frac{1}{2}| \leq h) = 0$.

The first two parts of Assumption 3a correspond to Conditions 0' and 0'' in [23], and Assumption 3a (5) above is precisely the margin condition imposed in Equation (5) of [35]; see, also, [31] and [27]. This concisely states that there is either a jump in the function at the margin or that the marginal distribution of $X$ is not supported near the margin; in either case, there is separation between the $Y = 0$ and $Y = 1$ cases, which makes the classification problem relatively easy. With these assumptions, we get the following Gibbs posterior asymptotic concentration rate result. Note that, in order to preserve the high-dimensionality in the asymptotic limit, we let the dimension $q = q_n$ increase with the sample size $n$. So, the data sequence actually forms a triangular array but, as is common in the literature, we suppress this formulation in our notation.

**Proposition 3a.** Consider a classification problem as described above, with $q \gg n$. Under Assumption 3a, the Gibbs posterior, with sufficiently small constant learning rate, concentrates at rate

$$
\varepsilon_n = (n^{-1}|S^*| \log q)^{1/2}
$$

with respect to $d(\theta, \theta^*) = \{R(\theta) - R(\theta^*)\}^{1/2}$.

This result shows that, even in very high dimensional settings, the Gibbs posterior concentrates on the optimal rule $\theta^*$ at a fast rate. For example, suppose that the dimension $q$ is polynomial in $n$, i.e., $q \sim n^b$ for any $b > 0$, while the “effective dimension,” or complexity, is sub-linear, i.e., $|S^*| \sim n^a$ for $a < 1$. Then we get that $\{\theta : R(\theta) - R(\theta^*) \lesssim n^{(1-a)} \log n\}$ has Gibbs posterior probability converging to 1 as $n \to \infty$. That is, rates better than $n^{-1/2}$ can easily be achieved, and even arbitrarily close to $n^{-1}$ is possible. Compare this to the rates in Propositions 2–3 in [23], also in terms of risk difference, that cannot be faster than $n^{-1/2}$. Further, the concentration rate in Proposition 3a is nearly the optimal rate corresponding to an oracle who has knowledge of $S^*$. That is, the Gibbs posterior concentrates at nearly the optimal rate adaptively with respect to the unknown complexity.
5.3.2. Tsybakov’s margin condition

Next, we consider classification under the more general Tsybakov margin condition [e.g., 44]. The problem set up is the same as above, except that here we consider the simpler low-dimensional case, with the number of predictors \((q + 1)\) small relative to \(n\). Since the dimension is no longer large, prior specification is much simpler. We will continue to assume, as before, that the \(x_0\) component of \(x\) has a constrained coefficient \(\alpha \in \{\pm 1\}\), to which we assign a discrete uniform prior. Otherwise, we simply require the prior \(\Pi\) have a (marginal) density, \(\pi\), for \(\beta\), with respect to Lebesgue measure on \(\mathbb{R}^q\).

Assumption 3b.

1. The marginal prior density for \(\beta\) is continuous and bounded away from 0 near \(\beta^*\).
2. There exists \(c > 0\) and \(\gamma > 0\) such that \(P(|2\eta(X) - 1| \leq h) \leq ch^\gamma\) for all sufficiently small \(h > 0\).

The concentration of the marginal distribution of \(\eta(X)\) around \(\frac{1}{2}\) controls the difficulty of the classification problem, and Condition 3b.2 concerns exactly this. Note that smaller \(\gamma\) implies \(\eta(X)\) is less concentrated around \(\frac{1}{2}\), so we expect our Gibbs posterior concentration rate, say, \(\varepsilon_n = \varepsilon_n(\gamma)\), to be a decreasing function of \(\gamma\). The following proposition confirms this.

**Proposition 3b.** Suppose Assumption 3b holds and, for the specified \(\gamma > 0\), let
\[
\varepsilon_n = (\log n)^\gamma/(2+2\gamma) n^{-\gamma/(3+2\gamma)}.
\]

Then the Gibbs posterior distribution, with learning rate \(\omega_n = \varepsilon_n^{1/\gamma}\), concentrates at rate \(\varepsilon_n\) with respect to \(d(\theta, \theta^*) = \{R(\theta) - R(\theta^*)\}^{1/2}\).

Note that Massart’s condition from Section 5.3.1 corresponds to Tsybakov’s condition above with \(\gamma = \infty\). In that case, the Gibbs posterior concentration rate we recover from Proposition 3b is \(\varepsilon_n = (\log n)^{1/2} n^{-1/2}\), which is achieved with a suitable constant learning rate. This is within a logarithmic factor of the optimal rate for finite-dimensional problems. Moreover, for both the \(\gamma < \infty\) and \(\gamma = \infty\) cases, we expect that the logarithmic factor could be removed following an approach like that described in Theorem 2, but we do not explore this possible extension here.

We should emphasize that this case is unusual because the learning rate \(\omega_n\) depends on the (likely unknown) smoothness exponent \(\gamma\). This means the rate in Proposition 3b is not adaptive to the margin. However, this dependence is not surprising, as it also appears in [16, Section 6]. The reason the learning rate depends on \(\gamma\) is that the Tsybakov margin condition in Assumption 3b.2 implies the Bernstein condition in (17) takes the form
\[
v(\theta, \theta^*) \lesssim m(\theta, \theta^*) \gamma/\gamma \gamma/\gamma/\gamma^{1+\gamma}.
\]

Therefore, in order to verify Condition 1 using the strategy in Section 3.4, we need \(\omega_n m(\theta, \theta^*)\) and \(\omega_n^2 v(\theta, \theta^*)\) to have the same order when \(d(\theta, \theta^*) \geq M_n \varepsilon_n\). This requires that the learning rate depends on \(\gamma\), in particular, \(\omega_n = \varepsilon_n^{1/\gamma}\).

5.4. Quantile regression curve

The \(\tau^{th}\) conditional quantile function \(\theta^*(x)\) of a response \(Y\) given a covariate \(X = x\) is modeled by a linear combination of basis functions \(f(x) = (f_1(x), \ldots, f_J(x))^T\):

\[Q_{Y|X=x}(\tau) = \beta^T f(x), \quad \beta \in \mathbb{R}^J.\]
In Section 5.2 we regarded the degree of smoothness, \( \alpha \), of \( \theta^* \) to be known and we chose the number of basis functions accordingly to achieve the optimal concentration rate. But, in practice, \( \alpha \) may not be known, which creates a challenge because, as mentioned, the number of terms needed in the basis function expansion modeling \( \theta^* \) depends on this unknown degree of smoothness. To achieve optimal concentration rates adaptive to unknown smoothness, the choice of prior is crucial. In particular, the prior must support a very large model space in order to guarantee it places sufficient mass near \( \theta^* \). Our approximation of \( \theta^* \) by a linear combination of basis functions suggests a hierarchical prior for \( \theta \equiv (J, \beta_J) \), similar to Section 5.3.1, with a marginal prior \( \pi \) for the number of basis functions \( J \) and a conditional prior \( \Pi_J \) for \( \beta_J \), given \( J \). The resulting prior for \( \theta \) is given by a mixture,

\[
\Pi(A) = \sum_{j=1}^{\infty} \pi(j) \Pi_j(\{\beta_j \in \mathbb{R}^j : \beta_j^T f \in A\}), \quad A \subseteq \Theta.
\]  

Then, in order for \( \Pi \) to place sufficient mass near \( \theta^* \), it is sufficient the marginal and conditional priors satisfy the following conditions: the marginal prior \( \pi \) for \( J \) satisfies for some \( c_1 > 0 \) for every \( J = j \)

\[
\pi(j) \geq e^{-c_1 j \log j};
\]

and, the conditional prior \( \Pi \) for \( \beta_J \) given \( J = j \) satisfies for every \( j \)

\[
\Pi(\{\beta : \|\beta - \beta'\|_2 \leq \epsilon\}) \geq e^{-C_j \log(1/\epsilon)}, \quad \text{for all } \beta' \in \mathbb{R}^J \text{ with } \|\beta'\|_\infty \leq H,
\]

for the same \( H \) as in (35) and for some constant \( C > 0 \) for all sufficiently small \( \epsilon > 0 \). Fortunately, many simple choices of \( (\pi, \Pi_J) \) are satisfactory for obtaining adaptive concentration rates, e.g., a Poisson prior on \( J \) and a \( J \)-dimensional normal conditional prior for \( \beta \), given \( J \); see Conditions (A1), (A2), and Remark 1 in [38]. Besides the conditions in (40) and (41) we need to make a minor modification of \( \Pi \) to make it suitable for our proof of Gibbs posterior concentration; see below.

We link the data and parameter through the check loss function

\[
\ell_\theta(u) = \frac{1}{2}(|\theta(x) - y| - |y|) + (1 - \tau)\theta(x),
\]

where \( \theta(x) = \beta^T f(x) \). See [26] for a proof that \( P \ell_\theta \) is minimized at \( \theta^* \). It is straightforward to show the check loss \( \theta \rightarrow \ell_\theta(u) \) is \( L \)-Lipschitz with \( L < 1 \), so, if \( Y \) were bounded we could use Condition 1 to find the concentration rate. However, to handle an unbounded response we need the flexibility of Condition 2 and Theorem 5. To verify (22), we found it necessary to limit the complexity of the parameter space by imposing a constraint on the prior distribution, namely that the sequence of prior distributions places all its mass on the set \( \Theta_n := \{\theta : \|\theta\|_\infty \leq \Delta_n\} \) for some \( \Delta_n \rightarrow \infty \); see Assumption 4.4. This constraint implies the prior depends on \( n \), and we refer to this sequence of priors by \( \Pi(n) \). Given the hierarchical prior \( \Pi \) in (39) one straightforward way to define a sequence of prior distributions satisfying the constraint is to restrict and renormalize \( \Pi \) to \( \Theta_n \), i.e., define \( \Pi(n) \) as

\[
\Pi(n)(A) = \Pi(A)/\Pi(\Theta_n), \quad A \subseteq \Theta \cap \Theta_n
\]

This particular \( \Pi(n) \) in (42) is not the only way to force a restriction to \( \Theta_n \), but it is convenient. That is, if \( \Pi \) places mass \( \eta \) on a sup-norm neighborhood of \( \theta^* \) (see the proof of Proposition 4), then, by construction, \( \Pi(n) \) in (42) places at least mass \( \eta \) on the same neighborhood.

We should emphasize that restricting the prior to \( \Theta_n \) is only a technical requirement needed for the proof, but it is not unreasonable. Since \( \theta^* \) is bounded, it is eventually in the growing support of the prior \( \Pi \). Similar assumptions appear in the literature on quantile regression; for example, Theorem 6 in [43] requires that the parameter space consists only of bounded functions.
Assumption 4.

1. The function \( \theta^* : \mathbb{X} \rightarrow \mathbb{R} \) is Hölder smooth with parameters \((\alpha, L)\) (see Assumption 2.1);
2. the basis functions satisfy the approximation property in (35);
3. the covariate space \( \mathbb{X} \) is compact and there exists a \( \delta > 0 \) such that the conditional density of \( Y \), given \( X = x \), is continuous and bounded away from 0 by a constant \( \beta > 0 \) in the interval \( (\theta^*(x) - \delta, \theta^*(x) + \delta) \) for every \( x \); and,
4. the sequence \( \Pi^{(n)} \) of prior distributions satisfies (42) for a sequence of subsets of the parameter space \( \Theta_n := \{ \theta : \|\theta\|_{\infty} \leq \Delta_n \} \) for some sequence \( \Delta_n > 0 \), for \( \Pi \) as defined in (39), and for marginal and conditional priors \((\pi, \tilde{\Pi})\) for \( J \) and \( \beta_J \) given \( J = j \) that satisfy (40) and (41).

Proposition 4. Define \( \varepsilon_n = (\log n)^{1/2} \Delta_n^2 n^{-\alpha/(1+2\alpha)} \). If the learning rate satisfies \( \omega_n = c \Delta_n^{-2} \) for some \( 0 < c < 1/2 \) and Assumption 4 holds, then the Gibbs posterior distribution concentrates at rate \( \varepsilon_n \) with respect to \( d(\theta, \theta^*) = \|\theta - \theta^*\|_{L_2(P)} \).

Since the mathematical statement does not give sufficient emphasis to the adaptation feature, we should follow up on this point. That is, \( n^{-\alpha/(2\alpha+1)} \) is the optimal rate [38] for estimating an \( \alpha \)-Hölder smooth function, and Proposition 4 says that the Gibbs posterior achieves nearly the optimal rate adaptively in the sense that it concentrates at nearly the optimal rate as if \( \alpha \) were known.

The concentration rate in Proposition 4 depends on the complexity of the parameter space as determined by \( \Delta_n \) in Assumption 4(3). For example, if the sup-norm bound on \( \theta^* \) were known, then \( \Delta_n \) and the learning rate \( \omega_n \) could be taken as constants and the rate would be optimal up to a \( (\log n)^{1/2} \) factor. On the other hand, if greater complexity is allowed, e.g., \( \Delta_n = (\log n)^p \) for some power \( p > 0 \), then the concentration rate takes on an additional \( (\log n)^{2p} \) factor, which is not a serious concern.

5.5. Personalized MCID

In the medical sciences, physicians who investigate the efficacy of new treatments are challenged to determine both statistically and practically significant effects. In many applications some quantitative effectiveness score can be used for assessing the statistical significance of the treatment, but physicians are increasingly interested also in patients’ qualitative assessments of whether they believed the treatment was effective. The aim of the approach described below is to find the cutoff on the effectiveness score that best separates patients by their reported outcomes. That cutoff value is called the minimum clinically important difference, or MCID.

For this application we follow up on the MCID problem discussed in [40] with a covariate-adjusted, or personalized, version. In medicine, there is a trend away from the classical “one size fits all” treatment procedures, to treatments that are tailored more-or-less to each individual. Along these lines, naturally, doctors would be interested to understand how that threshold for practical significance depends on the individual, hence there is interest in a so-called personalized MCID [19, 51].

Let the data \( U^n = (U_1, \ldots, U_n) \) be iid \( P \), where each observation is a triple \( U_i = (X_i, Y_i, Z_i) \) denoting the patient’s diagnostic measurement, their self-reported effectiveness outcome \( Y_i \in \{-1, 1\} \), and covariate value \( Z_i \in \mathbb{Z} \subseteq \mathbb{R}^q \), for \( i = 1, \ldots, n \) and \( q \geq 1 \). In practice the diagnostic measurement has something to do with the effectiveness of the treatment so one can imagine examples including blood pressure, blood glucose level, and viral load. Examples of covariates include a patient’s age, weight, and gender. The idea is that the \( z \)-scale cutoff for practical significance would depend on the covariate \( z \), hence the MCID is a function, say, \( \theta(z) \), and the goal is to learn this function.

The true MCID \( \theta^* \) is defined as the solution to an optimization problem. That is, if

\[
\ell_\theta(x, y, z) = \frac{1}{2}[1 - y \text{sign}\{x - \theta(z)\}], \quad (x, y, z) \in \mathbb{R} \times \{-1, 1\} \times \mathbb{Z},
\] (43)
then the expected loss is $R(\theta) = P[Y \neq \text{sign}(X - \theta(Z))]$, and the true MCID function is defined as the minimizer $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$, where the minimum is taken over a class $\Theta$ of functions on $\mathbb{Z}$. Alternatively, as in Section 5.3, the true $\theta^*$ satisfies $\eta_2(x) > \frac{1}{2}$ if $x > \theta^*(z)$ and $\eta_2(x) \leq \frac{1}{2}$ if $x \leq \theta^*(z)$, where $\eta_2(x) = P(Y = 1 \mid X = x, \theta = z)$ is the conditional probability function.

As described in Section 2, the Gibbs posterior distribution is based on an empirical risk function which, in the present case, is given by

$$R_n(\theta) = \frac{1}{2n} \sum_{i=1}^{n} [1 - Y_i \text{sign}(X_i - \theta(Z_i))], \quad \theta \in \Theta. \quad (44)$$

In order to put this theory into practice, it is necessary to give the function space $\Theta$ a lower-dimensional parametrization. In particular, we consider a true MCID function $\theta^*$ belonging to a Hölder class as in Assumption 2 but with unknown smoothness, as in Section 5.4. And, we model $\theta^*$ by a linear combination of basis functions $\theta(z) = \theta_{J,\beta}(z) := \sum_{j=1}^{J} \beta_j f_j(z)$, for basis functions $f_j, j = 1, \ldots, J$. Then, each $\theta$ is identified by a pair $(J, \beta)$ consisting of a positive integer $J$ and a $J$-dimensional coefficient vector $\beta$. We use cubic B-splines in the numerical examples in Section 2 of the supplementary material, but any basis capable of approximating $\theta^*$ will work, and see (35).

The prior setup is similar to that in Section 5.4, (39). That is, the prior is specified hierarchically with a marginal prior $\pi_J \Pi$ for $J$ and a suitable conditional prior $\Pi_J$ for $\beta_j$, given $J$. And, as mentioned before, very simple choices of the marginal and conditional priors achieve the desired adaptive rates.

Assumption 5 below concerns the smoothness of $\theta^*$ and requires the chosen basis satisfy the approximation property used previously; it also refers to the same mild assumptions on random series priors used in Section 5.4 sufficient to ensure adequate prior mass is assigned to a neighborhood of $\theta^*$; finally, it assumes a margin condition on the classifier like that used in Section 5.3.1 and Assumption 3a.(5). These conditions are sufficient to establish a Gibbs posterior concentration rate.

**Assumption 5.**

1. The true MCID function $\theta^*: \mathbb{Z} \rightarrow \mathbb{R}$ for a compact subset $\mathbb{Z}$ of $\mathbb{R}$ and $\theta^*$ is Hölder smooth with parameters $(\alpha, \ell)$ (see Assumption 2.1);
2. the basis functions satisfy the approximation property in (35);
3. the prior distribution $\Pi_J$ for $\theta$ is defined hierarchically as in (39) with marginal and conditional priors $(\pi, \Pi_J)$ for $J$ and $\beta_j$ given $J = j$ that satisfy (40) and (41); and,
4. there exists $h \in (0, 1)$ such that $P\{\|2n\eta_2(X) - 1\| \leq h\} = 0$; and,
5. the conditional distribution, $P_\theta$, of $X$, given $Z = z$, has a density with respect to Lebesgue measure that is uniformly bounded away from infinity.

**Proposition 5.** Suppose Assumption 5 holds, with $\alpha$ as defined there, and set $\varepsilon_n = (\log n)n^{-\alpha/(1+\alpha)}$. For any fixed $\omega > 0$ the Gibbs posterior concentrates at rate $\varepsilon_n$ with respect to the divergence

$$d(\theta, \theta^*) = P\{\theta(Z) \land \theta^*(Z) \leq X \leq \theta(Z) \lor \theta^*(Z)\}$$

$$= \int_{Z} \int_{\{\theta(z) \land \theta^*(z)\}} P_\theta(dx) P(dz). \quad (45)$$

The Gibbs posterior distribution we have defined for the personalized MCID function achieves the concentration rate in Proposition 5 adaptively to the unknown smoothness $\alpha$ of $\theta^*$. Mammen and Tsybakov [30] consider estimation of the boundary curve of a set, and they show that the minimax optimal rate is $n^{-\alpha/(\alpha+1)}$ when the boundary curve is $\alpha$-Hölder smooth and distance is measured
by the Lebesgue measure of the set symmetric difference. In our case, if \((X, Z)\) has a joint density, bounded away from 0, then our divergence measure \(d(\theta, \theta^*)\) is equivalent to

\[
\text{Leb}(\{(x, z) : x \leq \theta(z)\} \triangle \{(x, z) : x \leq \theta^*(z)\}),
\]

in which case our rate is within a logarithmic factor of the minimax optimal rate.

Hedayat et al. [19] also study the personalized MCID and derive a convergence rate for an \(M\)-estimator of \(\theta^*\) based on a smoothed and penalized version of (44). It is difficult to compare our result with theirs, for instance, because their rate depends on two user-controlled sequences related to the smoothing and penalization of their loss. But, as mentioned above, our rate is near optimal in certain cases, so the asymptotic results in [19] cannot be appreciably better than our rate in Proposition 5.

6. Concluding remarks

In this paper we focus on developing some simple, yet general, techniques for establishing asymptotic concentration rates for Gibbs posteriors. A key take-away is that the robustness to model misspecification offered by the Gibbs framework does not come at the expense of slower concentration rates. Indeed, in the examples presented here—and others presented elsewhere [42]—the rates achieved are the same as those achieved by traditional Bayesian posteriors and (nearly) minimax optimal. Another main point is that Gibbs posterior distributions are not inherently challenging to analyze; on the contrary, the proofs presented herein are concise and transparent. An additional novelty to the analysis presented here is that we consider cases where the learning rate can be non-constant, i.e., either a vanishing sequence or data-dependent, and prove corresponding posterior concentration rate results.

Here the focus has been on deriving Gibbs posteriors with the best possible concentration rates, and selection of learning rates has proceeded with these asymptotic properties in mind. In other works [41], random learning rates are derived for good uncertainty quantification in finite-samples. We conjecture, however, that the two are not mutually-exclusive, and that learning rates arising as solutions to the calibration algorithm in [41] also have the desirable concentration rate properties. Proving this conjecture seems challenging, but Section 3.3 provides a first step in this direction.

Acknowledgments

The authors sincerely thank the reviewers for their helpful feedback on previous versions of the manuscript. This work is partially supported by the U.S. National Science Foundation, DMS–1811802.

References

Gibbs posterior concentration rates


