Minimax Optimality in Online Learning under Logarithmic Loss with Parametric Constant Experts

Fares Hedayati

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-213
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-213.html

December 16, 2013

Acknowledgement

To my dearest Niknaz, Maryam, Hamid, Shirin, and Nazanin

and to the Bahá'í Institute for Higher Education

Minimax Optimality in Online Learning under Logarithmic Loss with Parametric Constant Experts

by

Fares Hedayati

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter L. Bartlett (Chair)
Professor Martin J. Wainwright
Professor David Aldous

Fall 2013

Minimax Optimality in Online Learning under Logarithmic Loss with Parametric Constant Experts

Abstract

Minimax Optimality in Online Learning under Logarithmic Loss with Parametric Constant Experts

by

Fares Hedayati

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Peter L. Bartlett (Chair)

We study online prediction of individual sequences under logarithmic loss with parametric experts. The goal is to predict a sequence of outcomes $x_t \in \mathcal{X}$, revealed one at a time, almost as well as a set of experts. At round t, the forecaster's prediction takes the form of a conditional probability density $q_t(\cdot \mid x^{t-1})$, where $x^{t-1} \equiv (x_1, x_2, \cdots, x_{t-1})$. The loss that the forecaster suffers at that round is $-\log q_t(x_t \mid x^{t-1})$, where x_t is the outcome revealed after the forecaster's prediction. The performance of the prediction strategy is measured relative to the best in a reference set of experts, a parametric class of i.i.d distributions. The difference between the accumulated loss of the prediction strategy and the best expert in the reference set is called the regret. We focus on the minimax regret, which is the regret of the strategy with the minimum of the worst-case regret over outcome sequences.

The minimax regret is achieved by the normalized maximum likelihood (NML) strategy. This strategy knows the length of the sequence in advance and the probability it assigns to each sequence is proportional to the maximum likelihood of the sequence. Conditionals are computed at each round by marginalization which is very costly for NML. Due to this drawback, much focus has been given to alternative strategies such as sequential normalized maximum likelihood (SNML) and Bayesian strategies. The conditional probability that SNML assigns to the next outcome is proportional to the maximum likelihood of the data seen so far and the next outcome. We investigate conditions that lead to optimality of SNML and Bayesian strategies. A major part of this thesis is dedicated to showing that optimality of SNML and optimality of a certain Bayesian strategy, namely the Bayesian strategy under Jeffreys prior are equivalent to each other, i.e. if SNML is optimal, then so is the Bayesian strategy under Jeffreys prior and if the Bayesian strategy under the Jeffreys prior is optimal then so is SNML. Note that Jeffreys prior in parametric families is proportional to the square root of the determinant of the Fisher information. Furthermore we show that optimality of SNML happens if and only if the joint distribution on sequences defined by SNML is exchangeable, i.e. the probability that SNML assigns to any sequence is invariant under any

permutation of the sequence. These results are proven for exponential families and any parametric family for which the maximum likelihood estimator is asymptotically normal. The most important implication of these results is that when SNML–exchangeability holds NML becomes horizon–independent, and it could be either calculated through a Bayesian update with Jeffreys prior or through a one step-ahead maximum likelihood calculation as in SNML. Another major part of this thesis is focused on showing that SNML–exchangeabilty holds for a large class of one–dimensional exponential family distributions, namely for Gaussian, the gamma, and the Tweedie exponential family of order $3/2$, and any one–to–one transformation of them and that it cannot hold for other one–dimensional exponential family distributions.

Finally in this thesis we investigate horizon–dependent priors when Jeffreys prior is not optimal. Only Jeffreys prior can make a Bayesian strategy optimal. This means that if Jeffreys prior is not optimal then nor is any other prior, except for possibly a horizon–dependent prior. This is because if there does not exist a prior that can make the Bayesian strategy optimal for all horizons then the only possibilities are priors that depend on the horizon of the game. We investigate the behavior of a natural horizon–dependent prior called the NML prior. We show that the NML prior converges in distribution to Jeffreys prior, which makes it asymptotically optimal, but not necessarily optimal for an arbitrary horizon. Furthermore we show that there are exactly three families, namely Gaussian, gamma and inverse Gaussian, where the NML prior is equal to Jeffreys prior and hence horizon–independent. Two of these families namely gamma and Gaussian have optimal NML prior. We also investigate the problem of finding an optimal horizon-dependent prior for online binary prediction with Bernoulli experts. We could not solve this problem, but we describe insights gained from our investigation and possible directions that researchers can take in tackling this open problem.

To my dearest Niknaz, Maryam, Hamid, Shirin, and Nazanin
and to the Bahá'í Institute for Higher Education

Contents

Acknowledgments

I did part of my undergraduate and my entire graduate studies here at UC Berkeley. I deeply appreciate this unique educational opportunity. I was very fortunate to have the mentorship and support of Peter Bartlett. His patience, knowledge, and encouragement have always motivated me to go forward. The academic freedom he gave me to explore different topics and choose what intrigued me most helped me find the right research path that shaped my thesis. Prior to working with Peter Bartlett, I was under the mentorship of Kurt Keutzer for my master thesis. I want to thank him as his mentorship and frequent advice on how to succeed in graduate school helped me in my subsequent years at UC Berkeley. I had the great privilege to collaborate with Peter Grünwald, Peter Harremoës, Wojciech Kotłowski, Suvirt Sra, Jike Chong, and Alan Malek. I did part of my undergraduate studies at the Bahá'í Institute for Higher Education (BIHE). BIHE was established by the Bahá'í community of Iran in response to the Iranian government's widespread and systematic denial of higher education to Iranian Bahá'ís. Words cannot sufficiently describe my gratitude for the sacrifices our educators and administrators at BIHE have made for us. Many of them are serving long prison sentences in Iran solely for educating Iranian Bahá'í youth. I would like to dedicate this thesis to them. Last but not least, I would like to express my deep gratitude toward my family, my lovely wife Niknaz Aftahi, whose immense love, warmth, and kindness have been the light of my life, my wonderful sisters Shirin and Nazanin and my dearest parents Maryam and Hamid whose presence and support have always been with me.

Chapter 1

Introduction, Definitions and Notation

1.1 Online Learning under Logarithmic Loss

The aim of online learning under logarithmic loss is to predict a sequence of outcomes $x_t \in \mathcal{X}$, revealed one at a time, almost as well as a set of experts. At round t, the forecaster's prediction takes the form of a conditional probability density $q_t(\cdot \mid x^{t-1})$, where $x^{t-1} \equiv (x_1, x_2, \cdots, x_{t-1})$ and the density is with respect to a fixed measure λ on $\mathcal{X}$. For example, if $\mathcal{X}$ is discrete, λ could be the counting measure; for $\mathcal{X} = \mathbb{R}^d$, λ could be Lebesgue measure. The loss that the forecaster suffers at that round is $-\log q_t(x_t \mid x^{t-1})$, where x_t is the outcome revealed after the forecaster's prediction. The performance of the prediction strategy is measured relative to the best in a reference set of experts. The difference between the accumulated loss of the prediction strategy and the best expert in the reference set is called the regret. The focus of this thesis is on parametric constant experts. A *parametric constant expert* is a parameterized probability density p_θ such that for all $t > 0$ and for all $x \in \mathcal{X}$, $p_\theta(x \mid x^{t-1}) = p_\theta(x)$.

Definition 1 (Parametric Constant Model)**.** *A constant expert is an i.i.d stochastic process, that is, a joint probability distribution p on sequences of elements of $\mathcal{X}$ such that for all $t > 0$ and for all x in $\mathcal{X}$, $p(x^t \,|x^{t-1}) = p(x_t)$. A parametric constant model $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ is a parameter set Θ, a measurable space $(\mathcal{X}, \Sigma)$, a measure λ on $\mathcal{X}$, and a parameterized function $p_\theta : \mathcal{X} \to [0, \infty)$ for which, for all $\theta \in \Theta$, p_θ is a probability density on $\mathcal{X}$ with respect to λ. It defines a set of constant experts via $p_\theta(x^t \,|x^{t-1}) = p_\theta(x_t)$.*

We call any *sequential probability assignment* of the form $q_t(\cdot \mid x^{t-1})$, a *strategy*. Note that $q_1(x_1 \mid x^0) = q_1(x_1)$ and that any sequential probability assignment of length n defines a joint density with respect to measure λ on the n outcomes and vice versa. This is because the product of the conditional probability densities over a sequence of length n, defined by a

sequential probability assignment, integrates to one, and from any joint density on sequences of length n, conditional densities can be computed via marginalization.

The *exponential family* is a natural and widely studied class of probability distributions (see, for example, [2]). A set of distributions parametrized by $\theta \in \Theta$ on a set of outcomes $\mathcal{X}$ is an exponential family if it can be written as a density $p_\theta(x) = h(x)e^{\theta^T \phi(x) - \psi(\theta)}$ defined relative to a $\phi : \mathcal{X} \to R^k$ and a $h : \mathcal{X} \to [0, \infty)$. Some of the major results of this thesis use the natural exponential families as the expert set.

Definition 2 (Natural Exponential Family)**.** *Natural exponential family distributions on $\mathbb{R}^d$, are defined with the following probability distribution:*

$$p_\theta(x) = h(x) \exp\left(x^\intercal \theta - A(\theta)\right),$$

where $\theta \in \{\theta \in \mathbb{R}^d | A(\theta) < \infty\}, x \in \mathbb{R}^d$, h is a reference measure, and the log normalization A ensures that p_θ is a probability distribution.

Definition 3 (Regret)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model and let $q^{(n)}$ denote the joint density with respect to measure λ defined by the product of the n sequential probability assignments $q_t(x_t \mid x^{t-1})$. For any sequence x^n from $\mathcal{X}$, the regret of a strategy $q^{(n)}$ with respect to Θ is given by*

$$R^\Theta(x^n, q^{(n)}) = \sum_{t=1}^{n} -\log q_t(x_t \mid x^{t-1}) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} -\log p_\theta(x_t \mid x^{t-1}) = \sup_{\theta \in \Theta} \log \frac{p_\theta(x^n)}{q^{(n)}(x^n)}.$$

We consider a generalization of the regret of Definition 3. This is because some strategies are only defined conditioned on a fixed initial sequence of observations x^{m-1}. Refer to Section 2.3 for two examples of these kinds of strategies. For such cases we define the conditional regret of x^n, given a fixed initial sequence x^{m-1}, in the following way [see 11, chap. 11].

Definition 4 (Conditional Regret)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model, let x^{m-1} be a fixed sequence from $\mathcal{X}$, and let $q^{(n)}$ denote the joint density with respect to measure λ defined by the product of the n sequential probability assignments $q_t(x_t \mid x^{t-1})$ for $t \geq m$. For any sequence x_m^n, the conditional regret of a strategy $q^{(n)}$ given x^{m-1} is given by*

$$\begin{aligned} R^\Theta(x_m^n, q^{(n)} \mid x^{m-1}) &= \sum_{t=m}^{n} -\log q_t(x_t \mid x^{t-1}) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} -\log p_\theta(x_t \mid x^{t-1}) \\ &= \sup_{\theta \in \Theta} \log \frac{p_\theta(x^n)}{q^{(n)}(x_m^n \mid x^{m-1})}. \end{aligned}$$

Notice that the strategy $q^{(n)}$ defines only the conditional distribution $q^{(n)}(x_m^n \mid x^{m-1})$. We call such a strategy a conditional strategy. In what follows, where we consider a conditional strategy, we assume that x^{m-1} is such that these conditional distributions are always well defined.

The interest is in strategies whose average regret $\frac{1}{n}R^{\Theta}(x^n, q^{(n)})$ diminishes to zero as n grows larger. A strategy with this property has a predicting power almost as good as someone who observes the entire sequence of data and picks the best predicting strategy. Equivalently these strategies do not lose much from not knowing the future and perform almost as well as one who knows it. We are interested in *minimax optimal* strategies, which are strategies $q^{(n)}$ that minimize the supremum over sequences x^n of the regret $R^{\Theta}(x^n, q^{(n)})$ (or, in the case of conditional regret, the supremum over sequences x_m^n of the conditional regret given x^{m-1}). The NML strategy is the unique minimax optimal strategy (see Section 1.3). NML is not naturally defined in terms of conditionals. Conditionals are computed at each round by marginalization of the joint distribution which makes the NML strategy very costly. Due to this major drawback much focus has been given to alternative strategies such as sequential normalized maximum likelihood (SNML) and Bayesian strategies (see Sections 1.4, and 1.5). Much of the work of this thesis is in characterization of the minimax optimality conditions of these alternatives.

The next section looks at online learning under logarithmic loss from a different perspective: the MDL principle.

1.2 The MDL principle

Online learning under logarithmic loss is equivalent to data compression in *the minimum description length (MDL) principle* context [9]. Before giving a brief overview of this concept, we review some basics of information theory. Let $p(\,\cdot\,)$ be a distribution over sequences of data of length n from $\mathcal{X}^n$. A well-known result in information theory says that there is a prefix code, namely the Shannon-Fano code, that assigns to $x^n \in \mathcal{X}^n$ a codeword of length $L(X^n) = \lceil -\log p(x^n) \rceil$. This code is optimal, as it minimizes the expected code-length function $E_p L(X^n)$ [9].

The concept of the MDL principle was first introduced by Jorma Rissanen [18]. In short, the concept looks at learning as finding regularity in data and explaining it more succinctly; the more succinctly the data is explained the more regularity is found and a better *learning* is achieved. Equivalently learning is viewed as the ability to compress data.

Let model $\mathcal{M}$ be a class of probability distributions over sequences of data of length n. Note that each distribution corresponds to a code. We use hypothesis and code interchangeably. The MDL principle looks for a hypothesis or a combination of hypotheses in $\mathcal{M}$ that compresses the data the most. Before seeing the data both the sender and the receiver should agree on a code. The sender cannot pick a code after observing the data, because the receiver does not know which code was picked and can not decipher the encoded data

correctly. Among all codes in $\mathcal{M}$, the one that compresses x^n the most is the one that corresponds to the maximum likelihood estimate. This is because the prefix code with the smallest code-length is the prefix code corresponding to a codeword of length $\min_{p\in\mathcal{M}}\lceil -\log p(x^n)\rceil$. This code however cannot be used, because for the maximum likelihood to be obtained the data should be observed first. Is there a hypothesis or a combination of hypotheses from model $\mathcal{M}$ that can compress the data asymptotically as well as the best hypothesis? The answer is positive. Codes with this property are called *universal codes* or *universal models* where the emphasis in the latter is on distributions. *Normalized maximum likelihood (NML in short)*, *Sequential normalized maximum likelihood*, and *Bayesian strategies* are some of the well-studied universal models. Sections 1.3, 1.4, and 1.5 go over these universal models in detail.

The logarithmic loss in online learning translates to the number of bits needed to compress the data in MDL. Online learning defines regret as the difference between the loss of the player and the loss of the best expert in hindsight, whereas MDL views regret as the number of extra bits needed to compress the data, in comparison with the best code in hindsight. For further study of the MDL principle, see the book [9].

1.3 Normalized Maximum Likelihood (NML)

Definition 5 (sup-integrable)**.** *We say that for a given horizon n, the model $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ is sup-integrable if for all sequences y^n from $\mathcal{X}$*

$$\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n) < \infty;$$

furthermore, we say that for an initial subsequence x^{m-1} the model is conditionally sup-integrable if for all sequences y^{n-m} from $\mathcal{X}$

$$\int_{\mathcal{X}^{n-m+1}} \sup_{\theta\in\Theta} p_\theta(x^{m-1}y^{n-m+1})\, d\lambda^{n-m+1}(y^{n-m+1}) < \infty.$$

We call this integral the Shtarkov integral.

The following definition of NML is based on the assumption that the model is sup-integrable or conditionally sup-integrable. Throughout the thesis, we will assume that the model is either sup-integrable or conditionally sup-integrable for some fixed initial x^{m-1}, which makes the conditional NML well-defined.

Definition 6 (NML)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model which is sup-integrable. Given a fixed horizon n, the normalized maximum likelihood (NML) strategy with respect to this parametric constant model is defined via the joint probability distribution $p^{(n)}_{nml}$, defined as*

$$p^{(n)}_{nml}(x^n) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n)}.$$

If the model is not sup-integrable but it is conditionally sup-integrable for some x^{m-1}, then the conditional NML is defined as

$$p^{(n)}_{nml}(x^n_m|x^{m-1}) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^{n-m+1}} \sup_{\theta\in\Theta} p_\theta(x^{m-1}y^{n-m+1})\, d\lambda^{n-m+1}(y^{n-m+1})}.$$

As an example, the Gaussian distribution with fixed variance of 1 and mean $\mu \in \mathbb{R}$ and $n = 2$ is not sup-integrable. However it is conditionally sup-integrable for any x_1 which makes the conditional $p^{(2)}_{nml}(\,\cdot\,|x_1)$ well-defined. The same phenomena happens for n greater than 2.

The following theorem plays a central role in many of the results of this thesis. It basically says that the optimality of a strategy means that it has equal regrets for all sequences of the same length. It further states that an optimal strategy is equivalent to an NML.

Theorem 1.3.1 (Optimality)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model. For a fixed horizon n, a strategy $p^{(n)}$ is minimax optimal if and only if it is an equalizer, i.e. the regret of the strategy stays the same for all sequences of length n. Moreover, NML is the only minimax optimal strategy and the conditional NML is the only minimax optimal conditional strategy.*

Proof. First note that regret of NML for an arbitrary sequence x^n is :

$$-\log \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n)} - \left(-\log \sup_{\theta\in\Theta}\ p_\theta(x^n)\right) = \log \int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n),$$

which is independent of x^n. Therefore, NML has the same regret for all sequences of length n. Let strategy $p^{(n)}$ be an equalizer and let $q^{(n)}$ be a strategy different from $p^{(n)}$. Then for some z^n we should have $p^{(n)}(z^n) > q^{(n)}(z^n)$ which in turn makes the regret of $q^{(n)}$ for z^n larger than that of $p^{(n)}_{nml}$. If sequence w^n maximizes the regret of $q^{(n)}$ then:

$$R^\Theta(w^n, q^{(n)}) > R^\Theta(z^n, q^{(n)}) > R^\Theta(z^n, p^{(n)}) = R^\Theta(w^n, p^{(n)})$$

This means that for any strategy $q^{(n)}$ different from $p^{(n)}$, the maximum regret of $q^{(n)}$ over all sequences of length n is strictly greater than the maximum regret of $p^{(n)}$, therefore $p^{(n)}$ has the minimum value of the maximum regret, i.e. it is minimax optimal. On the other hand if $p^{(n)}$ is minimax optimal then it should be an equalizer, because if it is not, then an equalizer strategy such as NML, has lower maximum regret than $p^{(n)}$, as was shown in the if-part of the proof. Note that NML should be the only minimax optimal strategy, because as was shown in this proof, any strategy $q^{(n)}$ different from $p^{(n)}_{nml}$ should end up having a maximum regret strictly greater than that of the NML. Finally the conditional NML attains the minimax conditional regret bound because the conditional $p^{(n)}_{nml}(\,\cdot \mid x^{m-1})$ has equal conditional regret for any sequence of length n with the first $m-1$ outcomes identical to x^{m-1}. This property guarantees optimality, because the same argument can be applied to conditional regret. If a

conditional strategy $q^{(n)}(\,\cdot\,|x^{m-1})$ is different from $p^{(n)}_{nml}(\,\cdot\,|x^{m-1})$ then we can conclude that the maximum conditional regret of $q^{(n)}(\,\cdot\,|x^{m-1})$ is strictly greater than that of $p^{(n)}_{nml}(\,\cdot\,|x^{m-1})$. This in turn proves the minimax optimality of the conditional NML strategy. □

One major drawback of NML is that it is defined in terms of a joint distribution, not conditionals. Conditionals should be calculated by marginalization at each time which is very costly. For example to marginalize the first k random variables in a joint distribution of n binary random variables, in general, 2^{n-k} sums are needed. Another drawback is the strategy's dependence on horizon n. This violates the spirit of online learning, which emphasizes the player's lack of knowledge of future events.

1.4 Sequential Normalized Maximum Likelihood (SNML)

Definition 7 (SNML)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model. The sequential normalized maximum likelihood (SNML) update, is defined as*

$$p_{snml}(x_t \mid x^{t-1}) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^t)}{\int_{\mathcal{X}} \sup_{\theta\in\Theta} p_\theta(x^t)\, d\lambda(x_t)}$$

under the assumption that the denominator is finite.

Note that in some cases the model is not sup-integrable and hence SNML is not well-defined. The results of this thesis are based on the assumption that in those cases there exists an initial sequence x^{m-1} such that $p_{snml}(\,\cdot\,|x^{m-1})$ is well-defined. Example 2.3.2 in Chapter 2 goes over one of these cases. SNML in that example is not well-defined for the exponential distribution but the problem goes away by conditioning on the first observation. Note that the SNML update does not depend on the horizon; it is naturally defined in terms of conditionals. For more information about this strategy and its origin refer to [21], [20] and [24].

In this thesis, the notion of exchangeability of stochastic processes plays an important role in characterizing conditions that lead to optimality of SNML.

Definition 8 (Exchangeable)**.** *A stochastic process is called* exchangeable *if the joint probability does not depend on the order of observations, that is, for any $n > 0$, any $x^n \in \mathcal{X}^n$, and any permutation σ on $\{1, \ldots, n\}$, the probability of x^n is the same as the probability of x^n permuted by σ.*

1.5 Bayesian Strategies

Definition 9 (Bayesian)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model and let $\pi(\,\cdot\,)$ be a probability distribution on Θ. In a Bayesian strategy, the joint probability for t obser-*

vations x^t, is defined in the following way:

$$p_\pi(x^t) \;= \int_{\theta\in\Theta} p_\theta(x^t)\, d\pi(\theta).$$

The conditional probability distribution is:

$$p_\pi(x_t \mid x^{t-1}) = \frac{p_\pi(x^t)}{p_\pi(x^{t-1})}.$$

We denote the conditional Bayesian strategy for a fixed x^{m-1} as $p_\pi(x_m^n \mid x^{m-1})$. In this thesis, Jeffreys prior [12] plays a central role regarding optimality of Bayesian strategies.

Definition 10 (Jeffreys prior)**.** *Let $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ be a parametric constant model where $\Theta \subseteq \mathbb{R}^d$. Jeffreys prior has density over the parameter space Θ that is proportional to $\sqrt{|I(\theta)|}$, where I is the Fisher information at θ (that is, the variance of the score, $\partial/\partial\theta \ln p_\theta(X)$, where X has density p_θ).*

Note that if the normalization factor of Jeffreys prior is not finite, the Bayesian strategy is not defined. We will always assume that in those cases, conditioning on an initial sequence x^{m-1}, makes the conditional Bayesian strategy under Jeffreys prior well-defined.

Under mild conditions, for exponential family distributions the regret of a Bayesian strategy is no more than a data–independent constant plus the minimax regret. Moreover under Jeffreys prior, the regret asymptotically approaches the minimax regret [see 11, chaps. 7,8]. Restriction to a fixed suitably bounded subset $\Theta_0 \subset \Theta$ is required for these results to hold. More specifically Θ_0 should be an *ineccsi* subset of Θ. *ineccsi* stands for "interior (is) nonempty; closure (is) compact subset of interior". An ineccsi subset of Θ is a subset $\Theta_0 \subset \Theta$ such that the interior of Θ_0 is nonempty and the closure of Θ_0 is a compact subset of the interior of Θ [see 9, p. 209].

1.6 Thesis Roadmap

NML is the unique optimal strategy in online prediction of individual sequences under logarithmic loss with parametric experts. The main focus of this thesis is on the optimality of alternative strategies with much lower computational costs, and hence on when they are equivalent to NML.

In Chapter 2, we show that in exponential family distributions the SNML is optimal, meaning it is equivalent to NML, if and only if the joint distribution on sequences defined by SNML is exchangeable. This property also characterizes the optimality of a Bayesian prediction strategy for an exponential family. The optimal prior distribution is Jeffreys prior. Furthermore, we show that the optimality of SNML implies its equivalence to the Bayesian strategy under Jeffreys prior and vice versa. The main proof technique in showing this result is using an extension of de Finetti's theorem on exponential families.

In Chapter 3, we further extend the results of Chapter 2 to a much broader class of parametric models, a class for which the maximum likelihood estimator is asymptotically normal. The optimal prediction strategy, normalized maximum likelihood, depends on the number n of rounds of the game, in general. However, when a Bayesian strategy is optimal, normalized maximum likelihood becomes independent of n. Our proof uses this to exploit the asymptotics of normalized maximum likelihood. The asymptotic normality of the maximum likelihood estimator is responsible for showing that an optimal Bayesian strategy should necessarily use the Jeffreys prior.

In Chapter 4, we focus on characterization of one–dimensional exponential family distributions that make the corresponding SNML exchangeable. Chapter 2 showed that in exponential families a Bayesian prediction strategy with Jeffreys prior and sequential normalized maximum likelihood coincide and are optimal if and only if the latter is exchangeable, which occurs if and only if the optimal strategy can be calculated without knowing the time horizon in advance. We show that for one-dimensional exponential families SNML is exchangeable only for three classes of natural exponential family distributions, namely the Gaussian, the gamma, and the Tweedie exponential family of order $3/2$, and any one–to–one transformation of them.

In Chapter 5, we study Bayesian strategies with horizon–dependent priors. We showed in Chapters 2 and 3 that if a Bayesian prediction strategy is optimal then it necessarily uses Jeffreys prior. As a result, if Jeffreys prior is not optimal then nor is any other prior, except for possibly a horizon–dependent prior. We investigate the behavior of a natural horizon–dependent prior called the NML prior. We show that the NML prior converges in distribution to Jeffreys prior, which makes it asymptotically optimal, but not necessarily optimal for an arbitrary horizon. Furthermore, we show that there are exactly three families, namely Gaussian, gamma and inverse Gaussian, where the NML prior is equal to Jeffreys prior and hence horizon–independent. In two of these families namely gamma and Gaussian the NML prior makes the corresponding Bayesian strategies optimal. Finally we show that in terms of the maximum regret in Bayesian strategies, Jeffreys prior is not always better than the NML prior, and that the NML prior is not always better than Jeffreys prior.

Finally, Chapter 6 talks about the problem of finding an optimal horizon–dependent prior for online binary prediction with Bernoulli experts. Even though we could not solve this problem, we shared insights gained from our investigation.

Chapter 2

Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior

In this chapter we investigate when the *sequential normalized maximum likelihood* strategy is optimal. We show that SNML is optimal if and only if the joint distribution on sequences defined by SNML is exchangeable. This property also characterizes the optimality of a Bayesian prediction strategy for an exponential family. The optimal prior distribution is Jeffreys prior. Note that the results of this chapter are from the paper [11].

2.1 Introduction

As we saw in Theorem 1.3.1, the optimal strategy for sequential probability assignment is the NML strategy (see Definition 6). NML suffers from two major drawbacks: the horizon n of the problem needs to be known in advance, and the strategy can be computationally expensive since it involves marginalizing over subsequences. In this chapter, we consider the optimality of two approaches that address these difficulties: Bayesian strategies, and sequential normalized maximum likelihood strategy (SNML) (see Definition 7 and 9). We consider the questions: for what classes is SNML optimal; for what classes does there exist a prior for which the Bayesian strategy is optimal; and, in those cases, what is the optimal prior? For certain parametric classes of experts, Bayesian prediction with a particular choice of prior namely the Jeffreys prior (Definition 10) has been shown to be asymptotically optimal [see 9, chaps. 7,8]. SNML is within a constant of the minimax regret [14]. We give characterizations of the optimality of these strategies in terms of an elementary property of the joint distribution defined by the SNML strategy. We show that SNML is optimal precisely when

its joint distribution is exchangeable (see Definition 8). In the case of natural exponential family distributions on $\mathbb{R}^d$ (Definition 2), we show that the optimal strategy is a Bayesian strategy iff SNML is exchangeable and in this case the optimal prior is Jeffreys prior.

2.2 Main Results

First, we show in Theorem 2.2.2 that SNML and NML are equivalent if and only if p_{snml} is exchangeable. This happens only if NML is horizon-independent. Then, we show in Theorem 2.2.4 that exchangeability of p_{snml} further implies the equivalence of NML, the Bayesian strategy with Jeffreys prior, and SNML. This theorem shows that the SNML strategy and the Bayesian strategy with Jeffreys prior are optimal in this case.

Note that Theorems 2.2.2 and 2.2.4 are based on the assumption that if NML is not sup-integrable then there exists an initial x^{m-1}, so that all of the relevant conditional distributions, i.e. conditional NML, conditional SNML and conditional Bayesian strategy under Jeffreys prior, are defined. From now on, each time we mention NML, SNML, or Bayesian strategies we mean NML, SNML, or Bayesian strategies conditioned on a suitable sequence of length $m-1$.

When we consider the conditional distribution $p(x_m^n \mid x^{m-1})$ defined by a conditional strategy, we are interested in exchangeability of the conditional stochastic process, that is, invariance under any permutation that leaves x^{m-1} unchanged. Now we are ready to state and prove the main results of this chapter. The first result applies to any class (countable or uncountable) for which the conditional strategies SNML and NML are defined.

Lemma 2.2.1. *The conditional regret under SNML is equal to*

$$R^{\Theta}_{snml}(x^n \mid x^{m-1}) = \log \frac{\int \sup_\theta p_\theta(x^{n-1}, x)\, dx}{p_{snml}(x_m^{n-1} \mid x^{m-1})}.$$

Proof. Write the conditional regret under SNML in the following way.

$$\begin{aligned} & R^{\Theta}_{snml}(x^n \mid x^{m-1}) \\ & \quad \equiv \; R^{\Theta}(x_m^n, p_{snml} \mid x^{m-1}) \\ & \quad = \; \log \sup_{\theta\in\Theta} p_\theta(x^n) \; - \; \log p_{snml}(x_m^n \mid x^{m-1}) \\ & \quad = \; \log \frac{p_{\hat\theta}(x^n)}{p_{snml}(x_m^n \mid x^{m-1})}, \end{aligned}$$

where $\hat\theta$ is the maximum likelihood estimator of x^n. One the other hand

$$\begin{aligned} & p_{snml}(x_m^n \mid x^{m-1}) \\ & \quad = \; p_{snml}(x_n \mid x^{n-1}) p_{snml}(x_m^{n-1} \mid x^{m-1}) \\ & \quad = \; \frac{p_{\hat\theta}(x^n)}{\int \sup_\theta p_\theta(x^{n-1}, x)\, dx} p_{snml}(x_m^{n-1} \mid x^{m-1}). \end{aligned}$$

Combining the two previous equations, we get:

$$R^{\Theta}_{snml}(x^n \mid x^{m-1}) = \log \frac{\int \sup_\theta p_\theta(x^{n-1}, x)\, dx}{p_{snml}(x_m^{n-1} \mid x^{m-1})}. \tag{2.1}$$

□

Theorem 2.2.2. *Fix $m > 0$ and x^{m-1}, and assume that $p^{(n)}_{nml}(x_m^n|x^{m-1})$, and $p_{snml}(x_m^n|x^{m-1})$ are well defined. SNML is equivalent to NML and hence is minimax optimal if and only if p_{snml} is exchangeable.*

Proof. By Lemma 2.2.1 $R^{\Theta}_{snml}(x^n \mid x^{m-1}) = \log \frac{\int \sup_\theta p_\theta(x^{n-1},x)\, dx}{p_{snml}(x_m^{n-1}|\, x^{m-1})}$, which means that the regret is independent of the last observation.

Now, we show that if p_{snml} is exchangeable, then the regret becomes independent of other observations, which implies that it is an equalizer and hence (by Theorem 1.3.1) equivalent to NML. Let $y^n = x^{m-1}z_m^n$ be a sequence of observations where z_m^n is different from x_m^n. We show that the regret of y^n is equal to that of x^n. Under any permutation of x_m^n, $\sup_{\theta\in\Theta} p_\theta(x^n)$ does not change due to the fact that $p_\theta(x^n) = \prod_{i=1}^n p_\theta(x_i)$. On the other hand $p_{snml}(\cdot \mid x^{m-1})$ is exchangeable meaning that $p_{snml}(x_m^n \mid x^{m-1})$ is permutation invariant. Consequently, for any permutation σ of x^n that leaves x^{m-1} fixed, $R^{\Theta}_{snml}(x^n \mid x^{m-1}) = R^{\Theta}_{snml}(\sigma(x^n) \mid x^{m-1})$. These two properties give us the following.

$$\begin{aligned}
&R^{\Theta}_{snml}(x^{m-1}, x_m^n \mid x^{m-1}) \\
&= R^{\Theta}_{snml}(x^{m-1}, x_m, \ldots, x_{n-1}, y_m \mid x^{m-1}) \\
&= R^{\Theta}_{snml}(x^{m-1}, y_m, x_{m+1}, \ldots, x_{n-1}, x_m \mid x^{m-1}) \\
&= R^{\Theta}_{snml}(x^{m-1}, y_m, x_{m+1}, \ldots, x_{n-1}, y_{m+1} \mid x^{m-1}) \\
&= R^{\Theta}_{snml}(x^{m-1}, y_m, y_{m+1}, x_{m+2}, \ldots, x_{n-1}, x_{m+1} \mid x^{m-1}).
\end{aligned}$$

Continuing inserting y_{m+i} at the last position and swapping it with x_{m+i} we see that $R^{\Theta}_{snml}(x^n \mid x^{m-1}) = R^{\Theta}_{snml}(y^n \mid y^{m-1})$. This means that SNML is an equalizer and hence it is equivalent to conditional normalized maximum likelihood.

Now, we prove the other direction. If SNML is equivalent to NML, meaning that for any $n \geq m$ and any x_m^n,

$$p_{snml}(x_m^n \mid x^{m-1}) = p^{(n)}_{nml}(x_m^n \mid x^{m-1}) = \frac{p^{(n)}_{nml}(x^n)}{p^{(n)}_{nml}(x^{m-1})},$$

then SNML is exchangeable. This is because

$$p^{(n)}_{nml}(x^n) \propto \sup_\theta \prod_{i=1}^n p_\theta(x_i),$$

and as the denominator is unchanged, the probability becomes permutation invariant and hence exchangeable. That is for any n and x_m^n the conditional probability $p_{snml}(x_m^n \mid x^{m-1})$ is invariant over permutations of x_m^n. □

The next theorem shows that some Bayesian strategy is optimal for a natural exponential family iff SNML is exchangeable. In that case, the optimal prior is Jeffreys prior. For the proof of this theorem we need a different notion of exchangeability that we call *sum − exchangeability* and was introduced originally in [7]. De Finetti's theorem says that a binary stochastic process p is exchangeable if and only if it is a mixture of Bernoulli distributions, i.e. there exists a prior π such that for any $n > 0$ and any $x \in \{0,1\}^n$,

$$p(x^n) = \int_{\theta \in [0,1]} \theta^{\left(\sum_{i=1}^n x_i\right)} (1-\theta)^{\left(n - \sum_{i=1}^n x_i\right)} \pi(\theta)\, d\theta$$

and the prior π in this equation is unique. Diaconis and Freedman extended this to exponential families [7], as in Lemma 2.2.3. We need two definitions for this lemma. Here $x_1, x_2, x_3, \ldots$ is a sample path of a stochastic process p.

Definition 11 (sum-compatible)**.** *Let h be a non-negative, finite, and locally integrable Borel function on $\mathbb{R}^d$. We call a general stochastic process p on $\mathbb{R}^d$, sum-compatible with respect to h if $\forall\, n > 0$*

$$p\left(0 < h^{(n)}\left(\sum_{i=1}^n x_i\right) < \infty\right) = 1, \tag{2.2}$$

where $h^{(n)}$ is the nth convolution of h, i.e.

$$h^{(n)}(s) = \int \left(\prod_{i=1}^{n-1} h(x_i)\right) h\left(s - \sum_{i=1}^{n-1} x_i\right) d\,x_1 \cdots d\,x_{n-1} \tag{2.3}$$

Definition 12 (sum-exchangeable)**.** *Let h be a non-negative, finite, and locally integrable Borel function on $\mathbb{R}^d$. We call a general stochastic process p on $\mathbb{R}^d$, sum-exchangeable with respect to h if $\forall\, n > 0$, $\forall\, s \in \mathbb{R}^d$*

$$p\left(x_1, \ldots, x_n \,\middle|\, \sum_{i=1}^n x_i = s\right) = \frac{\prod_{i=1}^n h(x_i)}{h^{(n)}(s)}\,, \tag{2.4}$$

where $h^{(n)}$ is the nth convolution of h.

Lemma 2.2.3 ([7])**.** *Consider a natural exponential family $p_\theta(x) = h(x)e^{x^\intercal \theta - A(\theta)}$ over $\Theta = \{\theta \in \mathbb{R}^d \mid A(\theta) < \infty\}$ where the reference measure h is a non-negative, finite, and locally integrable Borel function on $\mathbb{R}^d$. A stochastic process p is a mixture of distributions from this family, i.e. there exists a probability density π with respect to the Lebesgue measure on Θ such that for any x^n, $p(x^n) = \int_\Theta \pi(\theta)\, p_\theta(x^n)\, d\theta$, if and only if p is sum-compatible and sum-exchangeable with respect to the reference measure h.*

Theorem 2.2.4. *Suppose the class of parametric constant experts is a natural exponential family with the reference measure h as defined in Lemma 2.2.3. Also fix $m > 0$ and x^{m-1}, and assume that $p^{(n)}_{nml}(x^n_m|x^{m-1})$, $p_\pi(x^n_m|x^{m-1})$ and $p_{snml}(x^n_m|x^{m-1})$ are well defined, where π is the Jeffreys prior. Then the following are equivalent.*

(a) SNML is exchangeable

(b) SNML = NML:
For all n and all x^n_m,
$$p^{(n)}_{nml}(x^n_m|x^{m-1}) = p_{snml}(x^n_m|x^{m-1}).$$

(c) SNML = Bayesian:
There is a prior π on Θ such that for all n and all x^n_m,
$$p_{snml}(x^n_m|x^{m-1}) = p_\pi(x^n_m|x^{m-1}).$$

(d) SNML = Bayesian with Jeffreys prior :
For all n and all x^n_m,
$$p^{(n)}_{snml}(x^n_m|x^{m-1}) = p_{\pi_J}(x^n_m|x^{m-1}).$$

(e) NML = Bayesian:
There is a prior π on Θ such that for all n and all x^n_m,
$$p_{nml}(x^n_m|x^{m-1}) = p_\pi(x^n_m|x^{m-1}).$$

(f) NML = Bayesian with Jeffreys prior:
For all n and all x^n_m,
$$p^{(n)}_{nml}(x^n_m|x^{m-1}) = p_{\pi_J}(x^n_m|x^{m-1}).$$

Proof. We prove the equivalence by showing that $(a) \iff (b) \Rightarrow (c) \Rightarrow (d) \Rightarrow (e) \Rightarrow (b)$ and finally $(b) \iff (f)$.
$(a) \iff (b)$: We showed this in Theorem 2.2.2.
$(b) \Rightarrow (c)$: $p_{snml}(x^n_m \mid x^{m-1}) = p^{(n)}_{nml}(x^n_m \mid x^{m-1})$
For ease of notation we let $q(x^n_m) \equiv p_{snml}(x^n_m \mid x^{m-1}) = p^{(n)}_{nml}(x^n_m \mid x^{m-1})$. Let $\sum_{i=1}^{m-1} x_i = t$, and let $\sum_{i=m}^{n} x_i = s$. The maximum likelihood estimator is then $\hat{\theta} = (\nabla A)^{-1}\left(\frac{s+t}{n}\right)$. Writing $\bar{x}_n = s - \sum_{i=m}^{n-1} \bar{x}_i$ and $\bar{x}_1 = x_1, \ldots, \bar{x}_{m-1} = x_{m-1}$, we have

$$\begin{aligned}
&p^{(n)}_{nml}\left(\sum_{i=1}^{n} x_i = s \,\middle|\, x^{m-1}\right) \\
&= \frac{\int \prod_{i=1}^{n} h(\bar{x}_i) e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})} d\bar{x}_m \cdots d\bar{x}_{n-1}}{p^{(n)}_{nml}(x^{m-1})} \\
&= \frac{\left(\prod_{i=1}^{m-1} h(x_i) \times e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})}\right) \times h^{(n-m+1)}(s)}{p^{(n)}_{nml}(x^{m-1})}.
\end{aligned}$$

This is exactly the density of $Y_n \equiv X_m + \ldots + X_n | X^{m-1} = x^{m-1}$ where X_i are random variables generated by NML of horizon n. By Lemma 3.1a in [7] this density function should be finite and positive with probability one under $p^{(n)}_{nml}$. Since $e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})}$ and $p^{(n)}_{nml}(x^{m-1})$ and $\prod_{i=1}^{m-1} h(x_i)$ are finite, so is $h^{(n-m+1)}(s)$. Clearly $h^{(n-m+1)}(s) > 0$ almost surely under $p^{(n)}_{nml}$. Hence the conditional NML which is equivalent to the conditional SNML is sum-compatible with respect to h. Furthermore, with the same notation, we have

$$
\begin{aligned}
& q\left(x_m^n \,\middle|\, \sum_{i=m}^{n} x_i = s\right) \\
&= \frac{q(x_m^n)}{\int q(\bar{x}_m^n)\, d\bar{x}_m \cdots d\bar{x}_{n-1}} \\
&= \frac{p^{(n)}_{nml}(x_m^n \mid x^{m-1})}{\int p^{(n)}_{nml}(\bar{x}_m^n \mid x^{m-1}) d\,\bar{x}_m \cdots d\,\bar{x}_{n-1}} \\
&= \frac{p^{(n)}_{nml}(x^n)/p^{(n)}_{nml}(x^{m-1})}{\int p^{(n)}_{nml}(x^{m-1}, \bar{x}_m^n) d\,\bar{x}_m \cdots d\,\bar{x}_{n-1}/p^{(n)}_{nml}(x^{m-1})} \\
&= \frac{\prod_{i=m}^{n} h(x_i) e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})}}{\int \prod_{i=m}^{n} h(\bar{x}_i) e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})} d\bar{x}_m \cdots d\bar{x}_{n-1}} \\
&= \frac{\prod_{i=m}^{n} h(x_i)}{h^{(n-m+1)}(s)}.
\end{aligned}
$$

Therefore, the conditional $p_{snml}(\cdot \mid x^{m-1})$ is sum-exchangeable with respect to h as well. Since $p_{snml}(\cdot \mid x^{m-1})$ is sum-compatible and sum-exchangeable with respect to h, Lemma 2.2.3 tells us that $p_{snml}(\,. \,| x^{m-1})$ is a mixture of $h(x)e^{x^\intercal \theta - A(\theta)}$, i.e. there exists a probability density π with respect to Lebesgue measure on Θ such that:

$$
p_{snml}(x_m^n \mid x^{m-1}) = \int p_\theta(x_m^n)\pi(\theta)\, d\theta. \tag{2.5}
$$

Now we let

$$
\pi_1(\theta) = K \times \frac{\pi(\theta)}{p_\theta(x^{m-1})} \tag{2.6}
$$

for a $K > 0$ chosen so that π_1 is a density. Substituting this into Equation (2.5) we get:

$$
p_{snml}(x_m^n \mid x^{m-1}) = \frac{\int_\Theta p_\theta(x^n)\pi_1(\theta)\, d\theta}{\int_\Theta p_\theta(x^{m-1})\pi_1(\theta)\, d\theta}. \tag{2.7}
$$

$(c) \Rightarrow (d)$: Now, we consider the regret of $p_{snml}(x_m^{n-1} \mid x^{m-1})$. As $p_{snml}(x_m^{n-1} \mid x^{m-1})$ is a Bayesian probability density (Equation (2.5)) the results on regrets of Bayesian strategies can be applied here. If the maximum likelihood estimator $\hat{\theta}$ lies in a fixed, bounded, closed

subset of Θ which is bounded away from the boundary of Θ, then the regret of a Bayesian strategy with prior w is [see 11, chaps. 8]:

$$\frac{d}{2}\log\frac{n}{2\pi} - \log w(\hat{\theta}) + \log\sqrt{\det I(\hat{\theta})} + o(1),$$

where I is the Fisher Information (see Definition 10). We apply this theorem to $z^{n-m+1} \equiv x_m^n$ and π. Note that $\hat{\theta}_{x_m^n}$ is the maximum likelihood estimator of x_m^n. The reason we can apply Grünwald's theorem here is twofold. First, the maximum likelihood estimator always exists because the family is full rank and A invertible. Second, the parameter space Θ is open and for any maximum likelihood estimator there should exist a bounded subset that contains the maximum likelihood estimator and is bounded away from the boundary of the parameter space. Let's denote the regret of a Bayesian strategy with prior π on a sequence z^p by $R_\pi^\Theta(z^p)$ and the regret of SNML on z^p by $R_{snml}^\Theta(z^p)$. Then

$$R_\pi^\Theta(x_m^n) = R^\Theta\left(p_{snml}\left(\cdot \mid x^{m-1}\right), x_m^n\right) = \frac{d}{2}\log\frac{n_1}{2\pi} - \log\pi(\hat{\theta}_{x_m^n}) + \log\sqrt{\det I(\hat{\theta}_{x_m^n})} + o(1),$$

where $n_1 = n - m + 1$. However, here we are calculating the conditional regret. It is easy to verify the following relationship:

$$R^\Theta\left(p_{snml}\left(\cdot \mid x^{m-1}\right), x_m^n\right) = R_{snml}^\Theta\left(x_m^n \mid x^{m-1}\right) - \log\sup_\theta p_\theta(x^n) + \log\sup_\theta p_\theta(x_m^n).$$

Hence for conditional SNML we get the following:

$$\begin{aligned} R_{snml}^\Theta(x_m^n \mid x^{m-1}) &= R_\pi^\Theta(x_m^n) + \log\sup_\theta p_\theta(x^n) - \log\sup_\theta p_\theta(x_m^n) \\ &= \frac{d}{2}\log\frac{n_1}{2\pi} - \log\pi(\hat{\theta}_{x_m^n}) + \log\sqrt{\det I(\hat{\theta}_{x_m^n})} + o(1) + \log\frac{p_{\hat{\theta}_{x^n}}(x^n)}{p_{\hat{\theta}_{x_m^n}}(x_m^n)}. \end{aligned} \tag{2.8}$$

If conditional SNML is Bayesian then it is exchangeable and by $(a) \Rightarrow (b)$, conditional SNML is also equivalent to conditional NML and hence has equal regret for all x_m^n. Consequently the conditional regret in (2.8) should not vary for fixed n and different x_m^n. We denote the value of this regret as $c_{n_1}(x^{m-1})$, emphasizing the fact that it depends on n_1 and x^{m-1} only. Simplifying (2.8) we get

$$\pi(\hat{\theta}_{x_m^n}) = \left(\frac{n_1}{2\pi}\right)^{d/2} \times \sqrt{\det I(\hat{\theta}_{x_m^n})} \times \frac{e^{o(1)}}{c_{n_1}(x^{m-1})} \times \frac{p_{\hat{\theta}_{x^n}}(x^n)}{p_{\hat{\theta}_{x_m^n}}(x_m^n)}. \tag{2.9}$$

Fix $\theta_0 = \hat{\theta}_{x_m^n}$. We let $N = kn_1$ (k is a positive integer). There exists a sequence y^N whose maximum likelihood estimator is θ_0. This sequence is nothing but k copies of x_m^n,

concatenated. The family is of full rank, therefore A is strictly convex and its gradient invertible. This means $\hat{\theta}_{Y^N}$, the maximum likelihood of Y^N, is

$$\hat{\theta}_{Y^N} = (\nabla A)^{-1}\left(\frac{\sum_{i=1}^{N} y_i}{N}\right) = (\nabla A)^{-1}\left(\frac{k \times \sum_{i=m}^{n-1} x_i}{n_1 k}\right) = (\nabla A)^{-1}\left(\frac{\sum_{i=m}^{n} x_i}{n_1}\right) = \hat{\theta}_{x_m^n} = \theta_0.$$

As N grows to infinity then $\hat{\theta}_{(x^m Y^N)} \to \hat{\theta}_{Y^N} = \theta_0$. This means that $\frac{p_{\hat{\theta}_{x^n}}(x^n)}{p_{\hat{\theta}_{x_m^n}}(x_m^n)}$ in Equation (2.9) converges to $p_{\theta_0}(x^{m-1})$ as $N \to \infty$. Using this and Equation (2.9) $\lim_{N\to\infty} \pi(\hat{\theta}_{Y^N})$ converges to:

$$\pi(\theta_0)\sqrt{\det I(\theta_0)}p_{\theta_0}(x^{m-1})\left(\lim_{N\to\infty}\left(\frac{N}{2\pi}\right)^{d/2}\frac{1}{c_N(x^{m-1})}\right).$$

Since $c_N(x^{m-1})$ does not depend on θ_0, $\pi(\theta_0) = c(x^{m-1})p_{\theta_0}(x^{m-1})\sqrt{\det I(\theta_0)}$, for some function c. Hence $\pi(\theta) \propto p_\theta(x^{m-1})\sqrt{\det I(\theta)}$, which in turn by Equation (2.6) means $\pi_1(\theta) \propto \sqrt{\det I(\theta)}$.
$(d) \Rightarrow (e)$: This is because, SNML being Bayesian implies exchangeability of SNML and hence SNML is equal to NML (by $(a) \Rightarrow (b)$) which makes NML Bayesian too.
$(e) \Rightarrow (b)$: NML being Bayesian means that there exists a prior π, such that for any $n > m$ and x_m^n we have

$$p_{nml}^{(n)}(x_m^n \mid x^{m-1}) = \frac{\int p_\theta(x^n)\pi(\theta)d\theta}{\int p_\theta(x^{m-1})\pi(\theta)d\theta}.$$

For $n \geq m$, let $A(n)$ be:

$$A(n) = \int \sup_\theta p_\theta(x^{m-1}, z^{n-m+1})\, dz^{n-m+1}.$$

With this new definition we get :

$$p_{nml}^{(n-1)}(x_m^{n-1} \mid x^{m-1}) = \frac{\sup_\theta p_\theta(x^{n-1})}{A(n-1)}.$$

We can also get $p_{nml}^{(n-1)}$ by marginalizing $p_{nml}^{(n)}$ (remember NML is horizon–independent because it is Bayesian). Then for $n > m$:

$$p_{nml}^{(n-1)}(x_m^{n-1} \mid x^{m-1}) = \int_x p_{nml}^{(n)}(x_m^{n-1}, x \mid x^{m-1})dx = \int_x \sup_\theta \frac{p_\theta(x^{n-1}, x)}{A(n)}\, dx.$$

Therefore for $n > m$,

$$\frac{\sup_\theta p_\theta(x^{n-1})}{A(n-1)} = \int_x \sup_\theta \frac{p_\theta(x^{n-1}, x)}{A(n)}\, dx.$$

Hence

$$\int_x \sup_\theta p_\theta(x^{n-1}, x)\, dx = \frac{A(n)}{A(n-1)} \sup_\theta p_\theta(x^{n-1}). \tag{2.10}$$

This is also true for $n = m$ if we define

$$A(m-1) = \sup_\theta p_\theta(x^{m-1}).$$

We know from Lemma 2.2.1 that the conditional regret of x^n under SNML is

$$R^\Theta_{snml}(x^n \mid x^{m-1}) = \log \left(\frac{\int \sup_\theta p_\theta(x^{n-1}, x)\, dx}{p_{snml}(x_m^{n-1} \mid x^{m-1})} \right).$$

Using Equation (2.10) we get

$$\begin{aligned} R^\Theta_{snml}(x^n \mid x^{m-1}) &= \log \left[\frac{A(n)}{A(n-1)} \times \frac{\sup_\theta p_\theta(x^{n-1})}{p_{snml}(x_m^{n-1}|x^{m-1})} \right] \\ &= R^\Theta_{snml}(x^{n-1}|x^{m-1}) + \log \frac{A(n)}{A(n-1)}. \end{aligned}$$

Continuing this we get

$$\begin{aligned} R^\Theta_{snml}(x^n|x^{m-1}) &= R^\Theta_{snml}(x^{m-1}|x^{m-1}) + \sum_{i=m}^{n} \log \frac{A(i)}{A(i-1)} \\ &= \log \sup_\theta p_\theta(x^{m-1}) + \log \frac{A(n)}{A(m-1)} \\ &= \log A(n). \end{aligned}$$

This shows that the conditional regret is fixed for a fixed x^{m-1} and hence the conditional SNML is an equalizer and equivalent to conditional NML (Theorem 1.3.1).
$(e) \Rightarrow (f)$: If NML is Bayesian then it is equal to SNML and therefore SNML is Bayesian with Jeffreys prior and hence so is NML. This is by $(e) \Rightarrow (b) \Rightarrow (c) \Rightarrow (d)$.
$(f) \Rightarrow (e)$: This is trivial because Bayesian with Jeffreys prior is a special case of being Bayesian.

□

2.3 Examples

Example 2.3.1 (Bernoulli Distribution)**.** *In this setting, the experts are Bernoulli distributions,* $p_\mu(x^n) = \mu^{\left(\sum_{i=1}^n x_i\right)}(1-\mu)^{\left(n-\sum_{i=1}^n x_i\right)}$ *with parameter space* $(0,1)$*. Converting this to the natural form we get* $p_\theta = \exp\left(\sum_{i=1}^n x_i\theta - \log\left(e^\theta + 1\right)\right)$ *with* $\Theta = \mathbb{R}$*, where we use the*

transformation $\theta = \ln \frac{\mu}{1-\mu}$. *Consider* $x^5 = (10011)$ *and* $y^5 = (10110)$, x^5 *is a permutation of* y^5. *However* $p_{snml}(x^5) = 0.00930 \neq p_{snml}(y^5) = 0.00932$ *which in turn means that* $p_{snml}(\,.\,)$ *is not exchangeable. Therefore, SNML and NML cannot be equivalent and neither is equivalent to a Bayesian strategy. It turns out that the regret of SNML in this case is better than Bayesian with Jeffreys prior but worse than NML [1].*

Example 2.3.2 (Exponential Distribution)**.** *The distributions are of the form* $p_\theta(x) = \frac{1}{\theta} e^{-x/\theta}$ *with* $\Theta = (0, \infty)$. *It is easy to check that for* $n = 1$, $p_{snml}(x) \propto \frac{1}{x} e^{-x/x} \propto \frac{1}{x}$ *which is not integrable. Jeffreys prior is proportional to* $1/\theta$ *which is not integrable either. However for any* x_1, *subsequent conditionals for Bayesian with Jeffreys prior and SNML will be properly defined. For* $n > 1$ *the maximum likelihood estimator for* θ *is* $\frac{1}{\frac{\sum_{i=1}^n x_i}{n}}$ *and therefore* $p_{snml}(x_n \mid x^{n-1})$ *is proportional to*

$$\sup_\theta p_\theta(x^n) \;=\; \left(\frac{1}{\frac{\sum_{i=1}^n x_i}{n}}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i}{\frac{\sum_{i=1}^n x_i}{n}}\right) \propto \frac{1}{(\sum_{i=1}^n x_i)^n}.$$

Normalizing this we get

$$p_{snml}(x_n \mid x^{n-1}) = \frac{\left(\frac{1}{\sum_{i=1}^n x_i}\right)^n}{\int_0^\infty \left(\frac{1}{\sum_{i=1}^n x_i}\right)^n d\,x_n} = \frac{(n-1)\left(\sum_{i=1}^{n-1} x_i\right)^{n-1}}{\left(\sum_{i=1}^n x_i\right)^n}.$$

The conditional SNML becomes:

$$\begin{aligned} p_{snml}(x_2^n \mid x_1) &= \frac{(2-1)\left(\sum_{i=1}^{2-1} x_i\right)^{2-1}}{\left(\sum_{i=1}^2 x_i\right)^2} \times \frac{(3-1)\left(\sum_{i=1}^{3-1} x_i\right)^{3-1}}{\left(\sum_{i=1}^3 x_i\right)^3} \cdots \times \frac{(n-1)\left(\sum_{i=1}^{n-1} x_i\right)^{n-1}}{\left(\sum_{i=1}^n x_i\right)^n} \\ &= \frac{(n-1)!\,x_1}{\left(\sum_{i=1}^n x_i\right)^n}. \end{aligned}$$

As $p_{snml}(x_2^n \mid x_1)$ *depends on* $\sum_{i=1}^n x_i$ *only, we get exchangeability, which in turn implies that SNML and NML are equivalent. On the other hand, the exponential distribution can be converted to an instance of a natural exponential family distribution by the change of variable* $\lambda = \frac{-1}{\theta}$. *Hence Theorem 2.2.4 implies that SNML and NML are also equivalent to the Bayesian strategy with Jeffreys prior, conditioned on the first observation. It is straightforward to verify this.*

Chapter 3

Optimality of SNML and Bayesian Strategy under Jeffreys Prior in Parametric Families with Asymptotically Normal Maximum Likelihood Estimators

In this chapter we study online learning under logarithmic loss with regular parametric models. We show that a Bayesian strategy predicts optimally only if it uses Jeffreys prior. We showed this result for natural exponential families in the previous chapter; we extend that to parametric models for which the maximum likelihood estimator is asymptotically normal. The optimal prediction strategy, normalized maximum likelihood, depends on the number n of rounds of the game, in general. However, when a Bayesian strategy is optimal, normalized maximum likelihood becomes independent of n. Our proof uses this to exploit the asymptotics of normalized maximum likelihood. The asymptotic normality of the maximum likelihood estimator is responsible for the necessity of Jeffreys prior. Note that the results of this chapter are from the paper [10].

3.1 Introduction

The optimal strategy for sequentially assigning probability to outcomes is known to be normalized maximum likelihood (see Theorem 1.3.1). NML suffers from two major drawbacks: the horizon n of the problem needs to be known in advance, and the strategy can be computationally expensive since it involves marginalizing over subsequences. In this chapter, we investigate the optimality of two alternative strategies, namely the Bayesian strategy and the sequential normalized maximum likelihood strategy (see Definitions 7 and 9). We show

that for a very general class of parametric models (Definition 1), optimality of a Bayesian strategy means that the strategy uses Jeffreys prior. Furthermore we show that optimality of the Bayesian strategy is equivalent to optimality of sequential normalized maximum likelihood. The major regularity condition for these parametric families is that the maximum likelihood estimator is asymptotically normal. This classical condition holds for a broad class of parametric models.

3.2 Definitions and Notation

The asymptotic normality of the maximum likelihood estimator is the major regularity condition of the parametric models that is required for our main result to hold.

Definition 13 (Asymptotic Normality of MLE)**.** *Consider a parametric constant model* $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ *with* $\Theta \subseteq \mathbb{R}^d$*. We say that the parametric model has an asymptotically normal MLE if, for all* $\theta_0 \in \Theta$*,*

$$\sqrt{n}\left(\hat{\theta}_{(x^n)} - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}\left(\theta_0\right)\right),$$

where $I(\theta)$ *is the Fisher information at* θ*,* x^n *is a sample path of* p_{θ_0}*, and* $\hat{\theta}_{(x^n)}$ *is the maximum likelihood estimator of* θ *given* x^n*, that is,* $\hat{\theta}_{(x^n)}$ *maximizes* $p_\theta(x^n)$*.*

Asymptotic normality holds for regular parametric models; for typical regularity conditions, see for example, Theorem 3.3 in [16].

For parametric models whose maximum likelihood estimators take values in a countable set, we need the notion of a lattice MLE.

Definition 14 (Lattice MLE)**.** *Consider a parametric model* $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ *with* $\Theta \subseteq \mathbb{R}^d$*. The parametric model is said to have a lattice MLE with diminishing step-size* h_n*, if for any* θ*, the possible maximum likelihood estimators of* n *i.i.d random variables generated by* p_θ *are points in* Θ *that are of the form* $(b + k_1 h_n, b + k_2 h_n, \cdots, b + k_d h_n)$*, for some integers* $k_1, k_2, \cdots, k_d$ *and some real numbers* b *and* h_n*. Additionally* h_n *is positive and diminishes to zero as* n *goes to infinity.*

3.3 Main Result

We show that in parametric models with an asymptotically normal MLE, the optimality of a Bayesian strategy implies that the strategy uses Jeffreys prior. Furthermore, we show that the optimality of a Bayesian strategy is equivalent to the optimality of sequential normalized maximum likelihood. This extends the result for natural exponential family distributions from the previous chapter to regular parametric models. Note that NML is the unique

optimal strategy (Theorem 1.3.1), so when we say that some other strategy is equivalent to NML, that is the same as saying that strategy predicts optimally.

The results shown in this part are based on the assumption that if NML is not sup-integrable then there exists an initial sequence x^{m-1} such that conditional NML, conditional SNML and conditional Bayesian strategy under Jeffreys prior are all well-defined. From now on, each time we mention NML, SNML, or Bayesian strategies we mean NML, SNML, or Bayesian strategies conditioned on a suitable initial sequence of length $m-1$.

Theorem 3.3.1. *Suppose we have a parametric model $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ with an asymptotically normal MLE. Assume that the MLE has a density with respect to Lebesgue measure or that the model has a lattice MLE with diminishing step-size h_n. Also assume that $I(\theta)$, the Fisher information at θ is continuous in θ, and that, for all x, $p_\theta(x)$ is continuous in θ. Also fix $m > 0$ and x^{m-1}, and assume that $p^{(n)}_{nml}(x_m^n|x^{m-1})$, $p_\pi(x_m^n|x^{m-1})$ and $p_{snml}(x_m^n|x^{m-1})$ are well defined, where π is the Jeffreys prior. Then the following are equivalent.*

(a) NML = Bayesian:
There is a prior π on Θ such that for all n and all x_m^n,

$$p^{(n)}_{nml}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}).$$

(b) NML = SNML:
For all n and all x_m^n,

$$p^{(n)}_{nml}(x_m^n|x^{m-1}) = p_{snml}(x_m^n|x^{m-1}).$$

(c) NML = Bayesian with Jeffreys prior:
For all n and all x_m^n,

$$p^{(n)}_{nml}(x_m^n|x^{m-1}) = p_{\pi_J}(x_m^n|x^{m-1}).$$

(d) $p_{snml}(\,\cdot\,|x^{m-1})$ is exchangeable.

(e) SNML = Bayesian:
There is a prior π on Θ such that for all n and all x_m^n,

$$p_{snml}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}).$$

(f) SNML = Bayesian with Jeffreys prior:
For all n and all x_m^n,

$$p_{snml}(x_m^n|x^{m-1}) = p_{\pi_J}(x_m^n|x^{m-1}).$$

Proof. We prove that (a), (b), and (c) are equivalent, and that (d), (e), and (f) are equivalent. The equivalence of (b) and (d) is Theorem 2.2.2.

(a) ⇒ (b): NML being equivalent to a Bayesian strategy means that NML is horizon-independent. Hence for any $m-1 < t \leq n$,

$$p^{(n)}_{nml}(x_t|x^{t-1}) = p_\pi(x_t|x^{t-1}) = p^{(t)}_{nml}(x_t|x^{t-1}) = p_{snml}(x_t|x^{t-1}),$$

which means that NML is equivalent to SNML.
(b) ⇒ (c): We use the asymptotic normality property to prove this below.
(c) ⇒ (a): This is immediate.
(d) ⇒ (e): We know that (d) and (b) are equivalent by Theorem 2.2.2, and that (b) implies (a), but (b) and (a) together imply (e).
(e) ⇒ (d): Since SNML is Bayesian, $p_{snml}(x^n) = \int \prod_{i=1}^n p_\theta(x_i)\, d\,\pi(\theta)$ for some prior distribution π on Θ. As $\prod_{i=1}^n p_\theta(x_i)$ does not depend on the order of observations, SNML is exchangeable.
(e) ⇒ (f): (e) implies (d), which implies both (b) and (c), and together these imply (f).
(f) ⇒ (e): This is immediate.
The heart of the proof is verifying that
(b) ⇒ (c): NML being equivalent to SNML means that, for all $m-1 \leq t \leq n$,

$$\begin{aligned} p_{snml}(x^t \mid x^{m-1}) &= p^{(n)}_{nml}(x^t \mid x^{m-1}) \\ &= \frac{\int \sup_\theta p_\theta(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\int \sup_\theta p_\theta(x^{m-1}, y^{n-m+1}) d\,\lambda^{n-m+1}(y^{n-m+1})} \\ &= \frac{\int p_{\hat\theta_{(x^t,y^{n-t})}}(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\int p_{\hat\theta_{(x^{m-1},y^{n-m+1})}}(x^{m-1}, y^{n-m+1}) d\,\lambda^{n-m+1}(y^{n-m+1})}, \end{aligned} \tag{3.1}$$

where $\hat\theta_{(x^t,y^{n-t})}$ is the maximum likelihood estimator upon observing x^t, y^{n-t}. As n goes to infinity, $\hat\theta_{(x^t,y^{n-t})}$ converges to $\hat\theta_{y^{n-t}}$. This is because as n goes to infinity, $\frac{1}{n}\left[\sum_{i=1}^t \log p_\theta(x_i)\right]$ in the following equation goes to zero :

$$\hat\theta_{(x^t,y^{n-t})} = \arg\max_{\theta\in\Theta} \frac{1}{n}\left[\sum_{i=1}^t \log p_\theta(x_i) + \sum_{j=1}^{n-t} \log p_\theta(y_j)\right].$$

Now we rewrite Equation (3.1) in a different form. Let $C^{\theta_0}_{\Delta\theta}$ be a hypercube centered at θ_0 with all sides having length h, where $\Delta\theta = h^d$, is the volume of the hypercube. Define

$$S^n_{x^t}(\theta_0) = \left\{ z^{n-t} \,\middle|\, \hat\theta_{(x^t,z^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}} \right\},$$

where $C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}$ is a hypercube that has volume $\Delta\theta/\sqrt{n^d}$ with all sides having length equal to $h/\sqrt{n}$. Let $P^\Theta_{\Delta\theta/\sqrt{n^d}}$ be the largest collection of disjoint hypercubes $C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}$ that fit in Θ.

Note that as $\Delta\theta$ goes to zero $P^{\Theta}_{\Delta\theta/\sqrt{n^d}}$ covers the whole Θ. Define

$$g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^t}(\theta_0)} p_{\theta_0}(x^t) p_{\theta_0}(y^{n-t}) d\,\lambda^{n-t}\,(y^{n-t})}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^{m-1}}(\theta_0)} p_{\theta_0}(x^{m-1}) p_{\theta_0}(y^{n-m+1}) d\,\lambda^{n-m+1}\,(y^{n-m+1})}.$$

First of all we show that

$$\lim_{n\to\infty} \lim_{\Delta\theta\to 0} \mid g^n(x^t, x^{m-1}, \Delta\theta) - p^{(n)}_{nml}(x^t \mid x^{m-1}) \mid = 0.$$

Since for all n, we have $p_{snml}(x^t \mid x^{m-1}) = p^{(n)}_{nml}(x^t \mid x^{m-1})$ this implies that $g^n(x^t, x^{m-1}, \Delta\theta)$ converges to $p_{snml}(x^t \mid x^{m-1})$. Then we show that the limit of $g^n(x^t, x^{m-1}, \Delta\theta)$ as n goes to infinity and $\Delta\theta$ goes to zero is a Bayesian conditional under Jeffreys prior. Now, it is easy to see the following:

$$p^{(n)}_{nml}(x^t \mid x^{m-1}) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^t}(\theta_0)} p_{\hat{\theta}_{(x^t, y^{n-t})}}(x^t) p_{\hat{\theta}_{(x^t, y^{n-t})}}(y^{n-t}) d\,\lambda^{n-t}\,(y^{n-t})}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^{m-1}}(\theta_0)} p_{\hat{\theta}_{(x^{m-1}, y^{n-m+1})}}(x^{m-1}) p_{\hat{\theta}_{(x^{m-1}, y^{n-m+1})}}(y^{n-m+1}) d\,\lambda^{n-m+1}\,(y^{n-m+1})}.$$

The only difference between this and $g^n(x^t, x^{m-1}, \Delta\theta)$ is that instead of θ_0 we have the parameter $\hat{\theta}_{(x^{m-1}, y^{n-m+1})}$ for each hypercube. The distance between two points in each hypercube is at most $h\sqrt{d/n}$, hence

$$\left|\theta_0 - \hat{\theta}_{(x^t, y^{n-t})}\right| \leq h\sqrt{\frac{d}{n}}.$$

As $\Delta\theta$ and consequently h go to zero, θ_0 converges to $\hat{\theta}_{(x^t, y^{n-t})}$ for the expressions in the numerator and to $\hat{\theta}_{(x^{m-1}, y^{n-m+1})}$ for those in the denominator. Due to the continuity of the likelihood for each hypercube in the numerator, we have

$$\lim_{\Delta\theta\to 0} p_{\theta_0}\left(x^t, y^{n-t}\right) = p_{\hat{\theta}_{(x^t, y^{n-t})}}\left(x^t, y^{n-t}\right).$$

Similarly, for each hypercube in the denominator we have

$$\lim_{\Delta\theta\to 0} p_{\theta_0}\left(x^{m-1}, y^{n-m+1}\right) = p_{\hat{\theta}_{(x^{m-1}, y^{n-m+1})}}\left(x^{m-1}, y^{n-m+1}\right).$$

Hence $g^n(x^t, x^{m-1}, \Delta\theta)$ converges to $p^{(n)}_{nml}(x^t \mid x^{m-1})$. Furthermore as n goes to infinity the NML probability does not change, because it is equivalent to SNML and thus is horizon-independent. This means $\lim_{n\to\infty} \lim_{\Delta\theta\to 0} g^n(x^t, x^{m-1}, \Delta\theta) = p_{snml}(x^t \mid x^{m-1})$.

Next we show that the limit of $g^n(x^t, x^{m-1}, \Delta\theta)$ as n goes to infinity and $\Delta\theta$ goes to zero is a Bayesian conditional under Jeffreys prior, which completes the proof. The following is easy to see:

$$p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) = \int_{S^n_{x^t}(\theta_0)} p_{\theta_0}(y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right).$$

Moreover, we have

$$\begin{aligned} p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) &= p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0 \in C^0_{\Delta\theta/\sqrt{n^d}}\right) && (3.2)\\ &= p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in \sqrt{n}C^0_{\Delta\theta/\sqrt{n^d}}\right) && (3.3)\\ &= p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in C^0_{\Delta\theta}\right). && (3.4) \end{aligned}$$

Hence

$$\int_{S^n_{x^t}(\theta_0)} p_{\theta_0}(y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right) = p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in C^0_{\Delta\theta}\right).$$

Also,

$$g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t) p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in C^0_{\Delta\theta}\right)}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^{m-1}) p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^{m-1},Y^{n-m+1})} - \theta_0) \in C^0_{\Delta\theta}\right)}.$$

Let $F^n_{x^t,\theta_0}(.)$ be the cumulative distribution function of the random variable

$$\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0)$$

when the data is i.i.d. and generated by $p_{\theta_0}(\cdot)$. Define $F^n_{x^{m-1},\theta_0}(\cdot)$ similarly. With these definitions,

$$g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t) F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^{m-1}) F^n_{x^{m-1},\theta_0}\left(C^0_{\Delta\theta}\right)}.$$

Now we find the limit as $\Delta\theta$ goes to zero. There are two possibilities: either the MLE has a density with respect to Lebesgue measure or the model has a lattice MLE with diminishing step-size h_n. In the latter case, upon constructing $P^{\Theta}_{\Delta\theta/\sqrt{n^d}}$, we choose the hypercubes so that all points of the form $(b + k_1 h_n, b + k_2 h_n, \cdots, b + k_d h_n)$ in Θ are centers of some hypercubes. Furthermore we make sure that each of these hypercubes contains at most one point of the

form $(b + k_1h_n, b + k_2h_n, \cdots, b + k_dh_n)$, namely the center. Let $\Delta\theta_n$ be small enough to make this phenomenon hold. This construction makes many hypercubes $C^{\theta_0}_{\Delta\theta_n/\sqrt{n^d}}$ void of maximum likelihood points. Let us abbreviate $p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ in Equation (3.2) by $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$. Equation (3.2) shows that $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) = F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta_n}\right)$. Many of $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ are zero, namely those with θ_0 that do not correspond to a $\hat{\theta}_{(x^t,y^n-t)}$, hence:

$$\begin{aligned}\sum_{C^{\theta_0}_{\frac{\Delta\theta_n}{\sqrt{n^d}}}} p_{\theta_0}(x^t)F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta_n}\right) &= \sum_{C^{\theta_0}_{\frac{\Delta\theta_n}{\sqrt{n^d}}}} p_{\theta_0}(x^t)G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)\\ &= \sum_{\theta_0\in\hat{\Theta}^n_{x^t}} p_{\theta_0}(x^t)G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right),\end{aligned}$$

where $\hat{\Theta}^n_{x^t} = \left\{\theta \in \Theta \mid \exists\, y^{n-t} \ s.t.\ \hat{\theta}_{(x^t,y^n-t)} = \theta\right\}$. Furthermore we have the following.

$$g^n(x^t, x^{m-1}, \Delta\theta_n) = \frac{\sum_{\theta_0\in\hat{\Theta}^n_{x^t}} p_{\theta_0}(x^t)G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)}{\sum_{\theta_0\in\hat{\Theta}^n_{x^{m-1}}} p_{\theta_0}(x^{m-1})G^n_{x^{m-1},\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)}.$$

Note that $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ is the probability that $\hat{\theta}_{(x^t,Y^n-t)}$ equals θ_0 where Y^{n-t} are $n-t$ random variables generated by p_{θ_0} in an i.i.d fashion.

As n goes to infinity, the distribution of $\hat{\theta}_{(x^t,Y^n-t)}$ becomes independent of x^t. This is because $\frac{1}{n}\sum_{i=1}^t \log p_\theta(x_i)$ converges to zero for all θ, and $\hat{\theta}_{(x^t,Y^n-t)}$ converges in probability to θ_0. This along with the asymptotic normality of MLE implies that for all $\theta_0 \in \hat{\Theta}_{x^t,n}$, $G^n_{x^t,\theta_0}(\cdot)$ converges to the density of a multivariate normal distribution with mean θ_0 and covariance matrix $I^{-1}(\theta_0)$. A simple computation shows that the limit of $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ as n goes to infinity is $\sqrt{n^d\left|I(\theta_0)\right|/(2\pi)^d}$. Now we construct hypercubes of sides of length h_n and centers from $\hat{\Theta}^n_{x^t}$ for the numerator and from $\hat{\Theta}^n_{x^{m-1}}$ for the denominator. Let δ_n be the volume of each of these hypercubes. It is obvious that δ_n diminishes to zero as n goes to infinity. Using Riemann integral and the continuity of Fisher information and likelihood, we obtain:

$$\begin{aligned}\lim_{n\to\infty} g^n(x^t, x^{m-1}, \Delta\theta_n) &= \lim_{n\to\infty} \frac{\sum_{\theta_0\in\hat{\Theta}^n_{x^t}} p_{\theta_0}(x^t)G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)\delta_n}{\sum_{\theta_0\in\hat{\Theta}^n_{x^{m-1}}} p_{\theta_0}(x^{m-1})G^n_{x^{m-1},\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)\delta_n}\\ &= \frac{\int_\Theta p_\theta(x^t)\sqrt{|I(\theta)|}\,d\theta}{\int_\Theta p_\theta(x^{m-1})\sqrt{|I(\theta)|}\,d\theta},\end{aligned}$$

which shows that the strategy is Bayesian with Jeffreys prior. On the other hand if the MLE has a density with respect to Lebesgue measure then we get the following:

$$\begin{aligned}\lim_{\Delta\theta\to 0}\frac{1}{\sqrt{n^d}}\sum_{C^{\theta_0}_{\frac{\Delta\theta}{\sqrt{n^d}}}} p_{\theta_0}(x^t)F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right) &= \lim_{\Delta\theta\to 0}\frac{1}{\sqrt{n^d}}\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t)\left(\frac{F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)}{\Delta\theta/\sqrt{n^d}}\right)\frac{\Delta\theta}{\sqrt{n^d}}\\ &= \lim_{\Delta\theta\to 0}\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t)\left(\frac{F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)}{\Delta\theta}\right)\frac{\Delta\theta}{\sqrt{n^d}}\\ &= \int_\Theta p_{\theta_0}(x^t)f^n_{x^t,\theta_0}(0)d\,\theta_0,\end{aligned}$$

where $f^n_{x^t,\theta_0}(\cdot)$ is the density of $F^n_{x^t,\theta_0}$. This means that

$$g^n(x^t,x^{m-1}) \equiv \lim_{\Delta\theta\to 0} g^n(x^t,x^{m-1},\Delta\theta) = \frac{\int_\Theta p_{\theta_0}(x^t)f^n_{x^t,\theta_0}(0)d\,\theta_0}{\int_\Theta p_{\theta_0}(x^{m-1})f^n_{x^{m-1},\theta_0}(0)d\,\theta_0}. \tag{3.5}$$

As n goes to infinity, the distribution of $\hat{\theta}_{(x^t,Y^{n-t})}$ becomes independent of x^t. This is because $\frac{1}{n}\sum_{i=1}^t \log p_\theta(x_i)$ converges to zero for all θ, and $\hat{\theta}_{(x^t,Y^{n-t})}$ converges in probability to θ_0. This along with the asymptotic normality of MLE shows that as n goes to infinity we get the following convergence :

$$\sqrt{n}\left(\hat{\theta}_{(x^t,Y^{n-t})}-\theta_0\right)\xrightarrow{d} N\left(0,I^{-1}\left(\theta_0\right)\right).$$

Let $F_{\theta_0}(\cdot)$ be the cumulative distribution function of the multivariate normal distribution with mean 0 and covariance matrix $I^{-1}(\theta_0)$. Asymptotic normality implies that

$$F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)\to F_{\theta_0}(C^0_{\Delta\theta}).$$

This means that $f^n_{x^t,\theta_0}(\theta_0)$ converges to the density of a multivariate normal distribution with mean 0 and covariance matrix $I^{-1}(\theta_0)$. A simple computation shows that this value is $\sqrt{|I(\theta_0)|/(2\pi)^d}$. Now the only concern is whether we can take the limit of $n\to\infty$ inside the integral in Equation (3.5). We let $k^n_{x^t}(\theta)=\sqrt{(2\pi)^d}f^n_{x^t,\theta}(0)$, hence Equation (3.5) becomes:

$$g^n(x^t,x^{m-1}) = \frac{\int_\Theta p_\theta(x^t)k^n_{x^t}(\theta)d\theta}{\int_\Theta p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta}.$$

As $f^n_{x^t,\theta}(\theta)$ converges to $\sqrt{I(\theta_0)/(2\pi)^d}$ when n goes to infinity, $k^n_{x^{m-1}}(\theta)$ and $k^n_{x^t}(\theta)$ converge to $\sqrt{|\,I(\theta)\,|}$ as n goes to infinity. Now we use Lebesgue's dominated convergence theorem

[26] and Fatou's lemma [25] to show that limit and integral are interchangeable. Fatou's lemma shows that :

$$\int_\Theta p_\theta(x^{m-1})\sqrt{|\ I(\theta)\ |}d\theta \leq \lim_{n\to\infty}\int_\Theta p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta.$$

Let

$$h^n_{x^t}(\theta) = \frac{p_\theta(x^t)k^n_{x^t}(\theta)}{\lim_{s\to\infty}\int_\Theta p_\theta(x^{m-1})k^s_{x^{m-1}}(\theta)d\theta}.$$

As n goes to infinity, $k^n_{x^t}(\theta)$ approaches $\sqrt{|\ I(\theta)\ |}$. Hence for $\epsilon = \sqrt{|\ I(\theta)\ |}$ there exists an n_θ such that $|\ k^n_{x^t}(\theta) - \sqrt{|\ I(\theta)\ |}\ | \leq \epsilon$ for $n > n_\theta$.
Therefore for $n > n_\theta$ we have $k^n_{x^t}(\theta) \leq 2\sqrt{|\ I(\theta)\ |}$, and

$$h^n_{x^t}(\theta) \leq \frac{2p_\theta(x^t)\sqrt{|\ I(\theta)\ |}}{\int_\Theta p_\theta(x^{m-1})\sqrt{|\ I(\theta)\ |}d\theta}.$$

Now let $\bar{h}^n_{x^t}(\theta) = h^n_{x^t}(\theta)$ for $n > n_\theta$ and zero otherwise. For all n and $\theta \in \Theta$ we have :

$$\bar{h}^n_{x^t}(\theta) \leq \frac{2p_\theta(x^t)\sqrt{|\ I(\theta)\ |}}{\int_\Theta p_\theta(x^{m-1})\sqrt{|\ I(\theta)\ |}d\theta}.$$

It is obvious that the limits of both are equal as n goes to infinity. Furthermore, note that $\bar{h}^n_{x^t}(\theta)$ is upper bounded by an integrable function, namely twice the conditional Bayesian density of x^t under Jeffreys prior given x^{m-1}. We know that the conditional Bayesian density of x^t under Jeffreys prior given x^{m-1} is integrable from the assumption of the theorem. Consequently, Lebesgue's dominated convergence theorem is applicable here:

$$\begin{aligned}
\lim_{n\to\infty} g^n(x^t, x^{m-1}) &= \lim_{n\to\infty}\int_\Theta h^n_{x^t}(\theta)d\,\theta \\
&= \lim_{n\to\infty}\int_\Theta \bar{h}^n_{x^t}(\theta)d\,\theta \\
&= \int_\Theta \lim_{n\to\infty}\bar{h}^n_{x^t}(\theta)d\,\theta \\
&= \frac{\int_\Theta p_\theta(x^t)\sqrt{|\ I(\theta)\ |}}{\lim_{n\to\infty}\int_\Theta p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta}.
\end{aligned}$$

Also, we have

$$\begin{aligned}
\lim_{n\to\infty}\int_\Theta p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta &= \int_\Theta \lim_{n\to\infty} p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta \\
&= \int_\Theta p_\theta(x^{m-1})\sqrt{|\ I(\theta)\ |}d\theta,
\end{aligned}$$

because otherwise $p_{snml}(x^t \mid x^{m-1}) = \lim_{n\to\infty} g^n(x^t, x^{m-1}) = \lim_{n\to\infty} \int_\Theta \bar{h}^n_{x^t}(\theta) d\,\theta$ would not be a distribution. Hence we get:

$$\lim_{n\to\infty} \lim_{\Delta\theta\to 0} g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\int_\Theta p_\theta(x^t)\sqrt{I(\theta)}d\theta}{\int_\Theta p_\theta(x^{m-1})\sqrt{I(\theta)}d\theta}.$$

Notice that the proof does not use any properties of the Fisher information matrix. Thus, if the MLE is asymptotically normal with covariance $V(\theta)$, then an optimal Bayesian strategy has prior proportional to $\sqrt{|V(\theta)|}$. □

3.4 Examples

Example 3.4.1. *In this example the parametric family is the class of one-dimensional Gaussian distributions with unknown mean and variance μ and σ^2, i.e.*

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} + \log\sigma\right\}.$$

The MLE is

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i \quad \textit{and} \quad \hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}_n)^2.$$

The conditional SNML satisfies

$$\begin{aligned} p_{snml}(x_n|x^{n-1}) &\propto \left(2\pi\hat{\sigma}^2_n\right)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (x_i - \hat{\mu}_n)^2}{2\hat{\sigma}^2_n}\right\} \\ &= \frac{e^{-\frac{n}{2}} n^{\frac{n}{2}}}{(2\pi\,(n-1))^{\frac{n}{2}}} \frac{1}{\left(\hat{\sigma}^2_{n-1} + \frac{1}{n}\,(x_n - \hat{\mu}_{n-1})^2\right)^{\frac{n}{2}}}. \end{aligned}$$

Normalizing yields:

$$p_{snml}(x_n|x^{n-1}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \left(n\hat{\sigma}_{n-1}\right)^{-\frac{1}{2}} \left(1 + \frac{(x_n - \hat{\mu}_{n-1})^2}{n\hat{\sigma}^2_{n-1}}\right)^{-\frac{n}{2}}.$$

It can be shown [14] that for $n > 1$

$$\begin{aligned} &R(x_2^n, p_{snml} \mid x_1) - R(x_2^{n-1}, p_{snml} \mid x_1) \\ &= \frac{n+1}{2}\log n - \frac{n}{2}\log(n-1) - \frac{1}{2}\log 2e + \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \end{aligned}$$

This shows that the conditional SNML is an equalizer and hence equivalent to the conditional NML. Moreover, asymptotic normality holds for any $\mu \in \mathbb{R}$ and any $\sigma \in \mathbb{R}^+$ and $p_{\mu,\sigma^2}(x)$ is

continuous in μ and σ^2, hence Theorem 3.3.1 can be applied. This shows that conditional SNML and NML are equivalent to a conditional Bayesian strategy under Jeffreys prior. A direct computation of the Bayesian strategy with Jeffreys prior verifies this. Note that since this example is not a natural exponential family, the results of the previous chapter cannot be applied here.

Example 3.4.2. *In this example, the parametric family is the class of one-dimensional asymmetric student-t distributions as defined in [28] with unknown skewness parameter $\alpha \in (0\,,1)$ and fixed left and right tail parameters $v_1 = v_2 = 1$, i.e.*

$$p_\alpha(x) = \begin{cases} \frac{1}{\pi}\left(1 + \left(\frac{x}{2\alpha}\right)^2\right)^{-1} & \text{for } x \leq 0\,, \\ \frac{1}{\pi}\left(1 + \left(\frac{x}{2(1-\alpha)}\right)^2\right)^{-1} & \text{for } x > 0\,. \end{cases}$$

Zhu and John Galbraith [28] established asymptotic normality of maximum likelihood estimators in asymmetric student-t distributions. Note that additionally for any x, $p_\alpha(x)$ is continuous in α, hence Theorem 3.3.1 is applicable to this example. Proposition 2 in [28] shows that the Fisher information of p_α is proportional to $\frac{1}{\alpha(1-\alpha)}$. This means that Jeffreys prior is proportional to $\frac{1}{\sqrt{\alpha(1-\alpha)}}$. After normalization we get $\frac{1}{\pi\sqrt{\alpha(1-\alpha)}}$. Calculating the regret of the Bayesian strategy under Jeffreys prior shows that for a fixed $n > 0$, the regret changes for different sequences of observations. For example, for $n = 3$, and sequence of observations $(1, 1, -1)$ the maximum likelihood estimator of α is 0.4136 and the regret of the Bayesian strategy under Jeffreys prior is 1.1472. On the other hand if we observe $(2, 2, -2)$, the maximum likelihood estimator is 0.3777 with 1.1851 for regret. This means that the Bayesian strategy under Jeffreys prior is not optimal because otherwise it should have resulted in equal regrets for sequences of equal length. Furthermore Theorem 3.3.1 shows that no prior distribution on $(0\,,1)$ can make the Bayesian strategy optimal and SNML can not be optimal either.

Chapter 4

Characterization of Exponential Families with Minimax Optimal SNML and Bayesian with Jeffreys, and with Horizon-independent NML

In Chapter 2, we showed that a Bayesian prediction strategy with Jeffreys prior and sequential normalized maximum likelihood (SNML) coincide and are optimal if and only if the latter is exchangeable, which occurs if and only if the optimal strategy can be calculated without knowing the time horizon in advance. In this chapter we show that for 1-dimensional exponential families SNML is exchangeable only for three classes of exponential family distributions, namely the Gaussian, the gamma, and the Tweedie exponential family of order $3/2$, and any one-to-one transformation of them. The results of this chapter are from the paper [3].

4.1 Introduction

In this chapter our set of experts are i.i.d. exponential families of distributions, examples of which include normal, Bernoulli, multinomial, gamma, Poisson, Pareto, geometric distributions and many others. In online learning under logarithmic loss as we saw in Theorem 1.3.1, the minimax regret is achieved by the *normalized maximum likelihood* strategy. If the parameter space of a d-dimensional exponential family is constrained to a compact subset of the parameter space, NML achieves regret $(d/2)\ln n + O(1)$ [22, 19]. For unconstrained parameter spaces, the NML strategy is often not defined because it relies on sup-integrability and in many cases the model is not sup-integrable. In these cases NML can be replaced by the *conditional normalized maximum likelihood strategy*, which acts like NML, except that a small initial segment of the sequence is observed before prediction starts and then the NML

strategy is calculated conditioned on that initial segment. Whereas NML is optimal in the sense of achieving minimax regret (whenever it is finite), conditional NML is optimal in the sense that it achieves minimax *conditional* regret. Unfortunately both conditional NML and (whenever it is defined) the original NML suffer from two major drawbacks: the horizon n of the problem needs to be known in advance, and the strategy can be computationally expensive since it involves marginalizing over all possible future subsequences up to iteration n. These drawbacks motivated researchers to come up with an approximation to NML, known as *sequential normalized maximum likelihood*, or SNML for short. See [24, 20, 21] and Definition 7.

SNML predictions coincide with those of the NML distribution under the assumption that the current iteration is the last iteration. Therefore, SNML can be viewed as an approximation to NML for which the time horizon of the game does not need to be known. Kotlowski and Grünwald [14] showed that for general exponential families SNML is optimal up to an $O(1)$–term. Interestingly, acting short-sighted and looking only one step ahead does not hurt much.

A natural question to ask is if there are cases in which looking one step ahead in the prediction game is *exactly* the best one can do, even if the time horizon is known? In other words, when do SNML and NML coincide? This question is of fundamental importance for online learning for at least the following two reasons. First, we know that in a general sequential decision process, obtaining the optimal strategy requires recursive solution of the Bellman equation by a backward induction. A positive answer to the question above implies that we can avoid the backward induction altogether, because the optimal strategy in that case is independent of the time horizon: we get the same, optimal strategy no matter how far into the future we look. Thus, we only need to analyze the worst case regret with respect to the current outcome to be predicted. Second, as it was shown in the last two chapters, when NML and SNML coincide, they become Bayesian strategies and the prior of the Bayesian strategy must be Jeffreys prior. In other words, if NML is time-horizon independent, then the Bayesian strategy with Jeffreys prior is the minimax strategy. This happens if and only if the SNML strategy is *exchangeable* (see Theorems 3.3.1 and 2.2.4). Testing the exchangeability of the sequential strategy is, however, hard. What exponential families have exchangeable sequential normalized maximum likelihood strategies, and therefore have SNML=NML?

In this chapter we give a complete answer to this question, when the reference set of experts is a 1-dimensional natural exponential family. We show that there are essentially only three exponential families with a time-horizon independent minimax strategy (and hence both SNML and the Bayesian strategy with Jeffreys prior are equivalent to NML and thus optimal). These families are gamma, Gaussian and compound Poisson families (but also included are those families that can be obtained by a fixed transformation of the random variable from any of the three above, e.g. Pareto, Laplace, Rayleigh and many others). This implies that only in these three families is NML independent of the horizon, so that predicting by optimizing one–step ahead becomes equivalent to predicting by optimizing n–steps ahead, where n is the amount of data that the player is eventually going to observe.

The chapter is organized as follows. We introduce the mathematical context for our results in Section 4.2. We then give our main result in Section 4.3, showing that gamma, Gaussian and a compound Poisson family are the only families with time-horizon independent minimax strategies. Short versions of the proofs are given in Section 4.3. We end with a short discussion in Section 4.4.

4.2 Setup

We work with 1-dimensional *i.i.d. natural exponential families*. For these families $\mathcal{X}$ can be identified with a subset of $\mathbb{R}$ and the set of 'experts' is a set of distributions $\{p_\theta \mid \theta \in \Theta\}$ on $\mathbb{R}$, each of which is of the form

$$p_\theta(x) = h(x)\mathrm{e}^{\theta x - A(\theta)}, \ \theta \in \Theta. \tag{4.1}$$

Here h is a reference measure, given as a density relative to the underlying measure λ, and A is the cumulant generating function given by $A(\theta) = \ln \int e^{\theta x} \, \mathrm{d}h(x)$. The so-called natural parameter space of the family is the set

$$\Theta_{\text{full}} = \{\theta \in \mathbb{R} \,|\, A(\theta) < \infty\} \tag{4.2}$$

We will generally work with potentially restricted families with parameters sets Θ that may be proper subsets of Θ_{full} *and that we always require to have nonempty interior* (so for example, we do not consider finite subfamilies). Families with $\Theta = \Theta_{\text{full}}$ are called *full*.

According to the standard general definition of exponential families [2], we can have $\theta f(x)$ instead of θx in the exponent of 4.1 , for an arbitrary fixed function f. Families with $f(x) = x$ are called natural exponential families relative to random vector X (dened as $X(x) = x$). However, as long as f is smooth and one-to-one, a general exponential family with statistic $f(x)$ can always re-expressed as a natural exponential family relative to a different random variable $Y = f(X)$ (i.e. it defines exactly the same distributions on the underlying space), so our restriction to natural families is actually quite mild; see also the discussion right after our main result Theorem 4.3.9.

4.3 Main Results

We now provide a sequence of lemmas and theorems that lead up to our main result, Theorem 4.3.9. We provide a full proof of Lemma 4.3.1 and the final Theorem 4.3.9 in the main text, since, while not at all the most difficult ones, these results contain the key ideas for our reasoning. All other results are followed by a short proof sketch/idea.We first provide a number of definitions that will be used repeatedly.

4.3.1 Definitions

From now on, whenever we refer to an 'exponential family', unless we explicitly state otherwise, we mean an i.i.d. natural 1-dimensional family as in (4.1).

Our analysis below involves various parameterizations of natural exponential families, in particular the natural, the mean (see below) and the geodesic (see Section 4.3.3 below) parameterization. We typically use Θ for (a subset of) the natural parameter space, M for (a subset of) the corresponding mean-value space and B for the geodesic space, but if statements hold for general diffeomorphic parameterizations we use Γ to denote (subsets of) the parameter space (mean, geodesic and natural parameterizations are all instances of 'diffeomorphic' parameterizations [see 9, p. 611]). We then denote parameters by γ and we let $\hat{\gamma}(x^n)$ be the maximum likelihood (ML) estimate for data x^n. If x^n has no or several ML estimates, $\hat{\gamma}(x^n)$ is undefined. We let $\hat{\Gamma}_n$ be the subset of ML estimates for data of length n, i.e. the set of $\gamma \in \Gamma$ such that $\gamma = \hat{\gamma}(x^n)$ for some data x^n of length n, and we let $\hat{\Gamma}^\circ$ be the set of γ in the *interior* of Γ that are contained in $\hat{\Gamma}_n$ for *some* n. (recall that we always assume that Γ is closed). We will also use symbols $\hat{M}_n, \hat{M}^\circ, \hat{B}_n, \hat{B}^\circ, \ldots$ to denote corresponding sets in particular parameterizations. $D(\gamma_0 \| \gamma_1) := D(p_{\gamma_0} \| p_{\gamma_1})$ denotes the KL divergence of γ_1 to γ_0.

We recall the standard fact that every natural exponential family can be parameterized by the mean value of X: for each θ in the natural parameter space Θ, we can define $\mu_\theta := E_{p_\theta}[X]$; then the mapping from θ to μ_θ is one-to-one and strictly increasing, and the image $\mu(\Theta)$ is the mean-value parameter space M. We use $\hat{\mu}(x^n)$ for the maximum likelihood estimator in the mean-value parameter space. We will frequently use the *variance function* $V(\mu)$ which maps the mean of the family to its variance, i.e. $V(\mu)$ is the variance of p_μ. We note that the Fisher information $I(\mu)$ in the mean-value parameterization is the inverse of $V(\mu)$ [see 9, chap. 18].

Definition 15 (convex core)**.** *Consider a natural exponential family as in (4.1). Let* $x_0 = \inf\{x : x \in$ *support of* $h\}$*, and* $x_1 = \sup\{x : x \in$ *support of* $h\}$*. The* ***convex core*** *is the interval from* x_0 *to* x_1 *with* x_0 *included if and only if* h *has a point mass at* x_0*, and with* x_1 *included if and only if* h *has a point mass at* x_1*. We denote the convex core by* ***cc****.*

For example for a Bernoulli model, the convex core is $[0, 1]$, with 0 and 1 included. The intuition is that the convex core includes mean–values that can be achieved by distributions corresponding to natural parameter values ∞ and/or $-\infty$, in the cases where these are well–defined.

Definition 16 (maximal)**.** *An exponential family with* ***maximal mean-value parameter space*** *is an exponential family where the mean value parameter space equals the convex core* cc*.*

For example, truncated exponential families such as Bernoulli $[0.2, 0.8]$ do not satisfy the maximal mean-value condition. Note also that if we take the Bernoulli model in the natural

parameter space with the full parameter set Θ_{full} then we get mean-value parameter space $\mu(\Theta_{\text{full}}) = (0, 1)$, without the boundary points included. The maximal mean-value parameter space does include the Bernoulli boundary points. In the Gaussian location family with varying μ and fixed variance however, the maximal mean-value parameter space coincides with $\mu(\Theta_{\text{full}}) = \mathbb{R}$. Thus, the maximal mean-value parameter space coincides sometimes, but not always with $\mu(\Theta_{\text{full}})$ (the name 'full', although standard in the exponential family literature, is therefore perhaps a bit misleading).

4.3.2 Characterizations of SNML-Exchangeability

We now present three lemmas, which give an abstract characterization of SNML exchangeability. Then in Section 4.3.3 we will make these concrete, leading to our main theorem.

We let m be the smallest n such that for all $x^n \in \mathcal{X}^n$,

$$\int p_\gamma(x^n) I(\gamma)^{1/2} \, \mathrm{d}\gamma < \infty \qquad \text{and}$$
$$\int_{\mathcal{X}^{n-m}} \sup_{\gamma \in \Gamma} p_\gamma(x^m, y^{n-m}) \, \mathrm{d}\lambda^{n-m}(y^{n-m}) < \infty,$$

that is, such that Jeffreys posterior $\pi(\gamma \mid x^n) := p_\gamma(x^n) I(\gamma)^{1/2} / \int p_\gamma(x^n) I(\gamma)^{1/2} \, \mathrm{d}\gamma$ is proper (integrates to 1) for any conditioning sequence of length $n \geq m$, and that the conditional minimax regret is finite. Note that this implies that NML, Bayes with Jeffreys prior, and SNML, conditioned on any initial sequence of length m, are well-defined. From now on, each time we mention NML or SNML we mean NML or SNML conditioned on an initial sequence of suitable length m. In most of our examples $m = 1$ suffices.

We call the distribution p_γ *regular* if, for all x^n with $\hat{\gamma}(x^n) = \gamma$, we have $\mu_\gamma = \hat{\mu}(x^n) = E_{p_\gamma}[X] = n^{-1} \sum_{i=1}^n x_i$, i.e., in the mean-value parameter space, the ML estimator is equal to the observed average. This is always the case if the ML estimate is in the interior of Γ [9, see chap. 18], but if the ML estimate is on the boundary there can be exceptions, e.g. if Γ is a truncated parameter set. The following lemma is central:

Lemma 4.3.1. *Consider a natural exponential family (4.1) where the parameter set Γ is an interval.*

1. *If the SNML distribution for such a family is exchangeable then for all $n > m$ there is a constant C_n such that for all* regular *$\gamma_0 \in \hat{\Gamma}_n$, we have:*

$$\int_\Gamma \mathrm{e}^{-nD(\gamma_0 \| \gamma)} I(\gamma)^{1/2} \, \mathrm{d}\gamma = C_n. \tag{4.3}$$

2. *If furthermore the family has maximal mean-value parameter space, then the SNML distribution for such a family is exchangeable* if and only if *for all $n > m$ there is a*

constant C_n such that for all $\gamma_0 \in \hat{\Gamma}_n$

$$\int_\Gamma \mathrm{e}^{-nD(\gamma_0\|\gamma)} I(\gamma)^{1/2}\,\mathrm{d}\gamma = C_n. \tag{4.4}$$

The essence of the lemma is that C_n remains constant as γ_0 varies. This will be key to proving our main result.

Proof. As discussed in Theorem 3.3.1, if Γ is an interval, then SNML exchangeability is equivalent to the fact that Bayes with Jeffreys prior and NML coincide. Thus, equivalently, we must have, for *all* $x_1, \cdots, x_n \in \mathcal{X}^n$, and *all* t, such that $m \le t < n$,

$$p_\pi\left(x_{t+1}^n \left| x^t\right.\right) = p_{NML}^{(n)}\left(x_{t+1}^n \left| x^t\right.\right). \tag{4.5}$$

Since

$$p_\pi\left(x_{t+1}^n \left| x^t\right.\right) = \int_\Gamma p_\gamma\left(x_{t+1}^n\right)\,\mathrm{d}\pi\left(\theta \left| x^t\right.\right) = \int_\Gamma p_\gamma\left(x_{t+1}^n\right) \frac{p_\gamma(x^t) I(\gamma)^{1/2}}{\int_\Gamma p_{\gamma'}(x^t) I\left(\gamma'\right)^{1/2}\,\mathrm{d}\gamma'}\,\mathrm{d}\gamma,$$

and

$$p_{NML}^{(n)}\left(x_{t+1}^n \left| x^t\right.\right) = \frac{p_{\hat{\gamma}(x^n)}(x^n)}{\int_{\mathcal{X}^{n-t}} p_{\hat{\gamma}(x^t, y^{n-t})}(x^t, y^{n-t})\,\mathrm{d}\lambda^{n-t}(y^{n-t})}$$

in the diffeomorphic parametrization Γ, (4.5) is equivalent to

$$\int_\Gamma p_\gamma(x^n) I(\gamma)^{1/2}\,\mathrm{d}\gamma = C(n, x^t) \times p_{\hat{\gamma}(x^n)}(x^n), \tag{4.6}$$

where

$$C(n, x^t) = \frac{\int_\Gamma p_{\gamma'}(x^t) I\left(\gamma'\right)^{1/2}\,\mathrm{d}\gamma'}{\int_{\mathcal{X}^{n-t}} p_{\hat{\gamma}(x^t, y^{n-t})}(x^t, y^{n-t})\,\mathrm{d}\lambda^{n-t}(y^{n-t})}.$$

We now prove that $C(n, x^t) = C_n$, i.e. it may depend on n but *it does not depend on* $x_1, \ldots, x_n$. The key observation is that (4.6) is satisfied for any $t \ge m$, in particular for $t = m$, so that $C(n, x^t)$ cannot depend on x_{m+1}^n. However, since $C(n, x^t)$ and all other terms in (4.6) are invariant under any permutation of x^t, we conclude that $C(n, x^t)$ does not depend on the whole sequence x^n.

Now we divide both sides of (4.6) by $p_{\hat{\gamma}(x^n)}(x^n)$ and we exponentiate inside the integral. This gives:

$$\int_\gamma \mathrm{e}^{-\ln \frac{p_{\hat{\gamma}(x^n)}(x^n)}{p_\gamma(x^n)}} I(\gamma)^{1/2}\,\mathrm{d}\gamma = C_n. \tag{4.7}$$

We have thus shown that, assuming Γ is an interval, SNML exchangeability is equivalent to the condition that (4.7) holds for a fixed C_n, for all $x^n \in \mathcal{X}^n$.

Now for Part 1, let $\gamma_0 = \hat{\gamma}(x^n)$. We now use the celebrated robustness property of exponential families [9, Section 19.3, Eq. 19.12]. This property says that for all γ_0 such that p_{γ_0} is regular, for all x^n with $\hat{\gamma}(x^n) = \gamma_0$, we have

$$nD(\gamma_0 \| \gamma) = \ln \frac{p_{\hat{\gamma}(x^n)}(x^n)}{p_\gamma(x^n)}; \tag{4.8}$$

the result follows.

For Part 2, we note that, if the mean-value parameter space is maximal, then it must be an interval, and all points in this space must be regular [9, Section 19.3, Eq. 19.10]. The only-if direction follows immediately by Part 1. To see the converse, note that if the mean-value parameter space is maximal, then the maximum likelihood estimator exists and is unique for all $x^n \in \mathcal{X}^n$ (see [5]), and all $\gamma \in \Gamma$ are regular. Hence Equation (4.8) holds for all $x^n \in \mathcal{X}^n$ so that (4.4) implies that (4.7) holds for all $x^n \in \mathcal{X}^n$ and therefore that SNML exchangeability holds. □

We will also need a second lemma relating SNML exchangeability to maximality:

Lemma 4.3.2. *Consider a natural exponential family as in (4.1). If SNML is exchangeable, then the mean-value parameter space is maximal.*

Proof Sketch In our definition of exponential families we require that the parameter set Γ has nonempty interior, thus we may assume that it contains an interval. We can then show by approximating the integral in (4.3) by a Gaussian integral using standard Laplace-approximation techniques (as in e.g. [9, chap. 7]) that, for general 1-dimensional exponential families, the integral in (4.3) converges to $(2\pi/n)^{1/2}$ for any γ_0 in the interior of Γ. If SNML exchangeability holds, then we can show using Lemma 4.3.1 and continuity that this must also hold for all boundary points of Γ. But if the parameter space is not maximal, then the same standard Laplace approximation of the integral in (4.3) gives that for boundary points of Γ, the integral converges to $(1/2)(2\pi/n)^{1/2}$ and we have a contradiction.

Proof. Without loss of generality consider the mean-value parameter space. Assume the given exponential family is SNML-exchangeable and, without loss of generality, that the parameter space contains an interval $[\mu_0, \mu_1]$ with $\mu_0 < \mu_1$. By Lemma 4.3.1 we have for all n, all regular points in $x \in \hat{M}_n \cap [\mu_0, \mu_1]$ that

$$\int_{[\mu_0,\mu_1]} \exp\left(-nD\left(x \| \mu\right)\right) V\left(\mu\right)^{-1/2} \, \mathrm{d}\mu = C_n \tag{4.9}$$

is independent of x. Note that all points in the interior of $[\mu_0, \mu_1]$ must be regular [9, Section 19.3, Eq. 19.10].

By a standard Laplace approximation of the integral in (1) (done by a Taylor approximation of the KL divergence, $D\left(x \| \mu\right) \approx \frac{1}{2}\left(x - \mu\right)^2 V\left(x\right)^{-1}$, so that for large n the integral

becomes approximately Gaussian) we get, for each closed interval M_c that is a subset of the convex core, for each x in the interior of M_c, that

$$\frac{\int_{M_c} \exp\left(-nD\left(x\|\mu\right)\right) V\left(\mu\right)^{-1/2} \, \mathrm{d}\mu}{\left(\frac{2\pi}{n}\right)^{1/2}} \to 1 \tag{4.10}$$

and

$$\frac{\int_{\{\mu \in M_c : \mu \geq x\}} \exp\left(-nD\left(x\|\mu\right)\right) V\left(\mu\right)^{-1/2} \, \mathrm{d}\mu}{\left(\frac{2\pi}{n}\right)^{1/2}} \to \frac{1}{2} \tag{4.11}$$

For a precise statement and proof of these results, see e.g. [9, Theorem 8.1 combined with Eq. (8.14)]. Combining (4.10) with (4.9), taking $M_c = [\mu_0, \mu_1]$, it follows that $C_n \to (2\pi/n)^{1/2}$. Now for each $\epsilon > 0$ there is an n such that $x \in \hat{M}_n \cap M$ and $|x - \mu_0| < \epsilon$. Hence by continuity the equality (4.9) also holds for $x = \mu_0$, so we get

$$\frac{\int_{[\mu_0, \mu_1]} \exp\left(-nD\left(\mu_0\|\mu\right)\right) V\left(\mu\right)^{-1/2} \, \mathrm{d}\mu}{\left(\frac{2\pi}{n}\right)^{1/2}} \to 1 \tag{4.12}$$

Now assume (for the purpose of establishing a contradiction) that the convex core *cc* includes an $x' < \mu_0$ with $x' \notin M$ (M being the parameter space of the family), and let $M' = [x', \mu_1]$. Then μ_0 is in the interior of M' and so, taking $M_c = M'$, (4.11) with $x = \mu_0$ gives that the same integral as in (4.12) converges to $1/2$; we have arrived at a contradiction.

In the same way, one proves that there can be no $x' > \mu_1$ with x' in the convex core. Thus, the interval must coincide with the convex core, which is what we had to prove. □

4.3.3 The Main Theorem

In the following we will use the *Tweedie exponential family* of order $3/2$ [13]. These are natural exponential families characterized by a variance function of the form $V\left(\mu\right) = k\mu^{3/2}$, where μ is the mean and $V\left(\mu\right)$ is the variance function defined earlier (i.e. $V(\mu)$ is the variance of p_μ). Each of the elements is a compound Poisson distribution obtained by adding a Poisson distributed number of gamma distributions [13]. It is interesting to note that such distributions have a point mass at 0 so that the left tail gives a finite contribution to the Shtarkov integral but the right tail is light and gives an infinite contribution to the Shtarkov integral. Hence this family does not have finite minimax regret.

Lemma 4.3.3. *The following three types of exponential families are SNML exchangeable: The full Gaussian location families with fixed $\sigma^2 > 0$, the full gamma distributions with shape parameter $k > 0$, and the full Tweedie family of order $3/2$.*

Proof Sketch It is straightforward to check that all three families have maximal mean-value parameter space. The result now follows by checking that Condition (4.4) holds for these

families, which is relatively straightforward by taking derivatives of the cumulant generating function.

Proof. For each of the families it is sufficient to prove that

$$\int_{cc} \frac{\exp\left(-nD\left(\gamma_0 \| \gamma\right)\right)}{V\left(\gamma\right)^{1/2}} \, \mathrm{d}\gamma$$

does not depend on $\gamma_0 \in cc$ where cc denotes the convex core of the family.

In the Gaussian location family with variance σ^2 we have $D\left(\gamma_0 \| \gamma\right) = D\left(0 \| \gamma - \gamma_0\right)$, and $V\left(\gamma\right) = \sigma^2$, so the integral is invariant because of the invariance of the Lebesgue integral.

The scaling property of the gamma families implies that $D\left(\gamma_0 \| \gamma\right) = D\left(1 \left\| \frac{\gamma}{\gamma_0} \right.\right)$. For the gamma family with shape parameter k we have $V\left(\gamma\right) = \gamma^2/k$. Hence the integral equals

$$\begin{aligned} \int_0^\infty \frac{\exp\left(-nD\left(\gamma_0 \| \gamma\right)\right)}{\left(\gamma^2/k\right)^{1/2}} \, \mathrm{d}\gamma &= k^{1/2} \int_0^\infty \frac{\exp\left(-nD\left(1 \| \frac{\gamma}{\gamma_0}\right)\right)}{\gamma} \, \mathrm{d}\gamma \\ &= k^{1/2} \int_0^\infty \frac{\exp\left(-nD\left(1 \| t\right)\right)}{t} \, \mathrm{d}t, \end{aligned}$$

where we have used the substitution $t = \gamma/\gamma_0$. Hence the integral does not depend on γ_0.

We consider the Tweedie family of order $3/2$.Then the divergence can be calculated as

$$\begin{aligned} D\left(\mu_0 \| \mu_1\right) &= \int_{\mu_0}^{\mu_1} \frac{\mu - \mu_0}{2\mu^{3/2}} \, \mathrm{d}\mu \\ &= \left[\mu^{1/2} + \mu_0 \mu^{-1/2}\right]_{\mu_0}^{\mu_1} \\ &= \mu_1^{1/2} + \mu_0 \mu_1^{-1/2} - 2\mu_0^{1/2} \\ &= \frac{\left(\mu_1^{1/2} - \mu_0^{1/2}\right)^2}{\mu_1^{1/2}}. \end{aligned}$$

Therefore we have to prove that the following integral is constant

$$\begin{aligned} \int_0^\infty \exp\left(-nD\left(\gamma_0 \| \gamma\right)\right) V\left(\gamma\right)^{-1/2} \, \mathrm{d}\gamma &= \int_0^\infty \exp\left(-n \frac{\left(\gamma^{1/2} - \gamma_0^{1/2}\right)^2}{\gamma^{1/2}}\right) \gamma^{-3/4} \, \mathrm{d}\gamma \\ &= \int_0^\infty \exp\left(-\frac{\left(n\gamma^{1/2} - n\gamma_0^{1/2}\right)^2}{n\gamma^{1/2}}\right) \gamma^{-3/4} \, \mathrm{d}\gamma. \end{aligned}$$

The substitution $\gamma = t^4 n^{-2}$ gives

$$\frac{4}{n^{1/2}} \int_0^\infty \exp\left(-\frac{\left(t^2 - n\gamma_0^{1/2}\right)^2}{t^2} \right) \, \mathrm{d}t.$$

This integral is independent of γ_0, which proves the theorem. □

Remark 4.3.4. What we call "gamma" here includes also Pareto, Laplace, Rayleigh, Levy, Nakagami and many other families of distribution that are derived from the gamma family by a smooth one-to-one transformation. As the next lemma shows, smooth one-to-one transformations preserve SNML exchangeability.

Lemma 4.3.5. *Suppose* $\{p_\gamma(\,\cdot\,) \mid \gamma \in \Gamma\}$ *indexes an exponential family for a r.v.* X *that is SNML exchangeable. Let* $Y = f(X)$ *for some smooth one-to-one function* f *and let* $q_\gamma(\,\cdot\,)$ *be the density of* Y *under* $p_\gamma(\,\cdot\,)$*. Then the family* $\{q_\gamma(\,\cdot\,) \mid \gamma \in \Gamma\}$ *is SNML exchangeable as well.*

Proof. Since the family $p_\gamma(\,\cdot\,)$ is SNML exchangeable, hence for any $n > m$ the following joint distribution is invariant under permutations of x^n that leaves x^m invariant:

$$p_{snml}(x_{m+1}^n | x^m) = \prod_{t=m+1}^{n} \frac{\sup_\gamma p_\gamma(x^t)}{\int_{\mathcal{X}} \sup_\gamma p_\gamma(x^{t-1}, x) \, \mathrm{d}x} \tag{4.13}$$

Now under the $Y = f(X)$ transformation the density of Y becomes

$$q_\gamma(y) = p_\gamma(f^{-1}(y)) \left| \frac{\mathrm{d}\, f^{-1}(y)}{\mathrm{d}\, y} \right| . \tag{4.14}$$

For the ease of notation we let $v(y) = \left|\frac{\mathrm{d}\, f^{-1}(y)}{\mathrm{d}\, y}\right|$. Hence $q_\gamma(y) = p_\gamma\left(f^{-1}(y)\right) v(y)$ and

$$\begin{aligned}
p_{snml}(y_{m+1}^n|y^m) &= \prod_{t=m+1}^{n} \frac{\sup_\gamma q_\gamma(y^t)}{\int_{\mathcal{X}} \sup_\gamma q_\gamma\left(y^{t-1}, y\right)\, \mathrm{d}y} \\
&= \prod_{t=m+1}^{n} \frac{\sup_\gamma q_\gamma\left(f(x_1)\cdots f(x_t)\right)}{\int_{\mathcal{X}} \sup_\gamma q_\gamma\left(f(x_1)\cdots f(x_{t-1}), y\right)\, \mathrm{d}y} \\
&= \prod_{t=m+1}^{n} \frac{\sup_\gamma p_\gamma\left(x_1\cdots x_t\right) \prod_{j=1}^{t} v(y_j)}{\int_{\mathcal{X}} \sup_\gamma p_\gamma\left(x_1\cdots x_{t-1}, f^{-1}(y)\right) \prod_{j=1}^{t-1} v(y_j) v(y)\, \mathrm{d}y} \\
&= \prod_{t=m+1}^{n} \frac{\sup_\gamma p_\gamma(x^t) v(y_t)}{\int_{\mathcal{X}} \sup_\gamma p_\gamma(x^{t-1}, f^{-1}(y)) v(y)\, \mathrm{d}y} \\
&= \prod_{t=m+1}^{n} \frac{\sup_\gamma p_\theta(x^t) v(y_t)}{\int_{\mathcal{X}} \sup_\gamma p_\gamma(x^{t-1}, x)\, \mathrm{d}x} \\
&= p_{snml}(x_{m+1}^n|x^m) \prod_{t=m+1}^{n} v(y_t)\ .
\end{aligned}$$

Hence $p_{snml}(y_{m+1}^n|y^m)$ too is invariant under any permutation of y^n leaving y^m invariant, and hence exchangeable. Note that in the last but one equation we used the change of variable $f^{-1}(y) = x$ and the fact that $v(y)\mathrm{d}y = \mathrm{d}x$. □

As an example consider a random variable X with a gamma distribution of the form $Gamma\left(1/2, c/2\right)$. Now if X goes through the one-to-one transformation $f(X) = 1/X$ then $1/X \sim Inverse\text{-}Gamma\left(1/2, c/2\right)$ which is the same as $Levy\left(0, c\right)$, hence $Levy\left(0, c\right)$ is also SNML exchangeable. It is indeed easy to directly verify the SNML exchangeability of $Levy\left(0, c\right)$ using Lemma 4.3.1.

Now we are ready to state the next theorem which is simply a disjunction of two conditions necessary for SNML exchangeability in a parameterization called geodesic. The *geodesic parameterization* is the parameterization in which the Fisher information is constant. We will denote parameters in this parameterization by β with parameter set B. We can reparameterize from the natural parameter space Θ_{full} to the geodesic space by setting:

$$\beta = \int I(\theta)^{1/2}\, \mathrm{d}\theta, \tag{4.15}$$

so that $\mathrm{d}\beta = I(\theta)^{1/2}\, \mathrm{d}\theta$. Note that this is a bijection. This allows us to replace the integration measure in the condition of Lemma 4.3.1 and we get an equivalent condition: for any $n > m$ the following is independent of $\beta_0 \in \hat{B}^n$

$$\int_B \mathrm{e}^{-nD(\beta_0\|\beta)}\, \mathrm{d}\beta. \tag{4.16}$$

For an arbitrary parameterization, let

$$J^{(k)}(\gamma_0) = \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}\gamma^k} D(\gamma_0\|\gamma) \mid_{\gamma=\gamma_0}, \tag{4.17}$$

which is the coefficient of the k-th term in the Taylor series expansion of $D(\gamma_0\|\gamma)$.

Theorem 4.3.6. *Consider a natural exponential family (4.1). In terms of the geodesic parameterization, a necessary condition for SNML exchangeability is that there is a C s.t. for all $n \geq m$, and all $\beta_0 \in \hat{B}^\circ$ we have*

$$5(J^{(3)}(\beta_0))^2 - 4J^{(4)}(\beta_0)J^{(2)}(\beta_0) = C. \tag{4.18}$$

Proof Sketch A fifth-order Taylor expansion of (4.16) gives terms of different order in n, and each term should be constant. Equation (4.18) corresponds to the first non-trivial term in the expansion.

Proof. First of all, to obtain a better understanding of $J^{(k)}(\gamma_0)$, we list a few of its properties:

1. *For any parameterization*, it holds that $2J^{(2)}(\gamma_0)$ is equal to the Fisher information at γ_0 in the parameterization.
2. In the geodesic parameterization, $J^{(2)}(\beta_0)$ is constant over $\hat{B}^\circ$ and we will denote it as $J^{(2)}$.
3. In the *natural parameterization*, for $k \geq 2$,

$$J^{(k)}(\theta_0) = \frac{1}{k!}A^{(k)}(\theta_0) = \frac{1}{k!}I^{(k-2)}(\theta_0)\ , \tag{4.19}$$

   where $A^{(k)}$ is the kth derivative of the cumulant generating function, i.e. the kth cumulant, and $I^{(m)}$ is the m-th derivative of the Fisher information ($I^{(0)}$ is just the Fisher information).

A Taylor expansion of Equation (4.4) as a function of n gives that certain Taylor coefficients must equal zero and an elaborate calculation of the Taylor coefficient leads to Equation (4.18).

In the geodesic parameterization, the integral in Equation (4.4) becomes Equation (4.16). We denote this integral by $s(\beta_0, n)$. Using a fifth-order Taylor expansion we will show the following:

$$s(\beta_0, n) = \Phi + n^{-3/2} \cdot \frac{3}{4}\frac{\pi^{1/2}}{(J^{(2)})^{5/2}} \cdot u(\beta_0) + O(n^{-2}) \tag{4.20}$$

where

$$u(\beta_0) = \frac{5}{4} \cdot \frac{(J^{(3)}(\beta_0))^2}{J^{(2)}} - J^{(4)}(\beta_0), \tag{4.21}$$

$\Phi = \frac{\pi^{1/2}}{(nJ^{(2)})^{1/2}}$ is a Gaussian integral (scaled by n) and the n^{-2} remainder term may be both negative and positive. Condition (4.18) easily follows from Equation (4.20) as follows: take β_0, β_1 in $\hat{B}^\circ$. By Equation (4.16) we must have that $s(\beta_0, n) - s(\beta_1, n) = 0$ for all large n. But by Equation (4.20) this difference is equal to

$$cn^{-3/2} \cdot (u(\beta_0) - u(\beta_1)) + O(n^{-2})$$

for a constant $c > 0$ independent of β_0 and β_1. Since this must be 0 for all large n and since $u(\cdot)$ does not depend on n, this can only be true if $u(\beta_0) = u(\beta_1)$. We can do this for any β_0 and β_1 which makes Condition (4.18) follow.

Now we proceed to prove the claim in Equation (4.20). Define $A = [\beta_0 - c, \beta_0 + c]$ for some fixed $c > 0$, taken small enough so that A is a subset of the interior of B (this is why needed to restrict to $\hat{B}^\circ$ rather than $\hat{B}_n$). We can write

$$s(\beta_0, n) = f(\beta_0, n) + g(\beta_0, n) + h(\beta_0, n) \tag{4.22}$$

where we define:

$$f := \int_{\beta \in A} \mathrm{e}^{-nD(\beta_0 \| \beta)} \, \mathrm{d}\beta,$$

$$g := \int_{\beta > \beta_0 + c} \mathrm{e}^{-nD(\beta_0 \| \beta)} \, \mathrm{d}\beta \qquad h := \int_{\beta < \beta_0 - c} \mathrm{e}^{-nD(\beta_0 \| \beta)} \, \mathrm{d}\beta$$

(We write f instead of $f(\beta_0, n)$ whenever β_0 and n are clear from context; similarly for g, h).

We have

$$g \leq \sup_{\beta' > \beta_0 + c} \mathrm{e}^{-(n-m)D(\beta_0 \| \beta')} \int_{\beta > \beta_0 + c} \mathrm{e}^{-mD(\beta_0 \| \beta)} \, \mathrm{d}\beta \leq c_2 \mathrm{e}^{-c_3 n^{c_4}} \tag{4.23}$$

for some constants $c_2, c_3, c_4 > 0$. Here we used that $D(\beta_0 \| \beta')$ is increasing in β' so that the sup is achieved at $\beta_0 + c$, and the fact that by definition m was chosen such that the integral with $mD(\beta_0 \| \beta)$ in the exponent is finite. We can bound h similarly. Thus, the error we make if we neglect the integral outside the set A is negligible, and we can now concentrate on approximating f, the integral over A. We can write

$$f(\beta_0, n) = \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0 - \beta)^2} \left(\mathrm{e}^{-nJ^{(3)}(\beta_0)(\beta_0 - \beta)^3} \mathrm{e}^{-nJ^{(4)}(\beta_0)(\beta_0 - \beta)^4} \mathrm{e}^{-n \cdot O(\beta_0 - \beta)^5} \right) \, \mathrm{d}\beta \tag{4.24}$$

where the constant in front of the 5th-order term is bounded because we require A to be a compact subset of the interior of B. The fourth-order and fifth-order terms in the integral can themselves be well approximated by a first-order Taylor approximation of e^x and we can rewrite f as

$$\int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0 - \beta)^2} \left(\mathrm{e}^{-nJ^{(3)}(\beta_0)(\beta_0 - \beta)^3} (1 + V)(1 + W) \right) \, \mathrm{d}\beta$$

where $V = -nJ^{(4)}(\beta_0)(\beta_0-\beta)^4+O(n^2(\beta_0-\beta)^8)$ and $W = O(n(\beta_0-\beta)^5)$. Similarly, the second factor in the integral can be well-approximated by a second order Taylor approximation of $\mathrm{e}^x = 1 + x + (1/2)x^2 + O(x^3)$ so that we can further rewrite f as

$$\int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}(1+U)(1+V)(1+W)\,\mathrm{d}\beta = \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}(1+U+V+W+UV+UW+WV+UVW)\,\mathrm{d}\beta$$

where

$$U = -nJ^{(3)}(\beta_0)(\beta_0-\beta)^3 + \frac{1}{2}n^2(J^{(3)}(\beta_0))^2(\beta_0-\beta)^6 + O(n^3(\beta_0-\beta)^9).$$

Writing $\Phi_A := \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}\,\mathrm{d}\beta$ we can thus further rewrite f as

$$f = \Phi_A + \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}(U+V+R_1+R_2)\,\mathrm{d}\beta$$

where R_1 and R_2 are remainder terms,

$$\begin{aligned} R_1 = UV = {} & O(n^2|\beta_0-\beta|^7) + O(n^3(\beta_0-\beta)^{10}) \\ & + O(n^4|\beta_0-\beta|^{13}) + O(n^3|\beta_0-\beta|^{11}) + O(n^4(\beta_0-\beta)^{14}) \\ & + O(n^5|\beta_0-\beta|^{17}) \end{aligned}$$

and

$$R_2 = W(1+U+V+UV)) = O(n|\beta_0-\beta|^5).$$

Since $\int_{-\infty}^{\infty}|x|^m\mathrm{e}^{-nx^2}\,\mathrm{d}x = O(n^{(-m-1)/2})$, we have $\int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}(R_1+R_2)\,\mathrm{d}\beta = O(n^{-2})$, and hence we get

$$f = \Phi_A + \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}(U+V)\,\mathrm{d}\beta + O(n^{-2}).$$

Now, using the fact that $\int_{-a}^{a} x^3\mathrm{e}^{-nx^2}\,\mathrm{d}x = 0$ for all $a > 0$, the integral over the first term in U is 0. The final terms in U and V can be dealt with as the remainder terms above, and we can rewrite f further as

$$f = \Phi_A + \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}\left(\frac{1}{2}n^2(J^{(3)}(\beta_0))^2(\beta_0-\beta)^6 - nJ^{(4)}(\beta_0)(\beta_0-\beta)^4\right)\mathrm{d}\beta + O(n^{-2}).$$

If we integrate over the full real line rather than A then the error we make is of order $O(\mathrm{e}^{-cn}) \le O(n^{-2})$. The integrals over the real line can be evaluated whence we get:

$$\begin{aligned} f &= \Phi + \frac{n^2}{2}(J^{(3)}(\beta_0))^2 \cdot \left(\frac{15}{8}\frac{\pi^{1/2}}{(nJ^{(2)}(\beta_0))^{7/2}}\right) - nJ^{(4)}(\beta_0)\cdot\left(\frac{3}{4}\frac{\pi^{1/2}}{(nJ^{(2)}(\beta_0))^{5/2}}\right) + O(n^{-2}) \\ &= \Phi + n^{-3/2}\cdot\pi^{1/2}\cdot\left(\frac{15}{16}\cdot\frac{(J^{(3)}(\beta_0))^2}{(J^{(2)})^{7/2}} - \frac{3}{4}\cdot\frac{J^{(4)}(\beta_0)}{(J^{(2)})^{5/2}}\right) + O(n^{-2}). \end{aligned} \tag{4.25}$$

Combining with (4.22) and (4.23) Equation (4.20) follows. □

Note that originally the necessary and sufficient condition for SNML exchangeability was for all β_0 in $\hat{B}_n$, not just $\hat{B}^\circ$ as stated in Lemma 4.3.1. Here we slightly weakened it and only require $\beta_0 \in \hat{B}^\circ$. In this form the condition is not necessarily sufficient any more, but as we will see in the proof of Theorem 4.3.9, it is still sufficiently necessary for our purposes.

Theorem 4.3.7. *Consider a natural exponential family as in (4.1) with maximal mean-value parameter space. A necessary condition for SNML exchangeability is that the variance function is given by either*

$$V(\mu) = (c_1\mu + k)^2 \tag{4.26}$$

or

$$V(\mu) = (c_1\mu + k)^{3/2}, \tag{4.27}$$

for some constants c_1 and k.

Proof Sketch The differential equation (4.18) can be rephrased in terms of the mean value parameterization. Two solutions are (4.26) or (4.27). Other potential solutions are ruled out by a higher-order (in fact 7th-order) expansion.

Proof. The proof follows from Condition (4.18), which gives necessary conditions for exchangeability: there exists a constant C, such that for all $n \geq m$, all $\beta_0 \in \hat{B}^\circ$,

$$5\frac{(J^{(3)}(\beta_0))^2}{J^{(2)}} - 4J^{(4)}(\beta_0) = C.$$

To rephrase the above condition in terms of the natural parameterization, we use:

$$\frac{\partial}{\partial\beta}(\ldots) = \frac{\mathrm{d}\theta}{\mathrm{d}\beta}\frac{\partial}{\partial\theta}(\ldots) = I^{-1/2}(\theta)\frac{\partial}{\partial\theta}(\ldots),$$

because $\beta = \int I(\theta)^{1/2}\,\mathrm{d}\theta$, so that $\frac{\mathrm{d}\beta}{\mathrm{d}\theta} = I(\theta)^{1/2}$. We use the fact that $D(\beta_0\|\beta) = D(\theta_0\|\theta)$, where $\theta_0 = \theta(\beta_0)$ and $\theta = \theta(\beta)$ are corresponding parameters in the one-to-one mapping $\beta \mapsto \theta$. We also know that in the natural parameterization:

$$\frac{\partial^2}{\partial\theta^2}D(\theta_0\|\theta) = I(\theta).$$

We use the above properties to get:

$$\begin{aligned}
\frac{\partial D(\beta_0\|\beta)}{\partial\beta} &= \frac{\partial D(\theta_0\|\theta)}{\partial\theta}I^{-1/2}(\theta),\\
\frac{\partial^2 D(\beta_0\|\beta)}{\partial^2\beta} &= 1 - \frac{1}{2}\frac{\partial D(\theta_0\|\theta)}{\partial\theta}I^{-2}(\theta)\frac{\mathrm{d}I(\theta)}{\mathrm{d}\theta},\\
\frac{\partial^3 D(\beta_0\|\beta)}{\partial^3\beta} &= -\frac{1}{2}I^{-3/2}(\theta)\frac{\mathrm{d}I(\theta)}{\mathrm{d}\theta} + \frac{\partial D(\theta_0\|\theta)}{\partial\theta}\left(I^{-7/2}(\theta)\left(\frac{\mathrm{d}I(\theta)}{\mathrm{d}\theta}\right)^2 - \frac{1}{2}I^{-5/2}(\theta)\frac{\mathrm{d}^2I(\theta)}{\mathrm{d}\theta^2}\right),\\
\frac{\partial^4 D(\beta_0\|\beta)}{\partial\beta^4} &= \frac{7}{4}I^{-3}(\theta)\left(\frac{\mathrm{d}I(\theta)}{\mathrm{d}\theta}\right)^2 - I^{-2}(\theta)\frac{\mathrm{d}^2I(\theta)}{\mathrm{d}\theta^2} + \frac{\partial D(\theta_0\|\theta)}{\partial\theta}(\ldots).
\end{aligned}$$

From the above and (4.17) we get that:

$$\begin{aligned}
J^{(2)}(\beta_0) &= \frac{1}{2}, \\
J^{(3)}(\beta_0) &= \frac{1}{3!}\left(-\frac{1}{2}I^{-3/2}(\theta_0)\frac{\mathrm{d}I(\theta_0)}{\mathrm{d}\theta_0}\right), \\
J^{(4)}(\beta_0) &= \frac{1}{4!}\left(\frac{7}{4}I^{-3}(\theta_0)\left(\frac{\mathrm{d}I(\theta_0)}{\mathrm{d}\theta_0}\right)^2 - I^{-2}(\theta_0)\frac{\mathrm{d}^2I(\theta_0)}{\mathrm{d}\theta_0^2}\right),
\end{aligned}$$

where, as before, $\theta_0 = \theta(\beta_0)$, and we used the fact that $\frac{\partial D(\theta_0\|\theta)}{\partial\theta} = 0$ at $\theta = \theta_0$. Plugging the above into (4.18) and rearranging the terms gives the following differential equation for $I(\theta_0)$:

$$-\frac{4}{3}I^{-3}(\theta_0)\left(\frac{\mathrm{d}I(\theta_0)}{\mathrm{d}\theta_0}\right)^2 + I^{-2}(\theta_0)\frac{\mathrm{d}^2I(\theta_0)}{\mathrm{d}\theta_0^2} = \mathrm{const}(\theta_0) \tag{4.28}$$

for any $\theta_0 = \theta(\beta_0)$ for all $\beta_0 \in \hat{B}^\circ$. We now solve (4.28). Let us introduce a new variable $z(\theta_0) = I^{-1/3}(\theta_0)$. Then:

$$\begin{aligned}
\frac{\mathrm{d}z}{\mathrm{d}\theta_0} &= -\frac{1}{3}I^{-4/3}\frac{\mathrm{d}I}{\mathrm{d}\theta_0}, \\
\frac{\mathrm{d}^2z}{\mathrm{d}\theta_0^2} &= \frac{4}{9}I^{-7/3}\left(\frac{\mathrm{d}I}{\mathrm{d}\theta_0}\right)^2 - \frac{1}{3}I^{-4/3}\frac{\mathrm{d}^2I}{\mathrm{d}\theta_0^2}.
\end{aligned} \tag{4.29}$$

(we omit the dependence on θ_0 for the sake of clarity). The l.h.s. of Equation (4.28) becomes

$$-\frac{1}{3}I^{-2/3}\frac{\mathrm{d}^2z}{\mathrm{d}\theta_0^2} = -\frac{1}{3}z^2\frac{\mathrm{d}^2z}{\mathrm{d}\theta_0^2}.$$

Hence, the differential equation has simplified to:

$$\frac{\mathrm{d}^2z}{\mathrm{d}\theta_0^2} = \frac{c}{z^2}, \tag{4.30}$$

for some constant c.

Assume $c = 0$. Integrating Equation (4.30) once gives $\frac{\mathrm{d}z}{\mathrm{d}\theta_0} = a$. Using that $z = I^{-1/3}(\theta_0)$ and $I(\theta) = V(\mu(\theta))$ in the natural parametrization, we can rewrite the equation as

$$-\frac{1}{3}V^{-4/3}\frac{dV}{d\mu}\frac{d\mu}{d\theta_0} = a,$$

or, equivalently,

$$V(\mu)^{-1/3}\frac{\mathrm{d}V(\mu)}{\mathrm{d}\mu} = \mathrm{const}(\mu).$$

This differential equation has solutions of the form:

$$V(\mu) = (c_1\mu + k)^{3/2}$$

for some constants c_1 and k.

To find other SNML exchangeable families we may from now on **assume that** $c \neq 0$. We need to take a closer look at higher-order terms in the Taylor expansion of the integral (4.16) and obtain a stronger necessary condition for exchangeability.

Similarly as in the proof Theorem 4.3.6, we expand the integral over $A = [\beta_0 - c, \beta_0 + c]$:

$$\begin{aligned} f(\beta_0, n) &= \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2} \left(\prod_{k=3}^{6} e^{-nJ^{(k)}(\beta_0)(\beta_0-\beta)^k} \right) e^{-nO((\beta_0-\beta)^7)} \, \mathrm{d}\beta \\ &= \int_A \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2} \left(\prod_{k=3}^{7} (1 + X_k) \right) \, \mathrm{d}\beta \ , \end{aligned}$$

where

$$\begin{aligned} X_3 &= -nJ^{(3)}(\beta_0)(\beta_0-\beta)^3 + \frac{1}{2}n^2(J^{(3)}(\beta_0))^2(\beta_0-\beta)^6 + \frac{1}{3!}n^3(J^{(3)}(\beta_0))^3(\beta_0-\beta)^9 \\ &\quad + \frac{1}{4!}n^4(J^{(3)}(\beta_0))^4(\beta_0-\beta)^{12} + O(n^5(\beta_0-\beta)^{15}) \ , \\ X_4 &= -nJ^{(4)}(\beta_0)(\beta_0-\beta)^4 + \frac{1}{2}n^2(J^{(4)}(\beta_0))^2(\beta_0-\beta)^8 + O(n^3(\beta_0-\beta)^{12}) \, , \\ X_5 &= -nJ^{(5)}(\beta_0)(\beta_0-\beta)^5 + O(n^2(\beta_0-\beta)^{10}) \ , \\ X_6 &= -nJ^{(6)}(\beta_0)(\beta_0-\beta)^6 + O(n^2(\beta_0-\beta)^{12}) \ , \\ X_7 &= -O(n(\beta_0-\beta)^7) \ . \end{aligned}$$

We assume that Condition (4.18) is satisfied, so that $O(n^{-3/2})$ term in the expansion (cf. Equation (4.20)) is constant in β_0. Since if we integrate over the full real line rather than A then the error we make is of order $O(\mathrm{e}^{-cn})$, and $(\beta_0 - \beta)^m$ under Gaussian integral over the full real line results in $O(n^{(-m-1)/2})$ if m is even, and 0 if m is odd, there will be no terms of order $O(n^{-2})$. Therefore, we need to look for terms of order $O(n^{-5/2})$. There are five of them and their sum must be independent of β_0 (using similar argument as for the $O(n^{-3/2})$ term in the proof of Theorem 4.3.6):

$$\begin{aligned} &\frac{1}{4!}n^4(J^{(3)}(\beta_0))^4(\beta_0-\beta)^{12} \ + \ \frac{1}{2}n^2(J^{(4)}(\beta_0))^2(\beta_0-\beta)^8 \ - \ nJ^{(6)}(\beta_0)(\beta_0-\beta)^6 \\ &- \ \frac{1}{2}n^3(J^{(3)}(\beta_0))^2J^{(4)}(\beta_0)(\beta_0-\beta)^{10} \ + \ n^2J^{(3)}(\beta_0)J^{(5)}(\beta_0)(\beta_0-\beta)^8 = \mathrm{const}(\beta_0) \ . \end{aligned}$$

All the terms appear in the Gaussian integral. Given the fact that for even m,

$$\int \mathrm{e}^{-nJ^{(2)}(\beta_0)(\beta_0-\beta)^2}(\beta_0-\beta)^m d\beta = (m-1)!!\,(2\pi)^{1/2}\left(2nJ^{(2)}\right)^{-\frac{m+1}{2}} \ ,$$

we can rewrite the condition on the $O(n^{-5/2})$ term as:

$$\begin{aligned}\frac{11!!}{4!}n^{-5/2}(J^{(3)}(\beta_0))^4\left(2J^{(2)}\right)^{-13/2} &+ \frac{7!!}{2}n^{-5/2}(J^{(4)}(\beta_0))^2\left(2J^{(2)}\right)^{-9/2}\\ - 5!!n^{-5/2}J^{(6)}(\beta_0)\left(2J^{(2)}\right)^{-7/2} &- \frac{9!!}{2}n^{-5/2}(J^{(3)}(\beta_0))^2J^{(4)}(\beta_0)\left(2J^{(2)}\right)^{-11/2}\\ &+ 7!!n^{-5/2}J^{(3)}(\beta_0)J^{(5)}(\beta_0)\left(2J^{(2)}\right)^{-9/2} = \text{const}(\beta_0).\end{aligned}$$

Given that $J^{(k)}(\beta_0) = \frac{1}{k!}\frac{\partial^k D(\beta_0\|\beta)}{\partial\beta^k}\Big|_{\beta=\beta_0}$, and denoting $D_k = \frac{\partial^k D(\beta_0\|\beta)}{\partial\beta^k}\Big|_{\beta=\beta_0}$, we rewrite the condition above again as:

$$\frac{11!!}{4!(3!)^4}D_3^4 + \frac{7!!}{2(4!)^2}D_4^2 - \frac{5!!}{6!}D_6 - \frac{9!!}{2(3!)^2 4!}D_3^2D_4 + \frac{7!!}{3!5!}D_3D_5 = \text{const}(\beta_0),$$

where we also skipped the $n^{-5/2}$ terms and used the fact that $D_2 = 1$ in the geodesic parameterization. Evaluating the factorials and multiplying by a constant gives:

$$385D_3^4 + 105D_4^2 - 24D_6 - 630D_3^2D_4 + 168D_3D_5 = \text{const}(\beta_0). \tag{4.31}$$

Now, we need to evaluate D_3, D_4, D_5, and D_6. For the sake of argument, we will write them down in terms of the previously defined variable $z(\theta) = I^{-1/3}(\theta)$. We will also use a shorthand notation $z_n = \frac{\mathrm{d}^n z}{\mathrm{d}\theta^n}$:

$$\begin{aligned}\frac{\partial^2 D(\beta_0\|\beta)}{\partial\beta^2} &= 1 + \frac{3}{2}\frac{\partial D(\theta_0\|\theta)}{\partial\theta}z^2z_1\ ,\\ \frac{\partial^3 D(\beta_0\|\beta)}{\partial\beta^3} &= \frac{3}{2}z^{1/2}z_1 + \frac{\partial D(\theta_0\|\theta)}{\partial\theta}\left(3z^{5/2}z_1^2 + \frac{3}{2}z^{7/2}z_2\right)\ ,\\ \frac{\partial^4 D(\beta_0\|\beta)}{\partial\beta^4} &= \frac{15}{4}zz_1^2 + 3z^2z_2 + \frac{\partial D(\theta_0\|\theta)}{\partial\theta}\left(\frac{15}{2}z^3z_1^3 + \frac{45}{4}z^4z_1z_2 + \frac{3}{2}z^5z_3\right)\ ,\\ \frac{\partial^5 D(\beta_0\|\beta)}{\partial\beta^5} &= \frac{45}{4}z^{3/2}z_1^3 + \frac{99}{4}z^{5/2}z_1z_2 + \frac{9}{2}z^{7/2}z_3\ ,\\ &\quad + \frac{\partial D(\theta_0\|\theta)}{\partial\theta}\left(\frac{45}{2}z^{7/2}z_1^4 + \frac{135}{2}z^{9/2}z_1^2z_2 + \frac{45}{4}z^{11/2}z_2^2 + \frac{75}{4}z^{11/2}z_1z_3 + \frac{3}{2}z^{13/2}z_4\right)\ ,\\ \frac{\partial^6 D(\beta_0\|\beta)}{\partial\beta^6} &= \frac{315}{8}z^2z_1^4 + \frac{1305}{8}z^3z_1^2z_2 + 36z^4z_2^2 + \frac{237}{4}z^4z_1z_3 + 6z^5z_4 + \frac{\partial D(\theta_0\|\theta)}{\partial\theta}(\ldots)\ .\end{aligned}$$

Now, if we assume Condition (4.18) is satisfied, then so is the differential equation (4.30) at θ_0. It follows from (4.30) that:

$$\begin{aligned}z_2(\theta_0) &= cz(\theta_0)^{-2},\\ z_3(\theta_0) &= -2cz(\theta_0)^{-3}z_1(\theta_0),\\ z_4(\theta_0) &= 6cz(\theta_0)^{-4}z_1(\theta_0)^2 - 2c^2z(\theta_0)^{-5}\ .\end{aligned}$$

Using the above and the fact that $\frac{\partial D(\theta_0\|\theta)}{\partial\theta}\Big|_{\theta=\theta_0} = 0$, we finally write:

$$\begin{aligned}
D_2 &= 1, \\
D_3 &= \frac{3}{2} z^{1/2} z_1, \\
D_4 &= \frac{15}{4} z z_1^2 + 3c, \\
D_5 &= \frac{45}{4} z^{3/2} z_1^3 + \frac{63c}{4} z^{1/2} z_1, \\
D_6 &= \frac{315}{8} z^2 z_1^4 + \frac{645c}{8} z z_1^2 + 24c^2.
\end{aligned}$$

(all values of z and z_1 on the r.h.s. are evaluated at θ_0). We plug this into (4.31), denote $x := z^{1/2} z_1$, and get:

$$\begin{aligned}
385\frac{81}{16}x^4 + 105\left(\frac{225}{16}x^4 + \frac{90}{4}cx^2 + 9c^2\right) - 24\left(\frac{315}{8}x^4 + \frac{645c}{8}x^2 + 24c^2\right) \\
- 630\frac{9}{4}x^2\left(\frac{15}{4}x^2 + 3c\right) + 168\frac{3}{2}x\left(\frac{45}{4}x^3 + \frac{63c}{4}x\right) = \text{const}(\theta_0),
\end{aligned}$$

and after rearranging the terms:

$$\begin{aligned}
&x^4\left(\frac{385\cdot 81}{16} + \frac{105\cdot 225}{16} - \frac{24\cdot 315}{8} - \frac{630\cdot 15\cdot 9}{16} + \frac{168\cdot 45\cdot 3}{8}\right) \\
&+ cx^2\left(\frac{105\cdot 90}{4} - \frac{24\cdot 645}{8} - \frac{630\cdot 9\cdot 3}{4} + \frac{168\cdot 3\cdot 63}{8}\right) \\
&+ c^2\left(105\cdot 9 - 24\cdot 24\right) = \text{const}(\theta_0)
\end{aligned}$$

Interestingly the $O(x^4)$ term disappears and we get:

$$144cx^2 + 369c^2 = \text{const}(\theta_0)\ . \tag{4.32}$$

Since we have assumed that $c \neq 0$, Equation (4.32) is satisfied only when $x(\theta_0) = \text{const}(\theta_0)$. Since $x(\theta_0) = z(\theta_0)^{1/2}\frac{\mathrm{d}z(\theta_0)}{\mathrm{d}\theta_0}$, this leads to

$$\left(\frac{dz(\theta_0)}{d\theta_0}\right)^2 = \frac{\text{const}(\theta_0)}{z(\theta)}.$$

Using that $z(\theta_0) = I^{-1/3}(\theta_0)$ and $I(\theta) = V(\mu(\theta))$ in the natural parametrization, and Equation (4.29), we get the following differential equation in terms of the variance function:

$$\left(\frac{\mathrm{d}V(\mu)}{\mathrm{d}\mu}\right)^2 = -18cV(\mu),$$

which has a general solution of the form:

$$V(\mu) = (c_1\mu + k)^2$$

for some constants c_1 and k. □

Now we are ready to state our main theorem. We need one more definition: we say that a full exponential family of form (4.1) is a *linear transformation* of another full exponential family if, for some fixed a, b, it is the set of distributions given by (4.1), with each occurrence of x replaced by $ax + b$.

Remark 4.3.8. By Remark 4.3.4, linear transformations preserve SNML exchangeability. In a Gaussian location family, translating a distribution by b gives another distribution in the same exponential family, and the Gaussian location families are the only families with this property. Scaling a gamma distribution by a positive a gives another distribution in the same exponential family, and the gamma family is the only exponential family with this property.

Theorem 4.3.9. *The only natural 1-dimensional i.i.d. exponential families that have exchangeable SNML are the following three:*

- *The full Gaussian location families with arbitrary but fixed $\sigma^2 > 0$*
- *The full gamma exponential family with fixed shape parameter, and linear transformations of this family.*
- *The full Tweedie exponential family of order $3/2$, and linear transformations of this family.*

Before we prove this theorem, let us briefly discuss its generality. As we already indicated in the third paragraph after Equation (4.1), every exponential family defined with respect to a sufficient statistic $f(X)$ can be re-expressed as a natural family with respect to X as long as f is smooth and one-to-one. Thus the theorem also determines SNML exchangeability for general 1-dimensional i.i.d. exponential families with such f. Namely, if such a family, when mapped to a natural family, becomes the gamma, Gaussian or Tweedie $3/2$ family, then it is SNML exchangeable; otherwise it is not. The former is the case, for, for example, the Pareto and other families mentioned in Remark 4.3.4; the latter is the case, for, for example, the Bernoulli and Poisson distributions.

Proof. Lemma 4.3.3 says that these three families are SNML exchangeable. As we know that SNML exchangeability can only happen for families with maximal mean-value parameter space (Lemma 4.3.2), we focus on these families only. Thus, it is left to show that no other family with maximal mean-value parameter space is SNML exchangeable.

Theorem 4.3.7 gives the necessary condition for SNML exchangeability in terms of the variance function. Now we look at each case separately. The first part of the disjunction is the Equation (4.26), where the variance function is quadratic. Exponential families with quadratic variance functions have been classified by [15]. His result is that modulo linear transformations the only exponential families with quadratic variance functions are Gaussian, Poisson, gamma, binomial, negative binomial and the exotic hyperbolic secant distribution. Of these only the Gaussian and the gamma families have the desired form. We note that the exponential distributions are special cases of gamma distributions.

Now we get to the second case where the variance function is given by Equation (4.27). If $c_1 = 0$ we get an exponential family where the variance is constant, i.e. the family is the Gaussian translation family. Then the term k corresponds to a translation of the exponential family and we may assume that $k = 0$. If $c_1 \neq 0$ we can scale up or down and obtain the equation

$$V(\mu) = 2\mu^{3/2}. \tag{4.33}$$

There exists an exponential family with this variance function, namely the Tweedie family of order $3/2$ with $V(\mu) = 2\mu^{3/2}$. Since exponential families are uniquely determined by their variance function [15], the Tweedie family of order $3/2$ is the only family satisfying (4.33). □

4.4 Discussion

The present chapter has focused on 1-dimensional exponential families, whose parameter spaces have a nonempty interior. Any model that admits a 1-dimensional sufficient statistic can be embedded in a one dimensional exponential family. One can prove that SNML exchangeability implies that the parameter space must have non-empty interior, thus strengthening our results further.

We do not have any general results for the multidimensional case, but we can make a few observations: products of models that are SNML exchangeable are also exchangeable. All multidimensional Gaussian location models can be obtained in this way by a suitable choice of coordinate system. The only other SNML exchangeable models we know of in higher dimensions are Gaussian models where the mean is unknown and the scaling of the covariance matrix is unknown. This can be seen from the fact that a sum of squared Gaussian variables has a gamma distribution. The Tweedie family of order $3/2$ does not seem to play any interesting role in higher dimensions, because it cannot be combined with the other distributions.

Finally, we note that for all three exponential family distributions that are SNML exchangeable, the model is not sup-integrable, so NML is not defined. Thus, if we consider NML (rather than conditional NML) in 1-dimensional exponential families, we see that NML, when it is defined, is always horizon dependent. We conjecture that this conclusion will hold for arbitrary models. SNML exchangeability arises only in conditional models, and this is restricted to a few, very important models.

Chapter 5

Optimal Horizon-Dependent Priors, the NML Prior

In Chapter 2, we showed that if a Bayesian prediction strategy is optimal then it necessarily uses Jeffreys prior. As a result, if Jeffreys prior is not optimal then nor is any other prior, except possibly a horizon–dependent prior. This chapter investigates the behavior of a natural horizon–dependent prior called the NML prior. We show that the NML prior converges in distribution to Jeffreys prior, which makes it asymptotically optimal, but not necessarily optimal for an arbitrary horizon. Furthermore we show that there are exactly three families, namely Gaussian, gamma and inverse Gaussian, where the NML prior is equal to Jeffreys prior and hence horizon–independent. Two of these families, namely gamma and Gaussian, have optimal NML prior. Finally we show that Jeffreys prior is not always better than the NML prior, and that the NML prior is not always better than Jeffreys prior.

5.1 Introduction

Online learning under logarithmic loss aims to predict a sequence of outcomes $x_t \in \mathcal{X}$, almost as well as the best expert from a reference se, which in this chapter, are i.i.d natural exponential families. These families are parametrized by *the natural parameter* θ from a parameter space Θ. See Equation (5.1) or by parameter μ, called *the mean-value parameter*, from the mean-value parameter space Γ. See Equation (5.3).

We are interested in optimal Bayesian strategies. As it was shown in Chapter 2, if a Bayesian strategy is optimal then the strategy necessarily uses Jeffreys prior. This leaves open the possibility that a horizon-dependent prior (that is, one that depends on n, the number of rounds of the game) is optimal. For instance, in the case of Bernoulli experts, the Bayesian strategy with Jeffreys prior is not optimal. Xie and Barron [27] studied horizon-dependent priors. Although they did not find a prior of this kind that gives optimal predictions, they demonstrated a horizon-dependent prior that makes the Bayesian strategy

behave asymptotically like NML (even for data sequences with maximum likelihood estimators on the boundary). In this chapter we investigate a natural horizon-dependent prior that is motivated by NML. We call this prior the *NML prior* (see Definitions 17 and 18). For a fixed horizon n and the expert set $\{p_\mu(\cdot) \mid \mu \in \Gamma\}$, the NML prior at μ_0 is defined to be proportional to the density of the maximum likelihood estimator of n i.i.d samples from $p_{\mu_0}(\cdot)$ at μ_0. As n goes to infinity the NML prior converges in distribution to Jeffreys prior, and hence makes its corresponding Bayesian strategy asymptotically behave like NML (Lemma 5.3.1). This chapter investigate the following questions:

- The NML prior is asymptotically optimal, but is it optimal for any horizon n? We show not in all cases.
- Is the NML prior ever horizon independent, and if so how is it related to Jeffreys prior? We show that there are exactly three natural exponential families—namely gamma, inverse Gaussian and Gaussian (Theorem 5.3.2)—for which the NML prior is horizon independent, and in those cases it is equal to Jeffreys prior.
- Does the NML prior lead to lower regret than Jeffreys prior or vice versa? We show examples of both cases.
- Are there families for which the Bayesian strategy with the NML prior is optimal? The answer is positive. Two of the three families that have their NML priors equal to Jeffreys prior namely, the gamma and Gaussian families, have optimal NML priors.
- Finally we leave open the question whether there are other parametric families whose Bayesian strategy under the NML prior is optimal. Our two positive examples, namely the gamma and Gaussian families, have horizon–independent NML priors. Is it possible for an NML prior to be horizon–dependent and be optimal?

5.2 Notations and Definitions

In this chapter our set of experts are i.i.d natural exponential families defined in the following way:

$$p_\theta(x) = h(x)e^{\theta^\intercal x - A(\theta)}, \tag{5.1}$$

where h is a reference measure, and A is the cumulant generating function given by $A(\theta) = \log \int h(x)e^{\theta^\intercal x}\, dy$. The parameter space also known as the natural parameter space is the set

$$\Theta = \{\theta \in \mathbb{R}^d \mid A(\theta) < \infty\}. \tag{5.2}$$

An alternative way of representing the family is via the mean-value parameterization. The mean of $p_\theta(\cdot)$ is [9, chap. 19]

$$\mu = \nabla A(\theta),$$

hence

$$p_\mu(x) = h(x)e^{[(\nabla A)^{-1}(\mu)]^\intercal x - A\left([(\nabla A)^{-1}(\mu)]\right)}, \tag{5.3}$$

with the mean-value parameter space

$$\Gamma = \{\mu \in \mathbb{R}^d \mid \exists\, \theta \in \Theta \text{ s.t. } \nabla A(\theta) = \mu\}. \tag{5.4}$$

We need the following two definitions for our main results.

Definition 17 (Density of the Maximum Likelihood Estimator)**.** *Let $X_1, X_2, \cdots, X_n$ be a sequence of i.i.d random variables associated with density p_μ and probability distribution P_μ, and let $\hat{\mu}(X^n)$ be the random variable of this sequence's maximum likelihood estimator. Furthermore, let Q_μ^n be the probability distribution of $\hat{\mu}(X^n)$, meaning for any $\Gamma_0 \subseteq \Gamma$*

$$Q_\mu^n(\Gamma_0) = P_\mu^n\left(\hat{\mu}(X^n) \in \Gamma_0\right),$$

where P_μ^n is the sequence's probability distribution. We denote the density of the maximum likelihood estimator by f_μ^n which is equal to

$$f_\mu^{(n)} = \frac{dQ_\mu^n}{d\lambda_n},$$

where λ_n is assumed to be either the counting measure, in the case of discrete data, or the Lebesgue measure in the case of continuous data.

Note that in the mean-value parametrization of natural exponential families (Equation (5.3)), the maximum likelihood estimator of the mean parameter μ is the empirical mean of the observations. Therefore $f_\mu^{(n)}(\bar{\mu})$ is just the *nth* convolution of $p_\mu(\,\cdot\,)$ at $n\bar{\mu}$. This is because upon observing x^n, the θ that maximizes the joint distribution is the same θ that maximizes the following:

$$(x_1 + x_2 + \cdots + x_n)^\intercal\theta - nA(\theta)$$

Therefore the maximum likelihood estimator of θ is a $\hat{\theta}$ that solves the equation

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \nabla A(\hat{\theta}) = \hat{\mu}.$$

Note that we used the one-to-one transformation $\nabla A(\theta) = \mu$ to go from natural parametrization to mean–value parametrization.

We are ready to define the *NML prior*. Let the set of all possible maximum likelihood estimators for a fixed n be:

$$\hat{\Gamma}_n = \{\mu \in \Gamma \mid \exists\ x^n \in \mathcal{X}^n \text{ s.t } \hat{\mu}(x^n) = \mu\}.$$

Definition 18 (NML prior)**.** *For a fixed n, the NML prior is a horizon–dependent prior denoted by $\pi_n(\cdot)$ and defined over $\hat{\Gamma}_n = \{\mu \in \Gamma \mid \exists\, x^n \in \mathcal{X}^n \ s.t\ \hat{\mu}(x^n) = \mu\}$ as*

$$\pi_n(\bar{\mu}) = \frac{f^{(n)}_{\bar{\mu}}(\bar{\mu})}{\int_{\hat{\Gamma}_n} f^{(n)}_{\theta}(\mu)\, d\lambda_n(\mu)}.$$

It is convenient to work in a parametrization that is slightly different from that of Equations (5.1) and (5.3). Our main results rely on properties of a parametrization of natural exponential families called *the exponential dispersion models*. These models are parametrized by θ and σ^2 [13]. It will suffice to treat σ^2 (called the dispersion parameter) as fixed. Hence Jeffreys prior is a distribution over θ only (or, in the mean–value parameterization, over μ only),

$$p(x;\theta,\sigma^2) = c(x,\sigma^2)\exp\left(\frac{1}{\sigma^2}\left[\theta y - \kappa(\theta)\right]\right) \tag{5.5}$$

Equation (5.5) can be easily reparametrized into Equation (5.1) by the change of variable $x = \frac{1}{\sigma^2}y$ which gives $A(\theta) = \frac{1}{\sigma^2}\kappa(\theta)$ and $h(x) = c(\sigma^2 x, \sigma^2)\sigma^2$.

The natural parameter θ and the mean parameter μ are related to each other via

$$E_\theta[Y] = \mu = \frac{\partial \kappa(\theta)}{\partial \theta} = \tau(\theta). \tag{5.6}$$

So we can reparameterize Equation (5.5) as

$$p(x;\mu,\sigma^2) = c(x,\sigma^2)\exp\left(\frac{1}{\sigma^2}[\tau^{-1}(\mu)y - \kappa(\tau^{-1}(\mu))]\right). \tag{5.7}$$

5.3 Main Result

Lemma 5.3.1. *There exists a C, such that for all closed subsets Γ_0 in the interior of Γ,*

$$\lim_{n\to\infty} \frac{\frac{1}{\sqrt{n}}\int_{\Gamma_0\cap\hat{\Gamma}_n} f^{(n)}_{\mu}(\mu) d\lambda_n(\mu)}{\int_{\Gamma_0}\sqrt{I(\mu)}d\mu} = C,$$

where $I(\mu)$ is the Fisher information at μ.

Proof. We consider the *constrained stochastic complexity* over a closed subset of the mean–value parameter space, defined as

$$\log \int_{\{x^n\in\mathcal{X}^n \mid \hat{\mu}(x^n)\in\Gamma_0\}} \sup_{\mu\in\Theta} p_\mu(x^n)\, d\lambda^n(x^n).$$

Grünwald [9, pp. 227, 302, 303] showed that the constrained stochastic complexity over Γ_0 is

$$\log \int_{\Gamma_0 \cap \hat{\Gamma}_n} f_\mu^n(\mu) d\lambda_n(\mu) = \log \left(e^{o(1)} \left(\frac{n}{2\pi} \right)^{\frac{1}{2}} \int_{\Gamma_0} \sqrt{I(\mu)} d\mu \right),$$

hence

$$\lim_{n\to\infty} \frac{\frac{1}{\sqrt{n}} \int_{\Gamma_0 \cap \hat{\Gamma}_n} f_\mu^{(n)}(\mu) d\lambda_n(\mu)}{\int_{\Gamma_0} \sqrt{I(\mu)} d\mu} = \frac{1}{\sqrt{2\pi}}.$$

□

Note that when the NML prior converges in distribution to Jeffreys prior, their Bayesian posteriors also converge to each other for any sequence of data.

Theorem 5.3.2. *In 1-dimensional natural exponential families, the NML prior is exactly equal to Jeffreys prior for any number of observations, if and only if the family is gamma, Gaussian or inverse Gaussian.*

Proof. Fix $\sigma^2 > 0$. We want to show that $f_\mu^n(\mu)$ is proportional to $\sqrt{I(\mu)}$ for any $n > 0$ and $\mu \in \Omega$ if and only if the distribution is Gaussian, inverse Gaussian or gamma. First we restrict ourselves to $n = 1$ and look for distributions whose $f_\mu^1(\mu)$ (which is the same as $p_\mu(\mu)$) is proportional to Jeffreys prior at μ. Then we show that for those families Jeffreys prior is also proportional to $f_\mu^n(\mu)$ for any n.

Define $d(x, \mu) = 2y\{\tau^{-1}(x) - \tau^{-1}(\mu)\} - 2\{\kappa(\tau^{-1}(x)) - \kappa(\tau^{-1}(\mu))\}$. Then the density in the mean-value parametrization (Equation (5.7)) becomes

$$\begin{aligned} p(x; \mu, \sigma^2) &= c(x, \sigma^2) \exp\left(\frac{1}{\sigma^2} \{y\tau^{-1}(x) - \kappa(\tau^{-1}(x))\} \right) \exp\left(-\frac{1}{2\sigma^2} d(x, \mu) \right) \\ &= a(x, \sigma^2) \exp\left(-\frac{1}{2\sigma^2} d(x, \mu) \right), \end{aligned} \tag{5.8}$$

where we have defined $a(x, \sigma^2) = c(x, \sigma^2) \exp\left(\frac{1}{\sigma^2} \{y\tau^{-1}(x) - \kappa(\tau^{-1}(x))\} \right)$.

Define the *unit variance function* [13, p. 4] in the following way.

$$V(\mu) = \frac{2}{\frac{\partial^2}{\partial \mu^2} d(\mu, \mu)}.$$

It can be easily verified that the variance of Y is $\sigma^2 V(\mu)$, which gives a direct relationship between the mean and the variance. Jeffreys prior in the mean-value parameterization is proportional to the inverse of the square root of the variance, i.e., $\frac{1}{\sigma} V(\mu)^{-\frac{1}{2}}$. We are looking for models for which $p(\mu; \mu, \sigma^2)$ in Equation (5.8) is proportional to Jeffreys prior, that is,

$$a(\mu, \sigma^2) \exp\left(-\frac{1}{2\sigma^2} d(\mu, \mu) \right) \propto V(\mu)^{-\frac{1}{2}}.$$

Since $d(\mu, \mu)$ is zero, we are looking for models where

$$a(\mu, \sigma^2) \propto V(\mu)^{-\frac{1}{2}},$$

or, equivalently, models where

$$a(x, \sigma^2) = h(\sigma^2)V(x)^{-\frac{1}{2}}$$

for some function h of σ^2. In other words we are looking for exponential dispersion models of the form

$$p(x; \mu, \sigma^2) = h(\sigma^2)V(x)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}d(x, \mu)\right). \tag{5.9}$$

To complete our proof we use the *renormalized saddle point approximation*, which is an approximation of the exponential dispersion model in Equation (5.7), defined in the following way [13, p. 27].

$$p_0(x; \mu, \sigma^2) = c_0(\mu, \sigma^2)V(x)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}d(x, \mu)\right),$$

where the constant $c_0(\mu, \sigma^2)$ ensures that the approximation p_0 is a density:

$$c_0(\mu, \sigma^2) = \left(\int_\Omega V(x)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}d(x, \mu)\right) dy\right)^{-1}.$$

Daniels [6] showed that the only exponential dispersion models (as defined in Equation (5.8)) that are exact in their renormalized saddle-point approximation are Gaussian, inverse Gaussian, and gamma distributions. If the renormalized saddle point approximation is exact, then the moment generating functions of the two distributions must match [13, p. 189], meaning

$$\exp\left[\frac{1}{\sigma^2}\{\kappa\left(\theta + \sigma^2 u\right) - \kappa(\theta)\}\right] = \exp\left[\frac{1}{\sigma^2}\{\kappa\left(\theta + \sigma^2 u\right) - \kappa(\theta)\}\right] \frac{c_0(\tau(\theta), \sigma^2)}{c_0(\tau\left(\theta + \sigma^2 u\right), \sigma^2)}$$

for all u with $\theta + u\sigma^2 \in \Theta$. In this case, c_0 does not depend on its first argument. Thus, if the renormalized saddle-point approximation is exact, then the density is of the form of Equation (5.9). The converse statement is clearly true. Since the only exponential dispersion models for which this approximation is exact are the Gaussian, inverse Gaussian and gamma distributions. We have proved our theorem for $n = 1$.

Next, we show the theorem holds for any n. Note that these three families are all instances of Tweedie distributions with different indices. Tweedie distributions are exponential dispersion models whose variance is proportional to μ^p for some real p outside the interval $(0, 1)$; p is called the *index*. Clearly, Jeffreys prior in a Tweedie model of index p is proportional to $\mu^{-\frac{p}{2}}$. The inverse Gaussian has $p = 3$, the Gaussian has $p = 0$, and the gamma

distribution has $p = 2$. Let $X_1, \ldots, X_n$ be chosen i.i.d. from the Tweedie distribution with index p, mean μ and dispersion parameter σ^2 (written $X_i \sim Tw_p(\mu, \sigma^2)$). Then [13, p. 128]

$$X_1 + X_2 + \cdots + X_n \sim Tw_p\left(n\mu, n^{2-p}\sigma^2\right)$$

Let $f_{Tw}\left(\,\cdot\,; n\mu, p, n^{2-p}\sigma^2\right)$ denote the density of $Tw_p\left(n\mu, n^{2-p}\sigma^2\right)$ where p is either 0, 2, or 1. The NML prior for horizon n is proportional to:

$$\begin{aligned} f_{Tw}\left(\frac{X_1 + \cdots + X_n}{n} = \mu; n\mu, p, n^{2-p}\sigma^2\right) &= f_{Tw}\left(X_1 + \cdots + X_n = n\mu; n\mu, p, n^{2-p}\sigma^2\right) \\ &= h(n^{2-p}\sigma^2)\left((n\mu)^p\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2n^{2-p}\sigma^2} d(n\mu, n\mu)\right) \\ &= h(n^{2-p}\sigma^2)\left((n\mu)^p\right)^{-\frac{1}{2}} \end{aligned}$$

This is proportional to $\mu^{-\frac{p}{2}}$ and hence to Jeffreys prior, as desired. □

Two of the three families in Theorem 5.3.2, namely the gamma and Gaussian families, have optimal Bayesian strategies under Jeffreys prior [14], hence the NML prior is optimal for these families. These are the only families that have horizon–independent optimal NML priors. It is easy to show that the Jeffreys prior of inverse Gaussian is not optimal, therefore the NML prior is not optimal for this family. This means that even though the NML prior is asymptotically optimal, it is not necessarily optimal for arbitrary horizon. To show that inverse Gaussian does not have an optimal Jeffreys prior we compute the regret for two sequences of data under Jeffreys prior, and show that the two regrets differ. This implies that the Bayesian strategy under Jeffreys prior is not optimal, since optimal strategies are equalizers (see Theorem 1.3.1). We choose $x^3 = (100, 1, 1)$ and $x^3 = (100, 2, 2)$. Jeffreys prior for the inverse Gaussian is proportional to $\mu^{-\frac{3}{2}}$ which has a normalization factor of infinity. As a result, the Bayesian strategy is not defined. However conditioning on the first observation resolves the problem. An easy calculation shows that the regret of x^3 conditioned on $x_1 = 100$ is 14.606240 and the regret of x^3 conditioned on $X_1 = 100$ is 13.025950.

For families where the NML prior is different from Jeffreys prior (families other than those in Theorem 5.3.2), there are cases where Jeffreys prior is better than the NML prior in terms of maximum regret and vice versa. Therefore no one is universally better than the other. A Tweedie family of order $\frac{3}{2}$ has an optimal Jeffreys prior [3], however the NML prior in this case is different from Jeffreys prior, and hence is not optimal. On the other hand, our experiments shows that the NML prior for Bernoulli is better than Jeffreys prior. Jeffreys prior for Bernoulli is $\frac{1}{\pi}\beta\left(\frac{1}{2}, \frac{1}{2}\right)$, where β is the Beta function, and the maximum likelihood upon seeing m ones out of n outcomes is $\left(\frac{m}{n}\right)^m \left(\frac{n-m}{n}\right)^{n-m}$. Therefore the NML prior would be:

$$\pi_n\left(\frac{m}{n}\right) = \frac{p_\mu\left(\hat{\mu}\left(X^n\right) = \frac{m}{n}\right)}{\sum_{k=0}^n p_\mu\left(\hat{\mu}\left(X^n\right) = \frac{k}{n}\right)} = \frac{\binom{n}{m}\left(\frac{m}{n}\right)^m \left(\frac{n-m}{n}\right)^{n-m}}{\sum_{k=0}^n \binom{n}{k}\left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}}.$$

The Bayesian probability upon observing m ones under Jeffreys prior is $\frac{1}{\pi}\beta\left(m+\frac{1}{2}, n-m+\frac{1}{2}\right)$ whereas the Bayesian probability under the NML prior is:

$$\sum_{i=0}^{n}\left(\frac{i}{n}\right)^{m}\left(\frac{n-i}{n}\right)^{n-m} \times \quad \frac{\binom{n}{i}\left(\frac{i}{n}\right)^{i}\left(\frac{n-i}{n}\right)^{n-i}}{\sum_{k=0}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^{k}\left(\frac{n-k}{n}\right)^{n-k}}.$$

Our calculations show that the worst regret of Bernoulli's Bayesian strategy under the NML prior is always lower than the case where Jeffreys prior is used instead. The worst data sequence for the case of NML prior is a sequence that has only one one and for the case of Jeffreys prior it is a data sequence that has zero ones.

The two optimal NML priors that we found in this chapter, NML priors for gamma and Gaussian families, are horizon–independent. The question whether there exist horizon–dependent NML priors that are optimal still remains open.

Chapter 6

Attempts in Finding Optimal Horizon–dependent Priors for Online Binary Prediction

The Bayesian strategy under Jeffreys prior is not optimal for Bernoulli experts. In this chapter, we investigate possible approaches to find horizon–dependent priors that makes the Bayesian prediction for online binary prediction under Bernoulli experts minimax optimal, i.e. equivalent to NML. Even though this chapter does not find such optimal horizon–dependent priors, it shows different routes researchers can take to possibly tackle the problem.

6.1 Introduction

Online binary classification is the game between an adversary and a forecaster. At each round the forecaster should reveal her belief about the binary outcome of an event in a form of a probability distribution, simultaneously the adversary reveals the true value of the event. The sample space could be thought of as the set {0,1}. At each round, there is no assumption on how the event is generated. The binary outcome could even be generated in a way to mislead the learner, i.e. in an adversarial way. More formally we let $x_1, x_2, \ldots$, be a sequence of binary outcomes, i.e. $x_i \in \mathcal{X} = \{0, 1\}$ revealed one at a time. We use x^t to denote $(x_1, x_2, \cdots, x_t)$. At round t, after observing x^{t-1}, the learner assigns a probability distribution on $\mathcal{X}$, denoted $q_t\left(\cdot \mid x^{t-1}\right)$. Then, after x_t is revealed, the forecaster incurs the *log loss* -ln $q_t\left(x_t \mid x^{t-1}\right)$.

We call any sequential probability assignment a *strategy*. The term is also used for joint distributions when the horizon n is clear from context. The performance of a strategy is measured relative to the best in a reference set of i.i.d Bernoulli experts, each parametrized by a point θ in $[0, 1]$ and denoted as $p_\theta(\cdot)$. The joint probability of an arbitrary sequence

x^n under $p_\theta(\cdot)$ is :

$$\theta^{\sum_{t=1}^n x_t}(1-\theta)^{n-\sum_{t=1}^n x_t}.$$

Note that θ is the probability of observing a 1. The difference between the accumulated loss of the prediction strategy and the best expert in the reference set is called the *regret*; [4]; see Definition 3.

Let $q^{(n)}(\cdot)$ be such a distribution over n binary random variables. The regret as defined in Definition 3, of $q^{(n)}(\cdot)$ upon observing x^n is :

$$R^{[0,1]}(x^n, q^{(n)}) = \frac{\sup_{\theta\in[0,1]} \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{q^{(n)}(x^n)} = \frac{\left(\frac{\sum_{i=1}^n x_i}{n}\right)^{\sum_{i=1}^n x_i}\left(1-\frac{\sum_{i=1}^n x_i}{n}\right)^{n-\sum_{i=1}^n x_i}}{q^{(n)}(x^n)}$$

6.2 NML with i.i.d Bernoulli Distributions

Since there is no assumption on data, the interest lies in minimax strategies, strategies that minimize the regret in the worst case over all possible data sequences. As it was shown in Theorem 1.3.1, normalized maximum likelihood, Definition 6, is the unique joint distribution that achieves minimax regret. More formally, for an arbitrary sequence x^n, NML is the maximum likelihood of the sequence normalized, which is the following:

$$p^{(n)}_{nml}(x^n) = \frac{\left(\frac{\sum_{t=1}^n x_t}{n}\right)^{\sum_{t=1}^n x_t}\left(\frac{n-\sum_{t=1}^n x_t}{n}\right)^{n-\sum_{t=1}^n x_t}}{\sum_{i=0}^n \binom{i}{n}\left(\frac{i}{n}\right)^i\left(\frac{n-i}{n}\right)^{n-i}} \tag{6.1}$$

6.3 Bayesian Strategy with i.i.d Bernoulli Distributions

The drawback of NML is its marginalization cost. NML is naturally defined in terms of a joint distribution. In order to compute conditionals at round t, 2^{n-t} joint probabilities should be summed up for marginalization. This makes the game extremely costly for the forecaster. Bayesian strategies on the other hand are much easier to compute. A Bayesian strategy is defined by a prior $\pi(\cdot)$ on the parameter space $[0,1]$; as more data are observed the posterior is updated and the Bernoulli experts are mixed. At time t, the Bayesian conditional under prior $\pi(\cdot)$ is computed in the following way :

$$p_\pi(X_t = x_t|x^{t-1}) = \int_{[0,1]} \theta^{x_t}(1-\theta)^{1-x_t}\,\pi\left(\theta|x^{t-1}\right)\,d\theta$$

where $\pi\left(\theta|x^{t-1}\right)$ is the posterior. The joint distribution over x^n will be :

$$p_\pi(x^n) = \int_{[0,1]} \theta^{\sum_{t=1}^n x_t}(1-\theta)^{n-\sum_{t=1}^n x_t}\,\pi(\theta)\,d\theta$$

The Bayesian strategy with Jeffreys prior has very interesting properties and behaves asymptotically optimally. It has been shown that as long as the maximum likelihood does not lie on the boundary or gets arbitrarily close to it, the regret of the Bayesian strategy under Jeffreys prior converges to that of the minimax regret, i.e. to the NML's regret [see 9, chap. 8]. In the Bernoulli setup in case of observing all zeros or all ones, the regret of the Bayesian strategy under Jeffreys prior is higher than NML by an additive factor of $\frac{1}{2}\ln 2$. Bayesian updating under Jeffrey prior has the nice property that it corresponds to the following simple process. Initially we put two balls, one black and one white, in an urn. At each round the learner's probability of observing a 1 corresponds to the proportion of black balls in the urn and her probability of observing a 0 is the proportion of the white balls in the urn. At each round two black balls are added to the urn if a 1 is observed and two white balls is added to the urn if a 0 is observed. The proof that this process is indeed a Bayesian updating under Jeffreys prior is very simple. First note that the process is exchangeable, hence by de Finite's theorem it should be a Bayesian mixture of i.i.d Bernoulli distributions. The prior for this Bayesian mixture is determined by its moments. The tth moment is

$$\mu_t = \int_{[0,1]} \theta^t \pi(\theta) = p(x_1 = x_2 = \cdots = x_t = 1) = \prod_{i=1}^{t} \frac{2 \times i - 1}{2 \times i},$$

which is the probability of observing t ones in a row. These are the moments of a *Beta* distribution with parameters $\frac{1}{2}$ and $\frac{1}{2}$, i.e. $\beta\left(\frac{1}{2}, \frac{1}{2}\right)$ which is exactly the Jeffreys prior :

$$\pi_J(\theta) = \frac{1}{\pi\sqrt{\theta\,(1-\theta)}}$$

As we showed in Chapters 2 and 3 if Jeffreys prior is not optimal, neither can other priors be, except for horizon–dependent priors. In this chapter we investigates ways to find optimal horizon–dependent priors for Bernoulli distributions. In Section 6.4 we investigate ways to find a polynomial prior for this purpose and in Section 6.5 we look at this problem from the perspective of a finite moment problem.

6.4 Attempt I : Polynomial Priors

For a fixed horizon n we want to find a prior that takes a polynomial form of degree n and achieves the minimax bound.

$$\pi(\theta) = \sum_{i=0}^{n} a_i \theta^i \tag{6.2}$$

For a given sequence of outcomes x^n let m be the number of ones and $n - m$ be the number of zeros, and let θ be the probability of observing a one and $1 - \theta$ be the probability

of observing a zero. Using Equation (6.1), we write $n+1$ equations (for $m = 0, \cdots, m = n$) to find coefficients of our polynomial.

$$\int_0^1 \left(\sum_{i=0}^{n} a_i\theta^i\right)\theta^m(1-\theta)^{n-m}\,d\theta = \sum_{i=0}^{n}\left(\int_0^1 a_i\theta^i\right)\theta^m(1-\theta)^{n-m}\,d\theta =$$
$$\sum_{i=0}^{n} a_i\theta^{i+m}(1-\theta)^{n-m}\,d\theta = \sum_{i=0}^{n} a_i\beta(i+m-1, n-m+1) =$$
$$\sum_{i=0}^{n} a_i\frac{(m+i)!\,(n-m)!}{(n+i+1)!} = \frac{\left(\frac{m}{n}\right)^m\left(\frac{n-m}{n}\right)^{n-m}}{\sum_{i=0}^{n}\binom{i}{n}\left(\frac{i}{n}\right)^i\left(\frac{n-i}{n}\right)^{n-i}}.$$

Now we let A, B, C and D be $n+1$ by $n+1$ matrices defined in the following way:

$$A_{m,i} = \frac{(m+i)!\,(n-m)!}{(n+i+1)!}$$

$$D_{m,i} = (m+i)!$$

$$C_{m,i} = \begin{cases} \frac{1}{(n+i+1)!} & \text{if } m = i \\ 0 & \text{otherwise} \end{cases}$$

$$B_{m,i} = \begin{cases} (n-m)! & \text{if } m = i \\ 0 & \text{otherwise} \end{cases}$$

We let b and a be $n+1$ by 1 vectors defined in the following way:

$$a_{m,1} = a_m,$$

$$b_{m,1} = \frac{\left(\frac{m}{n}\right)^m\left(\frac{n-m}{n}\right)^{n-m}}{\sum_{i=0}^{n}\binom{i}{n}\left(\frac{i}{n}\right)^i\left(\frac{n-i}{n}\right)^{n-i}}.$$

It could be easily shown that $A = BDC$ and $a = A^{-1}b$, hence we have to find the inverse of A. $A^{-1} = C^{-1}D^{-1}B^{-1}$. C and B are diagonal matrices, therefore their inverses are as follows.

$$C^{-1}_{m,i} = \begin{cases} (n+i+1)! & \text{if } m = i \\ 0 & \text{otherwise} \end{cases}$$

$$B^{-1}_{m,i} = \begin{cases} \frac{1}{(n-m)!} & \text{if } m = i \\ 0 & \text{otherwise} \end{cases}$$

The only hard part is the inverse of D. The following lemma shows that D is invertible and gives an explicit formula for its inverse.

Lemma 6.4.1.

$$Det(D) = \prod_{i=0}^{n} i!^2$$

And

$$D^{-1}_{m,i} = \frac{(-1)^{m+i}}{m!^2 i!^2} \sum_{k=\max(m,i)}^{n} \frac{k!^2}{(k-m)!(k-i)!}.$$

Proof. A Hankel matrix H is any square matrix such that $H_{i,j} = H_{i-1,j+1}$. Hence an n by n Hankel matrix can be represented by $2n-1$ numbers called the Hankel series. These numbers are the first row, except for the last entry, and the transpose of the last column of the matrix concatenated. The binomial transform of a Hankel matrix H is another Hankel Matrix $\hat{H}$ whose Hankel numbers are defined in the following way: $\hat{h}_i = \sum_{k=0}^{i} \binom{i}{k} h_k$, where $\{h_0 \equiv H_{0,0}, h_1 \equiv H_{1,1}, h_2 \equiv H_{2,2}, \cdots, h_{2n-2} \equiv H_{2n-2,2n-2}\}$ is the Hankel series. The determinant of a Hankel matrix equals the determinant of the binomial transform of the matrix. D is obviously a Hankel matrix and its Hankel series is $\{i!\}$. Let V be a Hankel matrix whose Hankel series is the derangement series, i.e. $\{i! \sum_{k=0}^{i} \frac{(-1)^k}{k!}\}$. It could be easily shown that the binomial transform of V is D. Hence their determinants coincide. Radoux [17] showed that the determinant of V is $\prod_{i=0}^{n} i!^2$. This should be the determinant of D as well. Furthermore, Slavnov [23] showed that the inverse of D equals:

$$\frac{(-1)^{m+i}}{m!^2 i!^2} \sum_{k=\max(m,i)}^{n} \frac{k!^2}{(k-m)!(k-i)!}$$

□

The next theorem uses Lemma 6.4.1 and finds the coefficients in Equation (6.2).

Theorem 6.4.2. *The inverse of A is the following:*

$$\frac{(n+m+1)!}{(n-i)!} \frac{(-1)^{m+i}}{m!^2 i!^2} \sum_{k=\max(m,i)}^{n} \frac{k!^2}{(k-m)!(k-i)!}$$

And the coefficients of our polynomial prior in Equation (6.2) are:

$$a_m = \sum_{i=0}^{n} \left(\frac{(-1)^{m+i}}{m!^2 i!^2} \sum_{k=\max(m,i)}^{n} \frac{k!^2}{(k-m)!(k-i)!} \right) \frac{\left(\frac{i}{n}\right)^i \left(\frac{n-i}{n}\right)^{(n-i)}}{\sum_{j=0}^{n} \binom{n}{j} \left(\frac{j}{n}\right)^j \left(\frac{n-j}{n}\right)^{(n-j)} \frac{(n+i+1)!}{(n-m)!}}$$

Proof.

$$A^{-1} = C^{-1} D^{-1} B^{-1}$$

And

$$a = A^{-1} b = C^{-1} D^{-1} B^{-1} b$$

□

The only problem with the polynomial prior in Equation (6.2) is the lack of a guarantee for its positivity. It seems that for horizons up to 11 the prior is positive. However the polynomial slightly goes negative in the vicinity of 0 and 1 for horizons larger than 11 (see Figure 6.1).

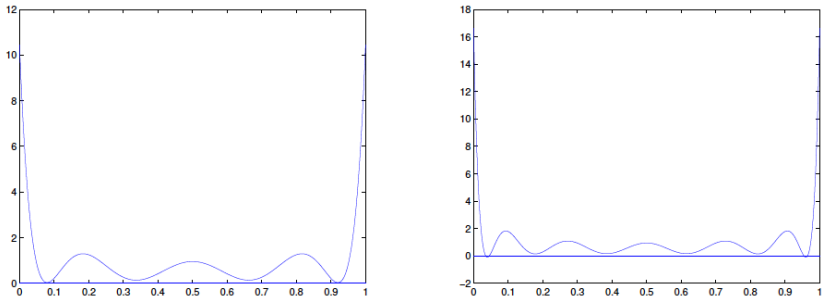

Figure 6.1: Polynomial prior for $n = 9$ (left), and for $n = 12$ (right).

6.5 Attempt II : Finite Hausdorff Moment Problem

In this section, we show that a well-known problem -called Finite Hausdorff moment problem– is a very natural approach to tackle the problem of finding optimal horizon–dependent priors. Note that for a fixed horizon n, if there exists a prior $\pi_n(\,\cdot\,)$ that makes Bayesian strategy equivalent to the NML strategy, then the first $n+1$ moments of this prior would be:

$$\mu_0 = 1, \mu_1 = p^{(n)}_{nml}(1), \mu_2 = p^{(n)}_{nml}(1^2), \cdots, \mu_n = p^{(n)}_{nml}(1^n)$$

This is because for any $i \leq n$, and $x^i = 1^i$,

$$p^{(n)}_{nml}(x^i) = \int_{[0,1]} \theta^{\sum_{k=1}^{i} x_k} (1-\theta)^{i-\sum_{k=1}^{i} x_k} \pi_n(\theta)\, d\theta = \int_{[0,1]} \theta^i \pi_n(\theta)\, d\theta = \mu_i = p^{(n)}_{nml}(1^i).$$

Thus, the problem of finding an optimal horizon–dependent prior boils down to finding a distribution $\pi_n(\,\cdot\,)$, such that its first $n+1$ moments coincide with $1, p^{(n)}_{nml}(1), p^{(n)}_{nml}(1^2), \cdots, p^{(n)}_{nml}(1^n)$. This is a well-known problem called *the finite Hausdorff moment problem*: What properties should a sequence of $n+1$ numbers $\mu_0, \mu_1, \cdots, \mu_n$ have to make them the first $n+1$ moments of some distribution? In other words does there exist a distribution with first $n+1$ moments $\mu_0, \mu_1, \cdots, \mu_n$? Such a distribution exists if a series of Hankel matrices defined below are positive definite [8].

If n is even we let

$$\Delta l_0 = \mu_0,$$

$$\Delta l_2 = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix},$$

$$\Delta l_4 = \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix},$$

$$\cdots,$$

$$\Delta l_n = \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_{\frac{n}{2}} \\ \mu_1 & \mu_2 & \cdots & \mu_{\frac{n}{2}+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{\frac{n}{2}} & \mu_{\frac{n}{2}+1} & \cdots & \mu_n \end{pmatrix}.$$

and let

$$\Delta u_0 = \mu_1 - \mu_2,$$

$$\Delta u_2 = \begin{pmatrix} \mu_1 - \mu_2 & \mu_2 - \mu_3 \\ \mu_2 - \mu_3 & \mu_3 - \mu_4 \end{pmatrix},$$

$$\Delta u_4 = \begin{pmatrix} \mu_1 - \mu_2 & \mu_2 - \mu_3 & \mu_3 - \mu_4 \\ \mu_2 - \mu_3 & \mu_3 - \mu_4 & \mu_4 - \mu_5 \\ \mu_3 - \mu_4 & \mu_4 - \mu_5 & \mu_5 - \mu_6 \end{pmatrix},$$

$$\cdots,$$

$$\Delta u_n = \begin{pmatrix} \mu_1 - \mu_2 & \mu_2 - \mu_3 & \cdots & \mu_{\frac{n}{2}} - \mu_{\frac{n}{2}+1} \\ \mu_2 - \mu_3 & \mu_3 - \mu_4 & \cdots & \mu_{\frac{n}{2}+1} - \mu_{\frac{n}{2}+2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{\frac{n}{2}} - \mu_{\frac{n}{2}+1} & \mu_{\frac{n}{2}+1} - \mu_{\frac{n}{2}+2} & \cdots & \mu_{n-1} - \mu_n \end{pmatrix}.$$

And if n is odd we let

$$\Delta l_1 = \mu_1,$$

$$\Delta l_3 = \begin{pmatrix} \mu_1 & \mu_2 \\ \mu_2 & \mu_3 \end{pmatrix},$$

$$\Delta l_5 = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \\ \mu_3 & \mu_4 & \mu_5 \end{pmatrix},$$

$$\cdots,$$

$$\Delta l_n = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_{\frac{n}{2}+1} \\ \mu_2 & \mu_3 & \cdots & \mu_{\frac{n}{2}+2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{\frac{n}{2}+1} & \mu_{\frac{n}{2}+2} & \cdots & \mu_n \end{pmatrix},$$

and let

$$\Delta u_1 = \mu_0 - \mu_1,$$

$$\Delta u_3 = \begin{pmatrix} \mu_0 - \mu_1 & \mu_1 - \mu_2 \\ \mu_1 - \mu_2 & \mu_2 - \mu_3 \end{pmatrix},$$

$$\Delta u_5 = \begin{pmatrix} \mu_0 - \mu_1 & \mu_1 - \mu_2 & \mu_2 - \mu_3 \\ \mu_1 - \mu_2 & \mu_2 - \mu_3 & \mu_3 - \mu_4 \\ \mu_2 - \mu_3 & \mu_3 - \mu_4 & \mu_4 - \mu_5 \end{pmatrix},$$

$$\cdots,$$

$$\Delta u_n = \begin{pmatrix} \mu_0 - \mu_1 & \mu_1 - \mu_2 & \cdots & \mu_{\frac{n}{2}} - \mu_{\frac{n}{2}+1} \\ \mu_1 - \mu_2 & \mu_2 - \mu_3 & \cdots & \mu_{\frac{n}{2}+1} - \mu_{\frac{n}{2}+2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{\frac{n}{2}} - \mu_{\frac{n}{2}+1} & \mu_{\frac{n}{2}+1} - \mu_{\frac{n}{2}+2} & \cdots & \mu_{n-1} - \mu_n \end{pmatrix}.$$

If $|\Delta l_i| > 0$ and $|\Delta u_i| > 0$ for all i's defined above, then there exists some distribution with its first $n+1$ moments equal to $\mu_0, \mu_1, \cdots, \mu_n$.

If such a distribution exists there are recipes for reconstruction, given the moments. One of these methods is the maximum entropy method. For further details refer to [8].
For different values of n, we constructed the Δl and Δu matrices on the NML moments :

$$1\,,\, p^{(n)}_{nml}(1)\,,\, p^{(n)}_{nml}(1^2)\,,\, p^{(n)}_{nml}(1^3)\,,\, \cdots\,,\, p^{(n)}_{nml}(1^n)$$

and all of them turned out positive definite. The question whether this is true for all n and how to reconstruct such a prior in an efficient manner remains still open.

Chapter 7

Conclusion

7.1 Overview

We have shown that NML, the unique minimax optimal strategy for online learning under logarithmic loss, is equivalent to Bayesian updating under Jeffreys prior and the SNML strategy if and only if the latter is exchangeable. This result holds for exponential families and more generally for any parametric family for which the maximum likelihood estimator is asymptotically normal. Moreover we showed if there is any optimal prior it must be Jeffreys prior, and optimality of this prior implies optimality of SNML and vice versa. In 1-dimensional exponential families this phenomenon holds only for three families and any 1-to-1 transformations of any of them. These families are: Gaussian, gamma, and the Tweedie families of order $3/2$. We further showed that for two of these families, i.e. the Gaussian and gamma families, the NML prior is equal to Jeffreys prior. The only other family that has this property is inverse Gaussian.

7.2 Open Problems

One general question which still remains open is the relationship between SNML and Bayesian updating with Jeffreys prior. Our results show that if either strategy is optimal so is the other one. This raises the performance comparison question: How does the regret of SNML compare to that of the Bayesian strategy under Jeffreys prior in non-optimal scenarios?

Optimal Bayesian strategies with horizon–dependent priors is another interesting direction that our research can take. The possibility of priors other than Jeffreys being optimal is non-existent. The only possible priors that can be optimal are those that are horizon–dependent, priors that depend on the number of outcomes that the forecaster will eventually see. For instance in online 0–1 prediction with Bernoulli experts, Jeffreys prior and consequently SNML are not optimal. Does there exist a horizon–dependent prior that makes the corresponding Bayesian strategy optimal? Does such a prior exists for all families with

non-optimal Bayesian strategies under Jeffreys prior?

Finally, for what multi-dimensional exponential families is SNML exchangeable?

Bibliography

[1] K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43:211–246, June 2001. ISSN 0885-6125. doi: 10.1023/A:1010896012157. URL http://portal.acm.org/citation.cfm?id=599611.599643.

[2] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. New York : John Wiley, 1978.

[3] P. Bartlett, P. Grünwald, P. Harremoës, F. Hedayati, and W. Kotlowski. Horizon-independent optimal prediction with log-loss in exponential families. *Proceedings of the Twenty Fifth Conference on Learning Theory (COLT' 13)*, 30:639–661, 2013.

[4] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[5] I. Csiszar and F. Matu. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003. ISSN 0018-9448. doi: 10.1109/TIT.2003.810633.

[6] H. Daniels. Saddlepoint Approximations in Statistics. *The Annals of Mathematical Statistics*, 25(4):631–650, December 1954. ISSN 00034851. doi: 10.2307/2236650. URL http://dx.doi.org/10.2307/2236650.

[7] P. Diaconis and D. A. Freedman. Cauchy's equation and De Finetti's theorem. *Scandinavian Journal of Statistics*, 17(3):pp. 235–249, 1990. ISSN 03036898. URL http://www.jstor.org/stable/4616171.

[8] M. Frontini and A. Tagliani. Maximum entropy in the finite Hausdorff moment problem.

[9] P. Grünwald. *The Minimum Description Length Principle*. MIT Press Books. The MIT Press, 2007. URL http://ideas.repec.org/b/mtp/titles/0262072815.html.

[10] F. Hedayati and P. Bartlett. The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. *Proceedings of the Twenty Fifth Conference on Learning Theory (COLT' 12)*, 2012.

[11] F. Hedayati and P. Bartlett. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior. *Journal of Machine Learning Research - Proceedings Track*, 22:504–510, 2012.

[12] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[13] B. Jørgensen. *The Theory of Dispersion Models*. London, 1997.

[14] W. Kotlowski and P. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In Sham Kakade and Ulrike von Luxburg, editors, *Proceedings of Annual Conference on Learning Theory 2011*. JMLR.org, July 2011. URL http://colt2011.sztaki.hu/colt2011_submission_52.pdf.

[15] C. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80, 1982. doi: 10.2307/2240499. URL http://dx.doi.org/10.2307/2240499.

[16] W. Newey and D. McFadden. Chapter 35: Large sample estimation and hypothesis testing. In Robert Engle and Dan. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier Science, 1994. ISBN 0-444-88766-0.

[17] C. Radoux. Déterminant de Hankel construit sur des polynômes liés aux nombres de dérangements. *European J. Combin., 12*, 43(2):327–329, 1991.

[18] J. Rissanen. Modeling By Shortest Data Description. *Automatica*, 14:465–471, 1978.

[19] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=481776.

[20] J. Rissanen and T. Roos. Conditional NML universal models. In *Information Theory and Applications Workshop, 2007*, pages 337–341, 2007. doi: 10.1109/ITA.2007.4357600.

[21] T. Roos and J. Rissanen. On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, 2008.

[22] Y. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987.

[23] N. Slavnov. The Fredholm determinant representation for the partition function of the six-vertex model. *Journal of Mathematical Sciences*, 115(1):2058–2065, 2003. ISSN 1072-3374. doi: 10.1023/A:1022664216120. URL http://dx.doi.org/10.1023/A%3A1022664216120.

[24] E. Takimoto and M. Warmuth. The last-step minimax algorithm. In *Proc. 11th International Conference on Algorithmic Learning Theory*, pages 279–290, 2000.

[25] E. Weisstein. Fatou's lemma., February 2012. URL http://mathworld.wolfram.com/FatousLemma.html.

[26] E. Weisstein. Lebesgue's dominated convergence theorem., February 2012. URL http://mathworld.wolfram.com/LebesguesDominatedConvergenceTheorem.html.

[27] Q. Xie and A. Barron. Minimax redundancy for the class of memoryless sources. *Information Theory, IEEE Transactions on*, 43(2):646–657, 1997. ISSN 0018-9448. doi: 10.1109/18.556120.

[28] D. Zhu and J. Galbraith. A generalized asymmetric student-t distribution with application to financial econometrics. CIRANO Working Papers 2009s-13, CIRANO, April 2009. URL http://ideas.repec.org/p/cir/cirwor/2009s-13.html.