Chapter 4

Bayesian Inconsistency: Explanations and Discussion

In this chapter, we give several explanations of how the Bayesian inconsistency seen in Chapter 4 may occur under 'bad' misspecification, and why SafeBayes provides a solution to this problem. We also discuss how our inconsistency example and the SafeBayes method relate to other work.

4.1 Bayes' behaviour explained

In this section we explain how anomalous behaviour of the Bayesian posterior may arise, taking a frequentist perspective. Section 4.1.1 is merely provided to give some initial intuition and may be skipped. The proof of Theorem 4.1 is given in Appendix 4.A.2.

4.1.1 Explanation I: Variance issues

Example 4.A. [Bernoulli] Consider the following very simple scenario: our 'model' consists of two Bernoulli distributions, $\mathcal{M} = \{P_\theta \mid \theta \in \{0.2, 0.8\}\}$, with P_θ expressing that $Y_1, Y_2, \ldots \sim$ i.i.d. $\textsc{Ber}(\theta)$. We perform Bayesian inference based on a uniform prior on $\mathcal{M}$. Suppose first that the data are, in fact, sampled i.i.d. from P_{θ^*}, where θ^* is the 'true' parameter. The model is misspecified, in particular we will take a $\theta^* \notin \{0.2, 0.8\}$. The log-likelihood ratio between the two distributions for data Y^n with n_1 ones and $n_0 = n - n_1$ zeroes, measured for convenience in bits (base 2), is given by

$$L = \log_2 \frac{f_{0.8}(Y^n)}{f_{0.2}(Y^n)} = \log_2 \frac{(0.8)^{n_1}(0.2)^{n_0}}{(0.2)^{n_1}(0.8)^{n_0}} = 2(n_1 - n_0). \tag{4.1}$$

With uniform priors, the posterior will prefer $\theta = 0.2$ as soon as $L < 0$.

First suppose $\theta^* = 1/2$. Then both distributions in $\mathcal{M}$ are equally far from θ^* in terms of KL divergence (or any other commonly used measure). By the

central limit theorem, however, we expect that the probability that $|L| > \sqrt{n}/2$ is larger than a constant for all large n; in this particular case we numerically find that, for all n, it is larger than 0.32.

This implies that, at each n, $\min_{\theta \in \{0.2,0.8\}} \pi(\theta \mid Y^n) \approx 2^{-\sqrt{n}/2}$ with 'true' probability at least 0.32. Thus, there is a nonnegligible 'true' probability that the posterior on one of the two distributions is negligibly small, and a naive Bayesian who adopted such a model would be strongly convinced that the other distribution would be better even though both distributions are equally bad. While this already indicates that strange things may happen under misspecification, we are of course more interested in the situation in which $\theta^* \neq 1/2$, so that one of the two distributions in $\mathcal{M}$ is truly 'better'. Now, if the 'true' parameter θ^* is within $O(1/\sqrt{n})$ of $1/2$, then, by the central limit theorem, the probability that $L < 0$ is nonnegligible. For example, if θ^* is exactly $1/2 + 1/\sqrt{n}$, then this probability is larger than 0.16 for all n. Thus, for values of θ^* this close to $1/2$, there is no way we can even expect Bayes to learn the 'best' value. For fixed (independent of n), larger values of θ^*, like 0.6, the posterior will concentrate at 0.8 at an exponential rate, but the sample size at which concentration starts is considerably larger than the sample sized needed when the true parameter is in fact 0.8. For example, at $n = 50$, $P_{0.6}(L < 0) \approx 0.1$, $P_{0.8}(L < 0) \approx 2 \cdot 10^{-5}$; both probabilities go to 0 exponentially fast but their ratio increases exponentially with n. So, under a fixed θ^*, with increasing n, Bayes may take longer to concentrate on the best $\tilde{\theta}$ if $\tilde{\theta} \neq \theta^*$ (misspecification) than if $\tilde{\theta} = \theta^*$, but it eventually 'recovers' (this was seen in the ridge experiments of Section 3.5.4). Now, for larger models, the consequence of slower concentration of the log-likelihood ratio L is that the probability that *some* 'bad' P_θ happens to 'win' is substantially larger than with a correct model. Grünwald and Langford (2007) showed that, in a classification context with an infinite-dimensional model, there are so many of such 'bad' P_θ that Bayes does not recover any more, and the posterior keeps putting most of its mass on a bad model for ever (although the particular bad model on which it puts its mass keeps changing). In Chapter 3 we empirically showed the same in a regression problem.

Now one might conjecture that the issues above are caused by the fact that the model $\mathcal{M}$ is 'disconnected'. In the Bernoulli example above, the problem indeed goes away if instead of the model $\mathcal{M}$, we adopt its 'closure' $\mathcal{M}' = \{P_\theta \mid \theta \in [0.2, 0.8]\}$. However, high-dimensional regression problems exhibit the same phenomenon, even if their parameter spaces are connected. It turns out that in general, to get concentration at the same rates as if the model were correct, the model must be *convex*, i.e. closed under taking any finite mixture of the densities, which is a much stronger requirement than mere connectedness. For standard Gaussian regression problems with $Y \mid X \sim N(0, \sigma^2)$, this would mean that we would have to adopt a model in which $Y \mid X$ can be any Gaussian mixture with arbitrarily many components — which is clearly not practical (note that 'convex' refers to the space of densities, not the space of regression functions (Grünwald and Langford, 2007, Section 6.3.5)).

4.1.2 Explanation II: Good vs. bad misspecification

Barron (1998) showed that sequential Bayesian prediction under a logarithmic score function shows excellent behaviour in a cumulative risk sense; for a related result see (Barron et al., 1999, Lemma 4). Although Barron (1998) focuses on the well-specified case, this assumption is not required for the proof and the result still holds even if the model $\mathcal{M}$ is completely wrong. For a precise description and proof of this result emphasizing that it holds under misspecification, see (Grünwald, 2007, Section 15.2). At first sight, this leads to a paradox, as we now explain.

A paradox? Let $\tilde{\theta}$ index the KL-optimal distribution in Θ as in Section 3.2.1. The result of Barron (1998) essentially says that, for arbitrary models Θ, for all n,

$$\mathbf{E}_{Z^n \sim P^*}\left[\sum_{i=1}^{n} \textsc{risk}^{\log}(\Pi \mid Z^{i-1}) - \textsc{risk}^{\log}(\tilde{\theta})\right] \leq \textsc{red}_n, \tag{4.2}$$

where $\textsc{risk}^{\log}(W)$, for a distribution W on Θ, is defined as the log-risk obtained when predicting by the W-mixture of f_θ, i.e.

$$\textsc{risk}^{\log}(W) = \mathbf{E}_{X,Y \sim P^*}[-\log \mathbf{E}_{\theta \sim W} f_\theta(Y \mid X)]. \tag{4.3}$$

In (4.2), this coincides with log-risk of the Bayes predictive density $\bar{f}(\cdot \mid Z^{i-1})$, as defined by (3.8). Here, as in the remainder of this section, we look at the standard Bayes predictive density, i.e. $\eta = 1$. $\textsc{red}_n$ is the so-called *relative expected stochastic complexity* or *redundancy* (Grünwald, 2007), which depends on the prior and for 'reasonable' priors is typically *small*. The result thus means that, when sequentially predicting using the standard predictive distribution under a log-scoring rule, one does not lose much compared to when predicting with the log-risk optimal $\tilde{\theta}$.

When $\mathcal{M}$ is a union of a finite or countably infinite number of parametric exponential families and $\tilde{p} < \infty$ is well-defined, then, under some further regularity conditions, which hold in the regression example of Chapter 3 (Grünwald, 2007), the redundancy is, up to $O(1)$, equal to the BIC term $(\tilde{k}/2) \log n$, where $\tilde{k}$ is the dimensionality of the smallest model containing $\tilde{\theta}$. In the regression case, $\mathcal{M}_{\tilde{p}}$ has $\tilde{p} + 2$ parameters $(\beta_0, \ldots, \beta_p, \sigma^2)$, so in the two experiments of Section 3.5, $\tilde{k} = 6$. Thus, in our regression example, when sequentially predicting with the standard Bayes predictive $\bar{f}(\cdot \mid Z^{i-1})$, the cumulative log-risk is at most $n \cdot \textsc{risk}^{\log}(\tilde{\theta})$ which is linear in n, plus a logarithmic term that becomes comparatively negligible as n increases. This is confirmed by Figure 4.2 on page 77. Now, for each individual $\theta = (p, \beta, \sigma^2)$ we know from Section 3.2.3 that, if its log-risk is close to that of $\tilde{\theta}$, then its square-risk must also be close to that of $\tilde{\theta}$; and $\tilde{\theta}$ itself has the smallest square-risk among all $\theta \in \Theta$. Hence, one would expect the reasoning for log-risk to transfer to square-risk: it seems that when sequentially predicting with the standard Bayes predictive $\bar{f}(\cdot \mid Z^{i-1})$, the cumulative square-risk should at most be n times the instantaneous square-risk of $\tilde{\theta}$ plus a term that hardly grows with n; in other words, the cumulative

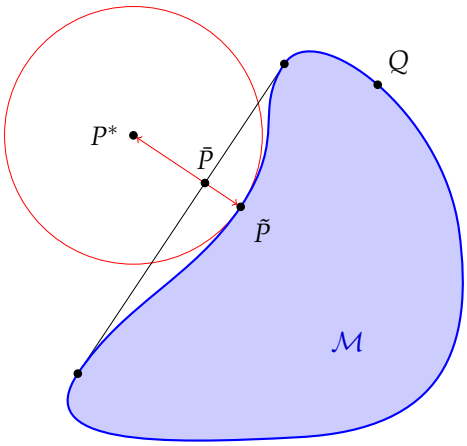

Figure 4.1: Benign vs. bad misspecification: $\tilde{P} = \arg\min_{P \in \mathcal{M}} D(P^* \| P)$ is the distribution in model $\mathcal{M}$ that minimizes KL divergence to the 'true' P^*, but, since the model is nonconvex, the Bayes predictive distribution $\bar{P}$ may happen to be very different from any $P \in \mathcal{M}$. When this happens, we can have 'bad misspecification' and then it may be necessary to decrease the learning rate (in this simplistic drawing $\bar{P}$ is a mixture of just two distributions; in our regression example it mixes infinitely many). Yet if P^* were such that $\inf_{P \in \mathcal{M}} D(P^* \| P)$ does not decrease if the infimum is taken over the convex hull of $\mathcal{M}$ (e.g. if Q rather than $\tilde{P}$ reached the minimum), then any learning rate $\eta < 1$ is fine ('benign' misspecification). In the picture, we even have $D(P^* \| \bar{P}) < D(P^* \| \tilde{P})$; in this case we can get hypercompression.

square-risk from time 1 to n, averaged over time by dividing by n, should rapidly converge to the constant instantaneous risk of $\tilde{\theta}$. Yet the experiments of Section 3.5 clearly show that this is *not* the case: Figure 3.3 shows that, until $n = 100$, it is about 3 times as large.

This 'paradox' is resolved once we realize that the Bayesian predictive density $\bar{f}(\cdot \mid^{i-1})$ is a *mixture* of various f_θ, and not necessarily similar to f_θ for any individual θ — the link between log-risk and square-risk (3.4) only holds for individual $\theta = (p, \beta, \sigma^2)$, not for mixtures of them. Indeed, if at each point in time i, $\bar{f}(\cdot \mid Z^i)$ would be very similar (in terms of e.g. Hellinger distance) to some particular f_{θ_i} with $\theta_i \in \Theta$, then there would really be a contradiction. Thus, the discrepancy between the good log-risk and bad square-risk results in fact *implies* that at a substantial fraction of sample sizes i, $\bar{f}(\cdot \mid Z^i)$ must be substantially different from *every* $\theta \in \Theta$. In other words, *the posterior is not concentrated at such i*. A cartoon picture of this situation is given in Figure 4.1: the Bayes predictive achieves small log-risk because it mixes together several distributions into a single predictive distribution which is very different from any particular single $f_\theta \in \mathcal{M}$. By Barron's bound, (4.2), the resulting $\bar{f}(\cdot \mid Z^i)$ must, averaged over i, have at most a risk almost as small as the risk of $\tilde{\theta}$. We can thus, at least informally, distinguish between "benign" and "bad" misspecifi-

cation. Bad misspecification occurs if there is a nonnegligible probability that for a range of sample sizes, the predictive distribution is substantially different from any of the distributions in $\mathcal{M}$. As Figure 4.1 suggests, 'bad' misspecification cannot occur for convex models $\mathcal{M}$ — and indeed, the results by Li (1999) suggest that for such models consistency holds under weak conditions for any $\eta < 1$, even under misspecification.

4.1.3 Hypercompression

The picture suggests that, if, as in our regression model, the model is nonconvex (i.e. the set of densities $\{f_\theta \mid \theta \in \Theta\}$ is not closed under taking mixtures), then $\bar{f}(\cdot \mid Z^i)$ might in fact be significantly *better* in terms of log-risk than the best $\tilde{\theta}$, and its individual constituents might even all be substantially worse than $\tilde{\theta}$. If this were indeed the case then, with high P^*-probability, we would also get the analogous result for an actual sample (and not just in expectation): the cumulative log-risk obtained by the Bayes predictive should be significantly smaller than the cumulative log-risk achieved with the optimal $\tilde{f}$. Figure 4.2 below shows that this indeed happens with our data, until $n \approx 100$.

The no-hypercompression inequality In fact, Figure 4.2 shows a phenomenon that is virtually impossible if the Bayesian's model and prior are 'correct' in the sense that data Z^n would behave like a typical sample from them: it easily follows from Markov's inequality (for details see Grünwald, 2007, Chapter 3) that, letting Π denote the Bayesian's joint distribution on $\Theta \times \mathcal{Z}^n$, for each $K \geq 0$,

$$\Pi\Big\{(\theta, Z^n) \,:\, \sum_{i=1}^{n} \Big(-\log \bar{f}(Y_i \mid X_i, Z^{i-1})\Big) \\ \leq \sum_{i=1}^{n} \Big(-\log f_\theta(Y_i \mid X_i, Z^{i-1})\Big) - K\Big\} \leq e^{-K},$$

i.e. the probability that the Bayes predictive $\bar{f}$ cumulatively outperforms f_θ, with θ drawn from the prior, by K log-loss units is exponentially small in K. Figure 4.2 below thus shows that at sample size $n \approx 90$, an a priori formulated event has happened of probability less than e^{-30}, clearly indicating that something about our model or prior is quite wrong.

Since the difference in cumulative log-loss between $\bar{f}$ and f_θ can be interpreted as the amount of bits saved when coding the data with a code that would be optimal under $\bar{f}$ rather than f_θ, this result has been called the *no-hypercompression inequality* by Grünwald (2007). The figure shows that for our data, we have substantial hypercompression.

The SafeBayes error measure As seen from (3.18), SafeBayes measures the performance of η-generalized Bayes not by the cumulative log-loss, as standard Bayes does, but instead by the cumulative posterior-expected error when

predicting by drawing from the posterior. One way to interpret this alternative error measure is that, at least in expectation, we cannot get hypercompression. Defining (compare to (4.3)!)

$$\text{RISK}^{\text{R-log}}(W) = \mathbf{E}_{X,Y\sim P^*}\, \mathbf{E}_{\theta\sim W}[-\log f_\theta(Y \mid X)], \tag{4.4}$$

we get by Fubini's theorem,

$$\begin{aligned}&\text{RISK}^{\text{R-log}}(W) - \text{RISK}^{\log}(\tilde{\theta}) \\ &\quad = \mathbf{E}_{\theta\sim W}\, \mathbf{E}_{X,Y\sim P^*}[[-\log f_\theta(Y \mid X)] - [-\log f_{\tilde{\theta}}(Y \mid X)]] \geq 0, \end{aligned} \tag{4.5}$$

where the inequality follows by definition of $\tilde{\theta}$ being log-risk optimal among Θ. There is thus a crucial difference between $\text{RISK}^{\text{R-log}}$ and $\text{RISK}^{\log}$ — for the latter we just argued that, under misspecification, $\text{RISK}^{\log}(W) - \text{RISK}^{\log}(\tilde{\theta}) \leq 0$ is very well possible. Thus, in contrast to predicting with the mixture density $\mathbf{E}_{\theta\sim W}\, f_\theta$, prediction by randomization (first sampling $\theta \sim W$ and then predicting with the sampled f_θ) cannot 'exploit' the fact that mixture densities might have smaller log-risk than their components. Thus, if the difference (4.5) is small, then W must put most of its mass on distributions $\theta \in \Theta$ that have small log-risk themselves. For *individual* θ, we know that small log-risk implies small square-risk. This implies that if (4.5) is small, then the (standard) posterior is concentrated on distributions with small square-risk.

Experimental demonstration of hypercompression for standard Bayes Figure 4.2 and Figure 4.3 show the predictive capabilities of standard Bayes in the wrong model example in terms of *cumulative* and *instantaneous log-loss* on a simulated sample. The graphs clearly demonstrate hypercompression: the Bayes predictive cumulatively performs *better* than the best single model / the best distribution in the model space, until at about $n \approx 100$ there is a phase transition. At individual points, we see that it sometimes performs a little worse, and sometimes (namely at the 'easy' $(0, 0)$ points) much better than the best distribution. We also see that, as anticipated above, randomized and in-model Bayesian prediction do *not* show hypercompression and in fact perform terribly on the log-loss until the phase transition at $n = 100$, when they become almost as good as standard Bayes. We see that for $\eta = 1$, they perform much worse. An important conclusion is that *if we are only interested in log-loss prediction, it is clear that we just want to use Bayes rather than randomized or in-model prediction with different* η.

> As an aside, we note that the first few outcomes have a dramatic effect on cumulative R- and I-log-loss (it disappears from Figure 4.2); this may be due to the fact that our densities — other than those considered by Grünwald (2012) — have unbounded support so that there is no v such that Theorem 4.1 below holds. This observation inspired the idea described in Section 5.2.1 about ignoring the first few outcomes when determining the optimal η. Also, we emphasize that the hypercompression phenomenon takes places more generally, not just in our regression setup — for example, the classification inconsistency noted by Grünwald and Langford (2007) can be understood in terms of hypercompression as well.

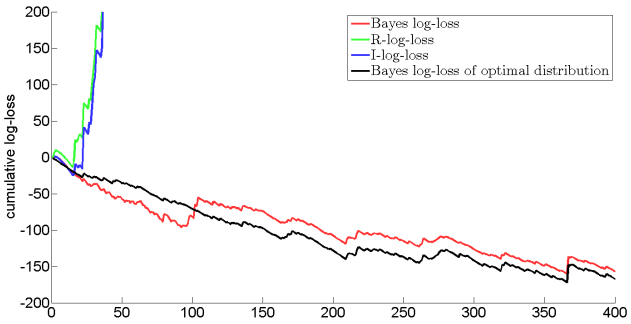

Figure 4.2: Cumulative standard, R-, and I-log-loss as defined in (3.18) and (3.22) respectively of standard Bayesian prediction ($\eta = 1$) on a single run for the model-averaging experiment of Figure 3.3. We clearly see that standard Bayes achieves *hypercompression*, being better than the best single model. And, as predicted by theory, randomized Bayes is never better than standard Bayes, whose curve has negative slope because the densities of outcomes are > 1 on average.

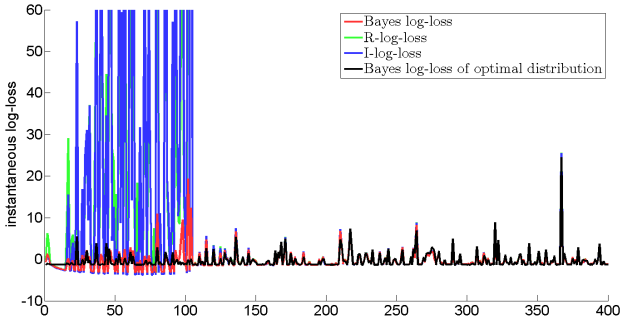

Figure 4.3: Instantaneous standard, R- and I-log-loss of standard Bayesian prediction for the run depicted in Figure 4.2.

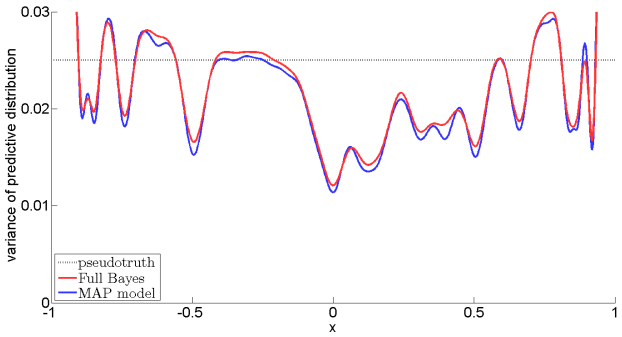

Figure 4.4: Variance of standard Bayes predictive distribution conditioned on a new input S as a function of S after 50 examples for the polynomial model-wrong experiment (Figure 3.1), shown both for the predictive distribution based on the full, model-averaging posterior and for the posterior conditioned on the MAP model $\mathcal{M}_{\check{p}_{\text{map}}}$. For both posteriors, the posterior mean of Y is incorrect for $x \neq 0$, yet $\bar{f}(Y \mid Z^{50}, X)$ still achieves small risk because of its small variance at $X = 0$.

How hypercompression arises in regression Figure 4.4 gives some clues as to how hypercompression is achieved: it shows the variance of the predictive distribution $\bar{f}(\cdot \mid Z^{50})$ as a function of $S \in [-1, 1]$ for the polynomial example of Figure 3.1 in the introduction, at sample size $n = 50$, where hypercompression takes place. Figure 3.1 gave the posterior mean (regression function) at $n = 100$; the function at $n = 50$ looks similar, correctly having mean 0 at $S = 0$ but, incorrectly, mean far from 0 at most other S. The predictive distribution conditioned on the MAP model $\mathcal{M}_{\check{p}_{\text{map}(Z^{50})}}$ is a t-distribution with approximately $\check{p}_{\text{map}(Z^{50})} \approx 50$ degrees of freedom, which means that it is approximately normal. Figure 4.4 shows that its variance is *much* smaller than the variance $\tilde{\sigma}^2$ at $S = 0$; as a result, its log-risk conditional on $U = 0$ is smaller than that of $\tilde{\theta} = (\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ by some large amount A. Conditioned at $S \neq 0$, its conditional mean is off by some amount, and its variance is, on average, slightly (but not much) smaller than $\tilde{\sigma}^2$, making its conditional log-risk given $U \neq 0$ larger than that of $\tilde{\theta}$ by an amount A' where, it turns out, A' is smaller than A. Both events $S = 0$ and $S \neq 0$ happen with probability $1/2$, so that the final, unconditional log-risk of $\bar{f}(\cdot \mid Z^{50})$ is smaller than that of $\tilde{\theta}$.

Summarizing, hypercompression occurs because the variance of the predictive distribution conditioned on past data and a new X is much smaller than $\tilde{\sigma}^2$ at $X = 0$. This suggests that, if instead of a prior on σ^2 we use models $\mathcal{M}_p$ with a fixed σ^2, we can only get hypercompression (and correspondingly bad square-risk behaviour) if $\sigma^2 \ll \tilde{\sigma}^2$, because the predictive variance based on linear models $\mathcal{M}_p$ with fixed variance σ^2 given $X = x$ is, for all x, lower bounded by σ^2. Our experiments in Section 5.1.1 confirm that this is indeed what happens.

4.1.4 Explanation III: The mixability gap & the Bayesian belief in concentration

As we indicated at the end of Section 4.1.2, bad misspecification can occur only if the standard ($\eta = 1$) posterior is *nonconcentrated*.[1] Intriguingly, by formalizing 'concentration' in the appropriate way, we will now show, under some conditions on the prior, that a *Bayesian a priori always believes that the posterior will concentrate very fast*. Thus, if we observe data Z^n, and for many $n' \leq n$, the posterior based on $Z^{n'}$ is not concentrated, then we can view this as an indication of bad misspecification. In the next section we will see that SafeBayes selects a $\hat{\eta} \ll 1$ iff we have such nonconcentration at $\eta = 1$. Thus, SafeBayes can partially be understood as a prior predictive check, i.e. a test whether the assumptions implied by the prior actually hold on the data (Box, 1980).

The mixability gap We express posterior nonconcentration in terms of the *mixability gap* (Grünwald, 2012; De Rooij et al., 2014). In this section we only consider the special case of $\eta = 1$ (standard Bayes), for which the mixability gap δ_i is defined as the difference between 1-R-log-loss (3.18) and standard log-loss as obtained by predicting with the posterior predictive, at sample size i:

$$\begin{aligned}\delta_i &:= \mathbf{E}_{\theta\sim\Pi|z^{i-1}}\left[-\log f(y_i \mid x_i, \theta)\right] - \left(-\log \mathbf{E}_{\theta\sim\Pi|z^{i-1}}[f(y_i \mid x_i, \theta)]\right) \\ &= \mathbf{E}_{\theta\sim\Pi|z^{i-1}}\left[-\log f_\theta(y_i \mid x_i)\right] - \left(-\log \bar{f}(y_i \mid x_i, z^{i-1})\right), \qquad (4.6)\end{aligned}$$

Straightforward application of Jensen's inequality as in (3.19) gives that $\delta_i \geq 0$. δ_i, which depends on $z_1, \ldots, z_i$, is a measure of the posterior's concentratedness at sample size i when used to predict y_i given x_i: it is small if $f_\theta(y_i \mid x_i)$ does not vary much among the θ that have substantial η-posterior mass; by strict convexity of $-\log$, it is 0 iff there exists a set Θ_0 with $\Pi(\Theta_0 \mid Z^{i-1}) = 1$ such that for all $\theta, \theta' \in \Theta_0$, $f_\theta(y_i \mid x_i) = f_{\theta'}(y_i \mid x_i)$.

We set the *cumulative mixability gap* to be $\Delta_n := \sum_{i=1}^n \delta_i$.

The Bayesian belief in posterior concentration As a theoretical contribution of this chapter, we now show that, under some conditions on model and prior, if the data are as expected by the model and prior, then the expected mixability gap goes to 0 as $O((\log n)/n)$, and hence a Bayesian automatically a priori believes that the posterior will concentrate fast. For simplicity we restrict ourselves to a model $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$ where Θ is countable, and we let all $\theta \in \Theta$ represent a conditional distribution for Y given X, extended to n outcomes by independence. We let π be a probability mass on Θ, and define the joint Bayesian distribution Π on $\Theta \times \mathcal{Y}^n \mid \mathcal{X}^n$ in the usual way, so that for measurable $\mathcal{A} \subset \mathcal{Y}^n$, $\Pi((\theta^*, \mathcal{A}) \mid X^n = x^n) = \pi(\theta^*) \cdot P_{\theta^*}(\mathcal{A} \mid X^n = x^n)$. The random variable θ^* refers to the θ chosen according to density π. We will

[1] Things would simplify if we could say 'bad misspecification can occur if and only if there is hypercompression', but we do not know whether that is the case; see Section 4.3.3.

look at the Bayesian probability distribution of the θ^*-expected mixability gap, $\bar{\delta}_n := \mathbf{E}_{\theta^*}[\delta_n]$.

Theorem 4.1. *Consider a countable model with prior Π as above. Suppose that the density ratios in Θ are uniformly bounded, i.e. there is a $v > 1$ such that for all $x, y \in \mathcal{X} \times \mathcal{Y}$, all $\theta, \theta' \in \Theta$, $f_\theta(y \mid x)/f_{\theta'}(y \mid x) \leq v$. Suppose that for some $\eta < 1$ we have $\sum_\theta \pi(\theta)^\eta < \infty$. Then for every $a > 0$ there are constants C_0 and C_1 such that, for all n,*

$$\Pi\left(\bar{\delta}_n \geq C_0 \cdot \frac{\log n}{n} \right) \leq C_1 \cdot \frac{1}{n^a}. \tag{4.7}$$

Moreover, for any $0 < a' \leq 1$, there exist C_2 and C_3 such that

$$\Pi\left(\Delta_n \geq C_2 \cdot n^{a'} \right) \leq C_3 \cdot \frac{(\log n)^2}{n^{a'}}, \tag{4.8}$$

i.e. the Bayesian believes that the mixability gap will be small on average and that the cumulative mixability gap will be small with high probability.

Thus, even though Δ_n is the difference between two quantities that are typically linear in n, with high probability it grows only polylogarithmically. This means that observing a large value of Δ_n strongly indicates misspecification.

> We hasten to add that the regularity conditions for Theorem 4.1 do *not* hold in the regression problem of Chapter 3; the theorem is merely meant to show that Δ_n is believed to be small in idealized circumstances that have been simplified so as to make mathematical analysis easier. Note however, that the regularity conditions do not constrain Θ in the most important respect: by allowing countably infinite Θ, we can approximate nonparametric models arbitrarily well by suitable covers (Cover and Thomas, 1991). In particular we do allow sets Θ for which maximum likelihood methods would lead to disastrous overfitting at all sample sizes. Also the condition that $\sum \pi(\theta)^\eta < \infty$ is standard in proving Bayesian and MDL convergence theorems (Barron and Cover, 1991; Zhang, 2006a). In fact, since the constants C_0 and C_1 scale logarithmically in v, we expect that Theorem 4.1 can be extended to the regression setting we are dealing with here as long as all distributions in the model have exponentially small tails, using methods similar to those in Grünwald (2014).

Example 4.B. [Cumulative nonconcentration can (and will) go together with momentary concentration: Example 4.A, Bernoulli, cont.] Consider the first instance of the Bernoulli Example 4.A again, where we again look at what happens if both distributions are equally bad: $\mathcal{M} = \{P_{0.2}, P_{0.8}\}$, whereas $Y_1, Y_2, \ldots$ are i.i.d. $\sim P_{\theta^*}$ with $\theta^* = 1/2$. As we showed in that example, at any given n, with P_{θ^*}-probability at least 0.32, $\min_{\theta \in \{0.2, 0.8\}} \pi(\theta \mid Y^n) \approx 2^{-\sqrt{n}/2}$: the posterior puts almost all mass on one θ. Lemma 6 of Van Erven et al. (2011) shows that in such cases δ_n is small; in this case, $\delta_n \leq 2(e-2) \min_{\theta \in \{0.2, 0.8\}} \pi(\theta \mid Y^n)$ $\approx 1.42 \cdot 2^{-\sqrt{n}/2}$. Thus, the posterior *looks* exceedingly concentrated at time n, with nonnegligible probability (this unwarranted confidence is a simplified

version of what was called the *fair balance paradox* by Yang (2007b), who conjectured it is the underlying reason for the problem of 'overconfident posteriors' in Bayesian phylogenetic tree inference). However, SafeBayes detects misspecification by looking at *cumulative* concentration, i.e. the sum of the δ's: L as in (4.1) can be interpreted as a random walk on $\mathbf{Z}$ starting at the origin, with equal probabilities to move to the left and to the right. By the central limit theorem, the random walk crosses the origin at time n with probability about $1/\sqrt{n\pi/2} = \tilde{O}(n^{-1/2})$, so that we may conjecture that, with high probability, it crosses the origin $\tilde{O}(n \cdot n^{-1/2}) = \tilde{O}(n^{1/2})$ times. Each time it crosses the origin, the posterior is uniform and hence as nonconcentrated as it can be, and Δ_n is increased by at least a fixed constant. One would therefore expect (under the 'true' θ^*) that Δ_n is of order $\sqrt{n}$, which by Theorem 4.1 is much larger than a Bayesian a priori expect it to be — the model fails the 'prior predictive check'.[2]

4.2 How SafeBayes works

In its simplest form, the in-model fixed variance case, SafeBayes finds the $\hat{\eta}$ that minimizes cumulative square-loss on the sample and thus can simply be understood as a pragmatic attempt to find a $\hat{\eta}$ that achieves small risk. However, the other versions of SafeBayes do not have such an easy interpretation. To explain them further, we need to generalize the notion of mixability gap in terms of the η'-*flattened* η-generalized Bayesian predictive density. The latter is defined, for $\eta, \eta' \leq 1$, as:

$$\bar{f}(y_i \mid x_i, z^{i-1}, \langle \eta' \rangle; \eta) := \left(\mathbf{E}_{\theta \sim \Pi \mid z^{i-1}, \eta} \left[f_\theta^{\eta'}(y_i \mid x_i) \right] \right)^{1/\eta'}. \tag{4.9}$$

By Jensen's inequality, we have $\bar{f}(y_i \mid x_i, z^{i-1}, \langle \eta' \rangle; \eta) \leq \bar{f}(y_i \mid x_i, z^{i-1}, \eta)$ for any $\eta' \leq 1$ and any (x_i, y_i). Indeed, intentionally, $\bar{f}(\cdot \mid \langle \eta' \rangle; \eta)$ is a 'defective' density in the sense that $\int_{\mathbf{R}} \bar{f}(y \mid x_i, z^{i-1}, \langle \eta' \rangle; \eta) dy < 1$. The log-loss achieved by η-generalized, η'-flattened Bayesian prediction is called (η, η')-*mix-loss* from now on, following terminology from De Rooij et al. (2014). For $0 < \eta \leq \eta' \leq 1$, the *mixability gap* $\delta_{i,\eta,\eta'}$ is defined as the difference between the η-R-log-loss and the η'-mix-loss:

$$\delta_{i,\eta,\eta'} := \mathbf{E}_{\theta \sim \Pi \mid Z^{i-1}, \eta} \left[-\log f_\theta(Y_i \mid X_i) \right] - \left(-\log \bar{f}(Y_i \mid X_i, Z^{i-1}; \langle \eta' \rangle; \eta) \right). \tag{4.10}$$

We once again define a cumulative version $\Delta_{n,\eta,\eta'} = \sum_{i=1}^n \delta_{i,\eta,\eta'}$, and note that the definitions are compatible with the special cases $\delta_i := \delta_{i,1,1}$ and $\Delta_n := \Delta_{n,1,1}$ defined in the previous subsection. Now we can rewrite the cumulative R-log-

[2] This heuristic argument can actually be formalized: if data are i.i.d. Bernoulli$(1/2)$, then the expected regret for every absolute loss predictor is of order $\tilde{O}(n^{1/2})$ (Cesa-Bianchi and Lugosi, 2006), which implies, via the connections between regret and Δ_n given by De Rooij et al. (2014), that Δ_n must also be of order $n^{1/2}$; we omit further details.

loss achieved by Bayes with the η-generalized posterior as

$$\sum_{i=1}^{n} \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[-\log f_\theta(y_i \mid x_i)\right] = \Delta_{n,\eta,\eta'} + \mathrm{CML}_{n,\eta,\eta'}, \tag{4.11}$$

where

$$\mathrm{CML}_{n,\eta,\eta'} = \left(\sum_{i=1}^{n} -\log \bar{f}(y_i \mid x_i, z^{i-1}, \langle\eta'\rangle; \eta)\right)$$

is the cumulative (η,η')-mix-loss. (4.11) holds for all $0 < \eta \leq \eta' \leq 1$. Consider first $\eta' = 1$. As was seen, if $\Delta_{n,1,1}$ is large, then this indicates potential bad misspecification. But (4.11) still holds for smaller $\eta' < 1$; by Jensen's inequality, for any fixed η, decreasing η' will make $\Delta_{n,\eta,\eta'}$ smaller as well. Indeed, for any fixed P^*, defining

$$\bar{\delta}_{\eta'} := \sup_W \mathbf{E}_{X,Y\sim P^*}\left[\mathbf{E}_{\theta\sim W}[-\log f_\theta(Y \mid X)] - \left(-\frac{1}{\eta'}\log \mathbf{E}_{\theta\sim W}[f_\theta(Y \mid X)^{\eta'}]\right)\right],$$

where the supremum is over *all* distributions on Θ, we have

$$\lim_{\eta'\downarrow 0} \bar{\delta}_{\eta'} = 0,$$

so we have an upper bound on the expectation of $\Delta_{n,\eta,\eta'}$ independent of the actual data that, for small enough η', will become negligibly small. But the left-hand side of (4.11) does not depend on η', so if, by decreasing η', we decrease $\Delta_{n,\eta,\eta'}$, $\mathrm{CML}_{n,\eta,\eta'}$ must increase by the same amount — so as yet we have gained nothing. Indeed, not surprisingly, Barron's bound does not hold any more for $\mathrm{CML}_{n,\eta,\eta'}$ with $\eta = 1$ and $\eta' < 1$ (and in general, it does not hold for η,η' whenever $\eta' < \eta$). *But*, it turns out, a version of Barron's bound still holds for $\mathrm{CML}_{n,\eta',\eta'}$, for all $\eta' > 0$: the cumulative log-risk of η'-flattened, η'-generalized Bayes is still guaranteed to be within a small RED_n of the cumulative log-risk of $\tilde{\theta}$, although RED_n does monotonically increase as η' gets smaller — simply because the prior becomes more important relative to the data (standard results in learning theory show that $\mathrm{CML}_{n,\eta,\eta}$ is monotonically decreasing in η, and can be upper bounded as $O(1/\eta)$; see e.g. (De Rooij et al., 2014, Lemma 1). Thus, it makes sense to consider the special case $\eta' = \eta$, and think of SafeBayes as finding the η minimizing

$$\sum_{i=1}^{n} \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[-\log f_\theta(y_i \mid x_i)\right] = \Delta_{n,\eta,\eta} + \mathrm{CML}_{n,\eta,\eta}, \tag{4.12}$$

since we have clear interpretations of both terms: the second indicates, by Barron's bound, how much worse the η-generalized posterior predicts in terms of log-loss compared to the optimal $\tilde{\theta}$; the first indicates how much is additionally lost if one is forced to predict by distributions inside the model. The second term decreases in η, the first has an upper bound which increases in

η. SafeBayes can now be understood as trying to minimize both terms at the same time.

Now broadly speaking, the central convergence result of Grünwald (2012) states that $\Delta_{n,\eta,\eta}$ will be 'sufficiently small' for all $\eta < 1$, and under some further conditions even for $\eta = 1$, if the model is correct or convex; and it will also be 'sufficiently small' if the model is incorrect, as long as η is smaller than some 'critical' value η_{crit} (which may depend on n though). Here 'sufficiently small' means that it is not the dominating term in (4.12). Intuitively, we would like the $\hat{\eta}$ determined by SafeBayes to be the largest η that is smaller than η_{crit}. Grünwald (2012) shows that SafeBayes indeed finds such an η, and that prediction based on the generalized posterior with this η achieves good frequentist convergence rates.

Experimental illustration Consider the main wrong-model experiment of Section 3.5. Figure 4.5 shows, as a function of η, in red, the cumulative η-R-log-loss measured by SafeBayes, averaged over 30 runs of the wrong-model experiment of Figure 3.3. In each individual run, SafeBayes picks the $\hat{\eta}$ minimizing this quantity; we thus get that on most runs, $\hat{\eta}$ is close to 0.4. In contrast to η-R-log-loss, and as predicted by theory, the η-mix-loss (in purple) decreases monotonically and coincides with the standard Bayesian log-loss at $\eta = 1$ and with the η-R-log-loss as $\eta \downarrow 0$. We also see hypercompression again: near $\eta = 1$, both the Bayesian log-loss and the mix-loss are smaller than the log-loss achieved by the best $\tilde{\theta}$ in the model. At $\eta = 0.5$, there is a sudden sharp rise in $\Delta_{n,\eta,\eta}$ (the difference between the red and purple curves). We can think of SafeBayes as trying to identify this 'critical' η_{crit}.

> Theorem 4.1 shows that, if both model and prior are well-specified, then the Bayesian posterior cumulatively concentrates in a very strong sense. More generally, if the model is correct but also if there is 'benign' misspe-

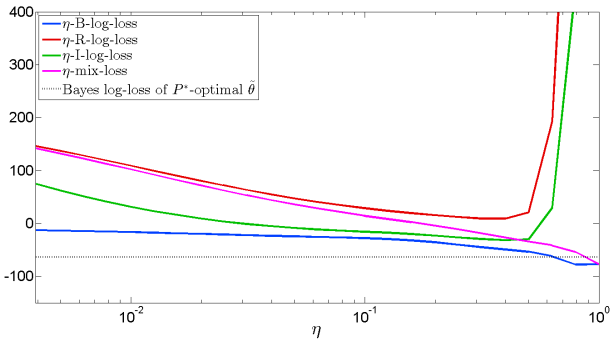

Figure 4.5: Cumulative losses up to sample 100 (where the posterior has not converged yet) as a function of η, averaged over 30 runs, for the experiment of Figure 3.3. η-B-log-loss is the cumulative log-loss achieved by standard Bayes with the η-generalized posterior.

cification, then, under some conditions on the prior, by the results of Grünwald (2012), the Bayesian posterior eventually cumulatively concentrates at $\eta = 1$. One might thus be tempted to interpret η_{crit} (the learning rate which SafeBayes tries to learn) as 'largest learning rate at which the posterior cumulatively concentrates'. However, this interpretation works only if $\eta_{\text{crit}} = 1$. If $\eta_{\text{crit}} < 1$, we can only show that, for every $\eta < \eta_{\text{crit}}$, $\Delta_{n,\eta,\eta}$ is small; true cumulative concentration would instead mean that $\Delta_{n,\eta,1}$ is small for such η (note we must have $\Delta_{n,\eta,\eta} \leq \Delta_{n,\eta,1}$ by Jensen). The figure shows that $\Delta_{n,\eta,1}$ (the difference between the red and blue curve) may indeed be large even at small η. A better interpretation is that, for every fixed η, with decreasing η', the geometry of the (η, η')-mix-loss changes, so that the loss difference between the mix loss and the R-log-loss obtained by randomization gets smaller. By then further using the generalized posterior for the same η', we guarantee that a version of Barron's bound holds for the (η', η')-mix-loss.

Replacing R- by I-loss Although the proofs of Grünwald (2012) are optimized for R-SafeBayes, the same story as above can be told for any fixed transformation from the posterior to a possibly randomized prediction, i.e. anything of the form (3.21); in particular for the most extreme transformation where we replace the posterior predictive by the distribution indexed by the posterior mean parameters so that instead of R-SafeBayes we end up with I-SafeBayes. In fact, the importance of the distinction between 'in-model' and 'out-model' prediction under model misspecification has been emphasized before (Grünwald, 2007; Barron and Hengartner, 1998; Kotłowski et al., 2010). In general, although we do not know how to exploit this intuition to strengthen the convergence proofs of Grünwald (2012), it seems more natural to replace the randomized predictions by deterministic, in-model predictions.

4.3 Discussion, open problems and conclusion

"If a subjective distribution Π attaches probability zero to a non-ignorable event, and if this event happens, then Π must be treated with suspicion, and *modified* or replaced" (emphasis added)
— A.P. Dawid (1982).

"Some models are obviously wrong, yet evidently useful"
— (very freely paraphrasing Box, 1979).

We already discussed the theoretical significance of the inconsistency result in the introduction. Extensive further discussion on Bayesian inference under misspecification is given by Walker (2013) and Grünwald and Langford (2007). For us, it remains to discuss the place of both the inconsistency result and our solution in Bayesian methodology.

Following the well-known Bayesian statisticians Box (1980), Good (1983), Dawid (1982, 2004) and Gelman (2004) (see also Gelman and Shalizi, 2012), we take the stance that model checking is a crucial part of successful Bayesian

practice. When there is a large discrepancy between a model's predictions and actual observations, it is not merely sufficient to keep gathering data and update one's posterior: something more radical is needed. In many such cases, the right thing to do is to go back to the drawing board and try to devise a more realistic model. However, we think this story is incomplete: in machine learning and pattern recognition, one often encounters situations in which the model employed is *obviously* wrong in some respects, yet there is a model instantiation (parameter vector) that is *pretty adequate* for the specific prediction task one is interested in. Examples of such obviously-wrong-yet-pretty-adequate models are, like in Chapter 3, assuming homoskedasticity in linear regression when the goal is to approximate the true regression function and the true noise is heteroskedastic,[3] but also the use of N-grams in language modelling (is the probability of a word given the previous three words really independent of everything that was said earlier?), logistic regression in e.g. spam filtering, and every single successful data compression method that we know of (see *Bayes and Gzip* (Grünwald, 2007, Chapter 17, page 537)). The difference with the more standard statistical (be it Bayesian or frequentist) mode of reasoning is eloquently described in Breiman's (2001) *the two cultures*.[4] Bayesian inference is among the most successful methods currently used in the obviously-wrong-yet-pretty-adequate-situation (to witness, state-of-the-art data compression methods such as Context-Tree-Weighting (Willems et al., 1995) have a Bayesian interpretation). Yet our results show that there is a danger: even *if* the employed model is pretty adequate (in the sense of containing a pretty good predictor), the Bayesian machinery might not be able to find it. The SafeBayesian algorithm can thus be viewed as an attempt to provide an alternative for the *data-analysis cycle* (Gelman and Shalizi, 2012) to this, in some sense, less ambitious setting: just like in the standard cycle, we do a model check, albeit a very specific one: we check whether there is 'cumulative concentration of the posterior' (see Section 4.1.4). If there is not, we know that we may not be learning to predict as well as the best predictor in our model, so we *modify* our posterior. Not in the strong sense of 'going back to the drawing board', but in the much weaker sense of making the learning rate smaller — we cannot hope that our model of reality has improved, because we still employ the same model — but we can now guarantee that we are doing the best we can with our given model, something which may be enough for the task at hand and which, as our experiments show, cannot always be achieved with standard Bayes.

[3] As long as, as in Chapter 3, the tails of the conditional distribution of Y given $X = x$ are sub-Gaussian, for each x; if they are not, there may be real outliers and then one cannot say that the model is 'pretty adequate' any more.

[4] The 'two cultures' does *not* refer to the Bayesian-frequentist divide, but to the modelling vs. prediction-divide. We certainly do not take the extreme view that statisticians should *only* be interested in prediction tasks such as classification and square-error prediction rather than density estimation and testing; our point is merely that in some cases, the goal of inference is clearly defined (it could be classification, but it could also be determination whether some random variables are (conditionally) (in)dependent), whereas part of our model is unavoidably misspecified; and in such cases, one may want to use a generalized form of Bayesian inference.

Benign vs. bad misspecification One might argue that the example of Chapter 3 is rather extreme, and that in practical situations, choosing a learning rate different from 1 may never be a useful thing to do. A crucial point here is that one can have 'benign' and 'bad' misspecification (Section 4.1.2). Under benign misspecification, standard Bayes with $\eta = 1$ will behave nicely under weak assumptions on the prior. While in our particular example, after 'eyeballing' the data one would probably have chosen a different, less misspecified model, it may be the case that 'bad' misspecification (as in Figure 4.1) also occurs, at least to some extent, in general, real-world data and is then not so easily spotted. Since we simply do not know whether such situations occur in practice, to be on the safe side, it seems desirable to have a theory about when we can get away with using standard Bayesian inference for a given prediction task even if the model is wrong, and how we can still use it with little modification if there is bad misspecification. Our work (esp. (Grünwald, 2014), the theoretical counterpart to Chapters 3–5) is a first step in this direction.

Towards a theory of Bayesian inference under misspecification What we have in mind is a theory of Bayesian inference under misspecification, in which the *goal* of learning plays a crucial role. The standard Bayesian approach is very ambitious: it can be used to solve every conceivable type of prediction or inference task. Every such task can be encoded as a loss or utility function, and, given the data and the prior, one merely has to calculate the posterior, and then makes an optimal decision by taking the act that minimizes expected loss or maximizes expected utility according to the posterior. Crucially, one uses the same posterior, independently of the utility function at hand, implying that one believes that one's own beliefs are correct *in every possible respect*. We envision a more modest approach, in which one acknowledges that one's beliefs are only adequate in some respects, not in others; how one proceeds then depends on how one's model and loss function interact. For example, if one is interested in data compression then, this problem being essentially equivalent to cumulative log-loss prediction, by Barron's (1998) bound one can simply use the standard $(\eta = 1)$ Bayesian predictive distribution — even under misspecification, this will guarantee that one predicts (at least!) as well as one could with the best element of one's model. If, on the other hand, one is interested in any of the KL-associated inference tasks (for linear models, these are square-loss and reliability, Section 3.2.3), then using $\eta = 1$ is not sufficient any more, and one may have to learn η from the data, e.g. in the SafeBayesian manner. Finally, if we are interested in an inference task that is not KL-associated under our model (i.e., a model instance can be good in the KL sense but bad in the task of interest), then a more radical step is needed: either go back to the drawing board and design a new model after all; or perhaps, the model can be changed in a more pragmatic way so that, for the right η, η-generalized Bayes once again will find the best predictor for the task at hand. Let us outline such a procedure for the case that the inference talk is simply prediction under some loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbf{R}$. In this case, if the ℓ-risk is not KL-associated this simply means that, for some data, the log likelihood is not a monotonic

function of the loss ℓ. To get the desired association, we may associate each conditional distribution $P_\theta(Y \mid X)$ in the model with its associated Bayes act δ_θ: $\delta_\theta(x)$ is defined as the act $\hat{y} \in \hat{\mathcal{Y}}$ which minimizes $P_\theta \mid X = x$-expected loss $\mathbf{E}_{Y \sim P_\theta | X=x}[\ell(y, \hat{y})]$. We can then define a new set of densities

$$f^{\text{NEW}}_{\theta,\gamma}(y \mid x) = \frac{1}{Z(\gamma)} e^{-\gamma \ell(y, \delta_\theta(x))}, \tag{4.13}$$

and perform (generalized) Bayesian inference based on these. Note that this effectively replaces, for each θ, the full likelihood by a 'likelihood' in which some information has been lost, and is thus reminiscent of what is done in *pseudo-likelihood* (Besag, 1975), *substitution likelihood* (Jeffreys, 1961; Dunson and Taylor, 2005), or *rank-based likelihood* (Gu and Ghosal, 2009) approaches (as a Bayesian, one may not want to lose information, but whether this still applies in nonparametric problems (Robins and Wasserman, 2000) let alone under misspecification (Grünwald and Halpern, 2004) is up to debate).

(4.13) can be made precise in two ways: either one just sets γ and $Z(\gamma)$ to 1, and allows the f^{NEW}_θ to be pseudo-densities, not necessarily integrating to 1 for each x. This is a standard approach in learning theory (Zhang, 2006b; Catoni, 2007). One could then learn η by, e.g., the basic SafeBayes algorithm with $\ell_\theta(x, y) := \ell(y, \delta_\theta(x))$ instead of log-loss. Or, one could define $Z(\gamma)$ so that the densities normalize (how to achieve this if $\int_y e^{-\gamma \ell(y, \delta_\theta(x))} dy$ depends on x is explained by Grünwald (2008)) and put a prior on γ as well (for linear models, this is akin to putting a prior on the variance). This will make the loss ℓ KL-associated and the KL-optimal $\tilde{\theta}$ will also have the reliability property, see again (Grünwald, 2008) for details. In this case we will get, with $z_i = (x_i, y_i)$, $\ell_\theta(z_i) := \ell(y_i, \delta_\theta(x_i))$, and using a prior on Θ and the scaling parameter γ, that the η-generalized posterior becomes

$$\pi(\theta, \gamma \mid z^n, \eta) \propto \frac{1}{Z(\gamma)^{\eta n}} e^{-\eta \gamma \sum_{i=1}^n \ell_\theta(z_i)} \cdot \pi(\theta, \gamma). \tag{4.14}$$

This idea was, in essence, already suggested by (Grünwald, 1998, Example 5.4) (see also Grünwald (1999)) under the name of *entropification* (however, Grünwald's papers wrongly suggest that, by introducing the scale parameter γ, it would be sufficient to only consider $\eta = 1$); see also (Lacoste-Julien et al., 2011; Quadrianto and Ghahramani, 2014).

Now both 'pure' subjective Bayesians and 'pure' frequentists might dismiss this programme as severe ad-hockery: the strict Bayesian would claim that nothing is needed on top of the Bayesian machinery; the strict frequentist would argue that Bayesian inference was never designed to 'work' under misspecification, so in misspecified situations it might be better to avoid Bayesian methods altogether rather than trying to 'repair' them. We strongly disagree with both types of purism, the reason being the ever-increasing number of successful applications of Bayesian methods in machine learning in situations in which models are obviously wrong. We would like to challenge the pure subjective Bayesian to explain this success, given that the statistician is using

a priori distributions that reflect beliefs which she knows to be false, and are thus not really her beliefs. We would like to challenge the pure frequentist to come up with better, non-Bayesian methods instead. In summary, we would urge both purists not to throw away the Bayesian baby with the misspecified bath water!

Moreover, from a prequential (Dawid, 1984), learning theory (citations see below) and Minimum Description Length (MDL (Barron et al., 1998)) perspective, the extension from Bayes to SafeBayes is *perfectly natural*. From the prequential perspective, SafeBayes seeks to find the largest η at which the generalized Bayesian predictions have a predictive interpretation in terms of the loss of interest rather than the log-loss. The learning theory and MDL perspectives are further explained in the next section.

4.3.1 Related work I: Learning theory and MDL

Learning theory From the learning theory perspective, generalized Bayesian updating as in (4.14) with $Z(\gamma)$ set to 1 can be seen as the result of a simple regularized loss minimization procedure (this was probably first noted by Williams (1980); see in particular (Zhang, 2006b)), which means that it continues to make sense if $\exp(-\gamma\ell_\theta)$ as in (4.13) does not have a direct probabilistic interpretation. Variations of such generalized Bayesian updating are known as "aggregating algorithm", "Hedge" or "exponential weights", and often have good worst-case optimality properties in nonstochastic settings (Vovk, 1990; Cesa-Bianchi and Lugosi, 2006) — but to get these the learning rate must often be set as small as $O(1/\sqrt{n})$. Similarly, PAC-Bayesian inference (Audibert, 2004; Zhang, 2006b; Catoni, 2007) (for a variation, see (Freund et al., 2004)) is also based on a posterior of form (4.13) and can achieve minimax optimal rates in e.g. classification problems by choosing an appropriate η, usually also very small. From this perspective, SafeBayes can be understood as trying to find a *larger* η than the worst-case optimal one, if the data indicate that the situation is not worst-case and faster learning is possible. Finally, Bissiri et al. (2013) give a motivation for (4.14) (with $Z(\gamma) \equiv 1$) based on coherence arguments that are more Bayesian in flavour.

MDL Of particular interest is the interpretation of the SafeBayesian method in terms of the MDL principle for model selection, which views learning as data compression. When several models for the same data are available, MDL picks the model that extracts the most 'regularity' from the data, as measured by the minimum number of bits needed to code the data *with the help of the model*. This is an interpretation that remains valid even if a model is completely misspecified (Grünwald, 2007). The resulting procedure (based on so-called *normalized maximum likelihood* codelengths) is operationally almost identical to Bayes factor model selection. Thus, it provides a potential answer to the question 'what does a high posterior belief in a model really mean, since one knows all models under consideration to be incorrect in any case?' (asked by, e.g., Gelman and Shalizi (2012)): even if all models are wrong, the information-

theoretic MDL interpretation stands. However, our work implies that there is a serious issue with these NML codes: note that any distribution P in a model $\mathcal{M}$ can be mapped to a code (the *Shannon-Fano code*) that would be optimal in expectation if data were sampled from P. Now, our work shows that if the data are sampled from some $P^* \notin \mathcal{M}$, then the codes based on Bayesian predictive distributions can sometimes compress substantially *better* in expectation than can be done based on any $P \in \mathcal{M}$ — this is the hypercompression phenomenon of Section 4.1.3. The same thing then holds for the NML codes, which assign almost the same codelengths as the Bayesian ones. Our work thus invalidates the interpretation of NML codelengths as 'compression with the help of (and only of!) the model', and suggests that, similarly to in-model SafeBayes one should design and use 'in-model' versions of the NML codes instead — codes that are guaranteed not to outperform, at least in expectation, the code based on the best distribution in the model.

4.3.2 Related work II: Analysis of Bayesian behaviour under misspecification

Consistency theorems The study of consistency and rate of convergence under misspecification for likelihood-based and specifically Bayesian methods go back at least to Berk (1966). For recent state-of-the-art work on likelihood-based, non-Bayesian methods see e.g. Dümbgen et al. (2011) and the very general Spokoiny (2012). Recent work on Bayesian methods includes Kleijn and Van der Vaart (2006), De Blasi and Walker (2013) and Ramamoorthi et al. (2013) who obtained results in quite general, i.i.d. nonparametric settings, non-i.i.d. settings (Shalizi, 2009), and more specific settings (Sriram et al., 2013); see also Grünwald (2014). Yet, as explicitly remarked by De Blasi and Walker (2013), the conditions on model and prior needed for consistency under misspecification are generally stronger than those needed when the model is correct. Essentially, if the data are i.i.d. both according to the model and the sampling distribution P^*, then Theorem 1 (in particular its Corollary 1) of De Blasi and Walker (2013) implies the following: if, for all $\epsilon > 0$, the model can be covered by a finite number of ϵ-Hellinger balls, then the Bayesian posterior eventually concentrates: for all $\delta, \gamma > 0$, the posterior mass on distributions within Hellinger distance δ of the $P_{\tilde{\theta}}$ that is closest to P^* in KL divergence will become larger than $1 - \gamma$ for all n larger than some n_γ. This implies that both in the ridge regression (finite p) and in the model averaging experiments (finite $p_{\max}$), Bayes eventually 'recovers' — as we indeed see in our experimental results. However, if $p_{\max} = \infty$, then the model has no finite Hellinger cover any more for small enough ϵ and indeed the conditions for Theorem 1 of De Blasi and Walker (2013) do not apply any more. Our results show that in such a case we can indeed have inconsistency if the model is incorrect. On the other hand, even if $p_{\max} = \infty$, we do have consistency in the setup of our correct-model experiment for the standard Bayesian posterior, as follows from the results by Zhang (2006a).

The limiting $\eta = 1$ Like several earlier results (Barron and Cover, 1991; Walker and Hjort, 2002), Zhang's consistency results for correct models hold under very weak conditions for generalized Bayes with any $\eta < 1$, and only under much stronger conditions for $\eta = 1$. Zhang provides an example of inconsistency-like behaviour in the well-specified case with $\eta = 1$ that automatically disappears as soon as one picks $\eta < 1$, leading Zhang (2006a) to claim that in general, generalized Bayesian methods $(\eta < 1)$ are more stable than standard Bayesian ones. Zhang's example, and the example of Bayesian model selection inconsistency in a well-specified model by Csiszár and Shields (2000) are closely related to ours, in that the Bayes predictive distribution for $\eta = 1$ becomes significantly different from any distribution in the model (see Figure 4.1). In their examples, the problem is resolved by taking any $\eta < 1$; in our misspecification case, η should even be taken much smaller.

Anomalous behaviour and modifications of Bayesian posterior under misspecification Anomalous behaviour of Bayesian inference under misspecification was, of course, observed before, e.g. (less dramatically than here) by Yang (2007b); Müller (2013) and (as dramatically, but involving a very artificial model) Grünwald and Langford (2007). Presumably also related is the 'brittleness' of Bayesian inference that has been observed by Owhadi and Scovel (2013). Not surprisingly then, we are not the first to suggest modification of likelihood-based estimators (see e.g. White, 1982; Royall and Tsou, 2003; Kotłowski et al., 2010) and posteriors (Royall and Tsou, 2003; Hoff and Wakefield, 2012; Doucet and Shephard, 2012; Müller, 2013). The latter three approaches (that extend the first) employ the so-called *sandwich posterior*, in which the covariance matrix of the posterior is changed based on a 'sandwich formula' involving the empirical variance; Müller (2013) provides extensive explanation and experimentation. Compared to the sandwich approach, our proposal, besides being applicable in fully nonparametric contexts, seems substantially more radical. This can be seen from the regression applications in Müller (2013), which involve a noninformative Jeffreys' prior on the regression coefficient vector β. With such a prior (as well as any normal prior scaled by variance σ^2), the posterior *mean* of β, and thus also the frequentist square-risk (which only depends on the posterior mean) remains unaffected by the sandwich modification, so for square-risk the method would perform like standard Bayes in our model-wrong experiments. Thus Müller (2013, Section 2.4) demonstrates its usefulness on other loss functions. Nevertheless, both the sandwich and the SafeBayesian methods can be thought of as methods for measuring the spread of a posterior, and it would be useful to compare the two in detail, both in theory and practice.

4.3.3 Future work and open problems

The results of these chapters raise several issues and prompt the following research agenda:

1. The misspecification in our example would presumably be easily spotted in practice. This raises the question whether 'bad' misspecification also arises for data sets that occur in practice and for which it would not be easily spotted. Currently, we know only of one experiment in this direction: Jansen (2013) applied the Bayesian Lasso (Park and Casella, 2008) to several real-world data sets, where the λ (i.e. $1/\eta$) is taken that minimizes the cumulative *square-loss* whereas at the same time σ^2 is a free parameter. Thus it is a hybrid of I-square-SafeBayes and I-log-SafeBayes, but equal to neither; the method was (somewhat) outperformed by standard Bayes on most data sets tried. However, we also tried this hybrid method in the model-wrong experiment of Chapter 3 and found that it is not competitive with either of the two 'true' in-model SafeBayes methods either; so the experiment does not 'really' test SafeBayes; more precise experiments are needed.

2. Our method has one major disadvantage: even if the data do not have a natural ordering, the $\hat{\eta}$ selected by SafeBayes will, in general, be order-dependent. Grünwald (2011) suggested a very different (and in fact, the first) method to learn $\hat{\eta}$, that does not have this problem. However, it is only applicable to countable models, and has no obvious computationally efficient implementation, so we do not know whether it has a future. Another method that is clearly related to I-square-SafeBayes is to determine η using leave-one-out cross-validation based on the squared error. This method is also order-independent and behaves comparably to I-square-SafeBayes (Section 5.1.1), but it is not clear how to extend it to general misspecified models. While we show in the same section that cross-validation based on log-loss of the Bayes predictive distribution fails dramatically, it may be that cross-validation based on log-loss of the Bayes posterior *mean* would generally work fine, and this method can be applied to general misspecified models, not just linear ones. Compared to I-log-SafeBayes this *in-model log-loss cross-validation* would have the advantage that it is order independent, and the disadvantage that it cannot (at least not straightforwardly) be used in an online setting and/or for non-i.i.d. models. Also, we suspect that if the number of models is exponential in the covariates (as in variable selection), cross-validation may be prone to overfitting whereas SafeBayes would not be, but this is just extrapolation from the well-specified case: it would be useful to investigate "in-model cross-validation" further.

3. What exactly are relations between the sandwich posterior (see above) and our approach? It would be good to test SafeBayes on the data sets used by Müller (2013).

4. It would be useful to establish exactly what properties of Bayesian updating remain valid for generalized Bayesian updating, and what properties do not hold any more. For example, *telescoping* (Cesa-Bianchi and Lugosi, 2006) holds for the standard posterior, for the η-flattened, η-generalized

posterior, but not for the (nonflattened) η-generalized posterior.

5. As discussed at the end of Section 4.2, the final term in (3.23) is lacking in the in-model versions of SafeBayes, and this does suggest that they should work better than the randomization versions — the corresponding $\Delta_{\eta,\eta}$ is always smaller. Yet we have no theoretical results to this end, and our empirical results confirm this to some extent (R-square-SafeBayes is not competitive), but not fully (R-log-SafeBayes is competitive), so more research is needed here.

6. As we indicated in Section 4.1.3, hypercompression implies nonconcentration, but we do not know whether the reverse implication holds as well, so we may perhaps have bad misspecification yet no hypercompression. It would give significant insight if we knew whether this indeed could happen.

7. In light of the discussion underneath (4.13), one would like to formulate a general theory of substitution likelihoods so that likelihoods can be determined based on the inference task of interest, so that this task becomes KL-associated, for *arbitrary* prediction tasks. Ideally, (4.13) and approaches such as pseudo-likelihood and rank-based likelihood would all become a special case. If this can be done, we would have a truly generalized Bayesian method.

Appendix 4.A More on mix loss

4.A.1 Implementing SafeBayes

To implement the SafeBayesian algorithm (page 52), generalized posteriors must be computed for different values of η, and the randomized loss (3.18) must be computed for each sample size. For linear models with conjugate priors as considered in our experiments, all required quantities can be computed analytically. We have already seen how to do this for models $\mathcal{M}_p$ with fixed dimension p. For unions of such models, it turns out that the mix-loss is a helpful tool.

Role of mix loss in generalized posterior over models The generalized posterior *across* a discrete set of models is given by (3.7), which, writing $\tau = (\beta, \sigma^2)$, is, via (3.10) and (3.9), equivalent to

$$\begin{aligned}\pi(p \mid z^n, \eta) &= \int_{\Theta_p} \pi(p, \tau \mid z^n, \eta)\, d\tau \\ &\propto \int (f(y^n \mid x^n, \tau, p))^\eta \pi(\tau \mid p)\, d\tau\, \pi(p). \end{aligned} \tag{4.15}$$

Here $\propto$ means 'proportional to' when p is varied and z^n and η are fixed. In practice we prefer to calculate this quantity incrementally: the posterior for z^{n+1} with prior Π is equal to the posterior for a single data point z_{n+1} when the posterior for z^n is used as prior (in this sense the generalized posterior behaves like the standard posterior): using this to further rewrite the second line of (4.15) gives

$$\begin{aligned}&\pi(p \mid z^n, \eta) \\ &\quad\propto \int (f(y^n \mid x^n, \tau, p))^\eta \pi(\tau \mid p)\, d\tau\, \pi(p) \\ &\quad= \int (f(y_n \mid x_n, \tau, p))^\eta \cdot (f(y^{n-1} \mid x^{n-1}, \tau, p))^\eta \pi(\tau \mid p)\, d\tau\, \pi(p) \\ &\quad= \int (f(y_n \mid x_n, \tau, p))^\eta \\ &\qquad\cdot \left(\pi(\tau \mid z^{n-1}, p, \eta) \cdot \int (f(y^{n-1} \mid x^{n-1}, \tau')^\eta \pi(\tau' \mid p) d\tau' \right) d\tau\, \pi(p) \\ &\quad\propto \int (f(y_n \mid x_n, \tau, p))^\eta \cdot \pi(\tau \mid z^{n-1}, p, \eta)\, d\tau \cdot \pi(p \mid z^{n-1}, \eta),\end{aligned}$$

where in the third inequality we used the definition of the generalized posterior and in the last we used (4.15).

The integral appearing in both the cumulative and the step-wise expression equals the expectation in (4.9) from the η-flattened η-generalized Bayesian predictive density for n and 1 outcome respectively; $-\log[(\cdot)^{1/\eta}]$ of this quantity is the mix loss of model p. We will now derive formulas for this quantity.

Model with fixed variance Use the notation of Section 3.3.1. Write $\sigma^2_{\text{mix}} = \sigma^2(1/\eta + x_{n+1}\Sigma_n x_{n+1}^\top)$. Then the mix loss for predicting one new data point y_{n+1} is

$$-\log \bar{f}(y_{n+1} \mid x_{n+1}, z^n, \langle\eta\rangle; \eta) = \frac{1}{\eta}\left[\frac{1}{2}(\eta-1)\log(2\pi\sigma^2) + \frac{1}{2}\log\eta + \frac{1}{2}\log(2\pi\sigma^2_{\text{mix}}) + \frac{1}{2\sigma^2_{\text{mix}}}(y_{n+1} - x_{n+1}\beta_n)^2\right].$$

Model with conjugate prior on variance Using the notation of Section 3.3.1, the mix loss is given by

$$-\log \bar{f}(y_{n+1} \mid x_{n+1}, z^n, \langle\eta\rangle; \eta) = \frac{1}{\eta}\left[\frac{1}{2}\eta\log\pi + \frac{1}{2}\log(1+\eta x_{n+1}\Sigma_n x_{n+1}^\top) + a_{n+1}\log(2b_n + \frac{(y_{n+1}-x_{n+1}\beta_n)^2}{1/\eta + x_{n+1}\Sigma_n x_{n+1}^\top}) - a_n \log 2b_n - \log\frac{\Gamma(a_{n+1})}{\Gamma(a_n)}\right].$$

4.A.2 Belief in concentration (proof of Theorem 4.1)

For simplicity, we only give the proof for the unconditional case, in which the θ represent distributions P_θ on $z \in \mathcal{Z}$; extension to the conditional case is straightforward. For $0 < \eta < 1$, let $d_\eta(\theta^* \| \theta)$ denote the Rènyi divergence of order $1-\eta$ (Van Erven and Harremoës, 2014), i.e. $d_\eta(\theta^*\|\theta) = -\frac{1}{\eta}\log \mathbf{E}_{Z\sim\theta^*}\left(\frac{f_\theta(Z)}{f_{\theta^*}(Z)}\right)^\eta$. We first state a lemma, proved further below. In the lemma, as in the remainder of the proof, (θ^*, Z^n) is the random variable distributed according to the Bayesian distribution Π.

Lemma 4.2. *Let Θ, Π and π be as in the statement of Theorem 4.1. For every $1/2 \le \eta < 1$, $\epsilon > 0$, let $\bar{\Theta}_{\eta,\epsilon} := \{\theta \in \Theta \mid d_\eta(\theta^*\|\theta) > \epsilon\}$. For every $b > 0$ and every sample size n and setting $\epsilon := (b\log n)/(n\eta)$ and $c_\eta = (1-\eta)/(1+\eta(1-\eta))$, we have:*

$$\Pi\left(\Pi(\bar{\Theta}_{\eta,\epsilon} \mid Z^n) \ge n^{-bc_\eta}\right) \le 2\left(\sum_{\theta\in\Theta}\pi(\theta)^\eta\right)\cdot n^{-bc_\eta}.$$

In particular, if π is summable for some $\eta < 1$, then using $b = 1/c_\eta$, we get that the Bayesian probability that the posterior probability of the set of θ farther than $b(\log n)/n$ from θ^* exceeds $1/n$, is $O(1/n)$.

We proceed to prove Theorem 4.1 using this lemma. By the information inequality (Cover and Thomas, 1991), we have for every probability density $f \neq f_{\theta^*}$ that

$$\begin{aligned} D(\theta^*\|\theta) &= \mathbf{E}_{Z_n\sim P_{\theta^*}}[-\log f_\theta(Z_n) + \log f_{\theta^*}(Z_n)] \\ &\ge \mathbf{E}_{Z_n\sim P_{\theta^*}}[-\log f_\theta(Z_n) + \log f(Z)]. \end{aligned}$$

In particular this holds with $f = \bar{f} \mid Z^n$, the Bayes predictive distribution based on the sample seen so far. It then follows from (4.6) that

$$\bar{\delta}_n \leq \mathbf{E}_{\theta \sim \Pi | Z^n}[D(\theta^* \| \theta)] \tag{4.16}$$

Since π^η is decreasing in η, we may without loss of generality assume that the η mentioned in the theorem statement is at least $1/2$. Now note (Van Erven and Harremoës, 2014, Theorem 16) that for every $1/2 < \eta < 1$, $d_{1/2}(\theta^* \| \theta) \leq (\eta/(1-\eta)) \cdot d_\eta(\theta^* \| \theta)$. We also know from (Yang and Barron, 1999, Lemma 4) that the KL divergence $D(\theta^* \| \theta)$ satisfies $D(\theta^* \| \theta) \leq (2 + \log v) d_{1/2}(\theta^* \| \theta)$. Since trivially $d_\eta(\theta^* \| \theta) \leq \log v$, we have, with $C = \frac{\eta}{1-\eta} \cdot (2 + 2 \log v)$, for every $\epsilon > 0$, using (4.16),

$$\begin{aligned}
\bar{\delta}_n &\leq C \cdot \mathbf{E}_{\theta \sim \Pi | Z^n}[d_\eta(\theta^* \| \theta)] \\
&\leq C\Pi\left(d_\eta > \epsilon \mid Z^n\right) \log v + C\left(1 - \Pi\left(d_\eta > \epsilon \mid Z^n\right)\right) \epsilon \\
&\leq C\left(\Pi\left(d_\eta > \epsilon \mid Z^n\right) \log v + \epsilon\right),
\end{aligned}$$

so that $\Pi\left(d_\eta > \epsilon \mid Z^n\right) \geq (C^{-1}\bar{\delta}_n - \epsilon)/(\log v)$ and by Lemma 4.2, we have for $\epsilon = b(\log n)/(n\eta)$ as in the lemma, that

$$\Pi\left(\frac{C^{-1}\bar{\delta}_n - \epsilon}{\log v} \geq n^{-bc_\eta}\right) \leq 2\left(\sum_{\theta \in \Theta} \pi(\theta)^\eta\right) \cdot n^{-bc_\eta}.$$

Rewriting this expression, plugging in the value of ϵ and using $\eta \geq 1/2$, gives

$$\Pi\left(\bar{\delta}_n \geq C\left((\log v) n^{-bc_\eta} + \frac{2b(\log n)}{n}\right)\right) \leq 2\left(\sum_{\theta \in \Theta} \pi(\theta)^\eta\right) \cdot n^{-bc_\eta}. \tag{4.17}$$

The first part of the result follows by setting $b = a/c_\eta$. For the second result, note that the first result implies (take $a = 2$), by the union bound over sample sizes $1, \ldots, n$, that the Bayesian probability that $\mathbf{E}_{Z^n \sim \theta^*}[\Delta_n]$ exceeds $C_0 \sum_{i=1}^n (\log i)/i \asymp (\log n)^2$ is $O(1/n)$. Thus there exists C', C_0' such that the Bayesian probability that $\mathbf{E}_{Z^n \sim \theta^*}[\Delta_n]$ exceeds $C_0'(\log n)^2$ is bounded by C'/n. Thus for the probability in (4.8) we have

$$\begin{aligned}
\Pi\left(\Delta_n \geq C_2 \cdot n^{a'}\right) &= \Pi\left(\Delta_n \geq C_2 \cdot n^{a'}, \mathbf{E}_{Z^n \sim \theta^*}[\Delta_n] \geq C_0'(\log n)^2\right) \\
&\quad + \Pi\left(\Delta_n \geq C_2 \cdot n^{a'}, \mathbf{E}_{Z^n \sim \theta^*}[\Delta_n] < C_0'(\log n)^2\right) \\
&\leq \Pi\left(\mathbf{E}_{Z^n \sim \theta^*}[\Delta_n] \geq C_0'(\log n)^2\right) \\
&\quad + \Pi\left(\Delta_n \geq C_2 \cdot n^{a'}, \mathbf{E}_{Z^n \sim \theta^*}[\Delta_n] < C_0'(\log n)^2\right) \\
&\leq \frac{C'}{n} + \frac{C_0'(\log n)^2}{C_2 n^{a'}},
\end{aligned}$$

where in the final step we used Markov's inequality. The second result follows.

Proof of Lemma 4.2 Fix $A > 0$ and $\gamma > 0$. We have

$$\begin{aligned}
\Pi\left(\Pi(\bar{\Theta}_{\eta,\epsilon} \mid Z^n) \geq A\right) &= \Pi\left(\frac{\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)\cdot f_\theta(Z^n)}{\sum_{\theta\in\Theta} \pi(\theta)\cdot f_\theta(Z^n)} \geq A\right) \\
&= \Pi\left(\frac{\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)\cdot f_\theta(Z^n)}{f_{\theta^*}(Z^n)} \cdot \frac{f_{\theta^*}(Z^n)}{\sum_{\theta\in\Theta} \pi(\theta)\cdot f_\theta(Z^n)} \geq A\right) \\
&\leq \Pi\left(\frac{\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)\cdot f_\theta(Z^n)}{f_{\theta^*}(Z^n)} \geq A^{1+\gamma}\right) + \Pi\left(\frac{f_{\theta^*}(Z^n)}{\sum_{\theta\in\Theta} \pi(\theta)\cdot f_\theta(Z^n)} \geq A^{-\gamma}\right),
\end{aligned} \tag{4.18}$$

where we used the union bound. The first term is equal to, and can be further bounded as

$$\begin{aligned}
&= \Pi\left(\frac{\left(\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)\cdot f_\theta(Z^n)\right)^\eta}{(f_{\theta^*}(Z^n))^\eta} \geq A^{\eta(1+\gamma)}\right) \\
&\leq \Pi\left(\frac{\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)^\eta\cdot (f_\theta(Z^n))^\eta}{(f_{\theta^*}(Z^n))^\eta} \geq A^{\eta(1+\gamma)}\right) \\
&= \sum_{\theta^*} \pi(\theta^*) P_{\theta^*}\left(\frac{\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)^\eta\cdot (f_\theta(Z^n))^\eta}{(f_{\theta^*}(Z^n))^\eta} \geq A^{\eta(1+\gamma)}\right) \\
&\leq \sum_{\theta^*\in\Theta} \pi(\theta^*)\, \mathbf{E}_{Z^n\sim P_{\theta^*}}\left[\frac{\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)^\eta\cdot (f_\theta(Z^n))^\eta}{(f_{\theta^*}(Z^n))^\eta}\right]\cdot A^{-\eta(1+\gamma)} \\
&= \sum_{\theta^*\in\Theta} \pi(\theta^*) \sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)^\eta \cdot \left(\mathbf{E}_{Z\sim P_{\theta^*}}\left[\frac{(f_\theta(Z))^\eta}{(f_{\theta^*}(Z))^\eta}\right]\right)^n \cdot A^{-\eta(1+\gamma)} \\
&\leq \left(\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)^\eta\right) e^{-n\eta\epsilon}\cdot A^{-\eta(1+\gamma)}.
\end{aligned}$$

where the first inequality follows by differentiation to η (or equivalently, by monotonicity of ℓ^p-norms), the second is Markov's, and the third is the definition of Rènyi divergence.

The second term in (4.18) can be bounded as

$$\begin{aligned}
&\leq \Pi\left(\frac{f_{\theta^*}(Z^n)}{\pi(\theta^*)\cdot f_{\theta^*}(Z^n)} \geq A^{-\gamma}\right) = \Pi(\pi(\theta^*)^{-1+\eta} \geq A^{-(1-\eta)\gamma}) \\
&\leq \mathbf{E}_{\theta^*\sim\Pi}[\pi(\theta^*)^{-1+\eta}] A^{\gamma(1-\eta)} = \sum_{\theta^*} \pi(\theta^*)^\eta A^{\gamma(1-\eta)}.
\end{aligned}$$

Combining the upper bounds on the two terms on the right in (4.18), we get:

$$\Pi\left(\Pi(\bar{\Theta}_{\eta,\epsilon} \mid Z^n) \geq A\right) \leq \left(\sum_{\theta\in\bar{\Theta}_{\eta,\epsilon}} \pi(\theta)^\eta\right)\left(e^{-n\eta\epsilon}\cdot A^{-\eta(1+\gamma)} + A^{\gamma(1-\eta)}\right).$$

Now we plug in the chosen value of $\epsilon = (b \log n)/(n\eta)$ and we set $A = n^{-b/(\gamma+\eta)}$. With these values the second factor on the right becomes

$$e^{-n\eta\epsilon} \cdot A^{-\eta(1+\gamma)} + A^{\gamma(1-\eta)} \\ = n^{-b} n^{b(\eta(1+\gamma))/(\gamma+\eta)} + n^{-b\gamma(1-\eta)/(\gamma+\eta)} = 2n^{-b\cdot\gamma\cdot\frac{1-\eta}{\gamma+\eta}}.$$

Since this holds for all $\gamma > 0$, it also holds for $\gamma = 1/(1-\eta)$, and the result follows.