JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–23

Horizon-Independent Optimal Prediction with Log-Loss in Exponential Families

Peter Bartlett BARTLETT@CS.BERKELEY.EDU
University of California at Berkeley,
Queensland University of Technology

Peter Grünwald PETER.GRUNWALD@CWI.NL
CWI P.O. Box 94079 NL-1090 GB Amsterdam The Netherlands

Peter Harremoës HARREMOES@IEEE.ORG
Copenhagen Business College Denmark

Fares Hedayati FARESHED@EECS.BERKELEY.EDU
University of California at Berkeley

Wojciech Kotłowski WKOTLOWSKI@CS.PUT.POZNAN.PL
Poznań University of Technology Poland

Abstract

We study online learning under logarithmic loss with regular parametric models. Hedayati and Bartlett (2012b) 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, and if and only if the optimal strategy can be calculated without knowing the time horizon in advance. They put forward the question what families have exchangeable SNML strategies. This paper fully answers this open problem for one-dimensional exponential families. The exchangeability can happen only for three classes of natural exponential family distributions, namely the Gaussian, Gamma, and the Tweedie exponential family of order $3/2$.

Keywords: SNML Exchangeability, Exponential Family, Online Learning, Logarithmic Loss, Bayesian Strategy, Jeffreys Prior, Fisher Information

1. Introduction

We work in the setting of online learning under logarithmic loss. Let $x_1, x_2, \ldots$, be a sequence of outcomes from $\mathcal{X}$ revealed one at a time. We use x^t to denote $(x_1, x_2, \cdots, x_t)$, and x_m^n to denote $(x_m, x_{m+1}, \cdots, x_n)$. At round t, after observing x^{t-1}, a forecaster assigns a probability distribution on $\mathcal{X}$, denoted $p\left(\cdot \mid x^{t-1}\right)$. Then, after x_t is revealed, the forecaster incurs the *log loss* $-\ln p\left(x_t \mid x^{t-1}\right)$. The performance of the strategy is measured relative to the best in a reference set of strategies (experts). The difference between the accumulated loss of the prediction strategy and the best expert in the reference set is called the *regret* (Cesa-Bianchi and Lugosi, 2006). The goal is to minimize the regret in the worst case over all possible data sequences.

In this paper 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. If there is a known time horizon n of the game (sequence

length), a well-known result in the literature states that the minimax regret is achieved by the *normalized maximum likelihood* strategy, NML for short (Shtarkov, 1987; Rissanen, 1996). If the parameter space of a d-dimensional exponential family is constrained to a compact subset of the parameter space, NML achieves regret $\frac{d}{2}\ln n + O(1)$. For unconstrained parameter spaces, the NML strategy is often not defined because it relies on finiteness of the Shtarkov sum (or integral) and in many application this sum is infinite. In these cases NML can be replaced by the *conditional normalized maximum likelihood strategy* (CNML), 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), CNML is optimal in the sense that it achieves minimax *conditional* regret. Unfortunately both CNML 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 CNML, known as *sequential normalized maximum likelihood*, or SNML for short (Takimoto and Warmuth, 2000a; Rissanen and Roos, 2007; Roos and Rissanen, 2008).

SNML predictions coincide with those of the CNML distribution under the assumption that the current iteration is the last iteration. Therefore, SNML can be viewed as an approximation to CNML for which the time horizon of the game does not need to be known. Kotlowski and Grünwald (2011) 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 CNML coincide? We believe answering this question is of fundamental importance for online learning at least from the two following 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 becomes time-horizon independent: we get the same, optimal strategy no matter how far to the future we look. Thus, we only need to analyze the worst case regret with respect to the current outcome to be predicted.

Secondly, it has been shown (Kotlowski and Grünwald, 2011; Hedayati and Bartlett, 2012a,b; Harremoës, 2013) that when CNML and SNML coincide, then they become Bayesian strategies and the prior of the Bayesian strategy must be Jeffreys prior. In other words, if CNML is time-horizon independent, then the Bayesian strategy with Jeffreys prior is the (conditional) minimax strategy. Hedayati and Bartlett (2012a,b) showed that this happens if and only if SNML strategy is *exchangeable*. Testing the exchangeability of the sequential strategy is, however, hard, and does not lead to a simple characterization of exponential families for which CNML=SNML holds. Therefore, Hedayati and Bartlett (2012b) put forward the following question: in the case of exponential families, what families have exchangeable sequential normalized maximum likelihood strategies?

In this paper we give a complete answer to the CNML=SNML question, when the reference set of experts is a single-parameter exponential family. We show that there are essentially only three exponential families with time-horizon independent minimax strategy (and hence optimal Bayesian strategy with Jeffreys prior and optimal SNML). These families are Gamma, Gaussian, and Tweedie 3/2 families (but also included are those families, which can be obtained by a fixed transformation of variable from any of the three above, e.g. Pareto, Laplace, Rayleigh and many others). This means that only in these families, a Bayesian strategy with Jeffreys prior is equivalent to SNML and both are equivalent to CNML and hence optimal in the minimax conditional regret sense. More interestingly this implies that only in these three families CNML becomes independent of the horizon, so that one–stead ahead lookup becomes equivalent to n–step ahead lookup, where n is the amount of data the player is eventually going to observe.

The paper is organized as follows. We introduce the mathematical context for our results in Section 2. We then give our main result in Section 3, showing that Gamma, Gaussian and Tweedie 3/2 family are the only families with time-horizon independent minimax strategies. Short versions of the proofs are given in Section 3, and detailed proofs can be found in the appendix.We end with a short discussion in Section 4.

2. Set-Up

We work in the setup of Hedayati and Bartlett (2012b) and use their definitions and notation except that we follow Grünwald (2007) in the distinction between NML and CNML.

A sequential prediction strategy (or just 'strategy' for short) p is any sequential probability assignment that, given a history x^{t-1}, defines $p\left(\cdot \left| x^{t-1}\right.\right)$, the conditional density of $x_t \in \mathcal{X}$ with respect to a fixed underlying measure λ on $\mathcal{X}$. As an example, we usually take λ to be the counting measure if $\mathcal{X}$ is discrete; and if $\mathcal{X} = \mathbb{R}^d$, λ is taken to be Lebesgue measure.

A prediction strategy defines a joint distribution p on sequences of elements of $\mathcal{X}$ in the obvious way,

$$p\left(x^n\right) = \prod_{t=1}^{n} p\left(x_t \left| x^{t-1}\right.\right).$$

Conversely, any probability distribution $p\left(x^n\right)$ on the set $\mathcal{X}^n$ defines a prediction strategy induced by its conditional distributions $p\left(\cdot \left| x^{t-1}\right.\right)$ for $1 \leq t \leq n$ (Cesa-Bianchi and Lugosi, 2006; Grünwald, 2007).

We try to come up with strategies which predict as well as the best element of a reference set of 'experts', which we take to be 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

$$p_\theta(x) = h(x)\mathrm{e}^{\theta x - A(\theta)}, \ \theta \in \Theta. \tag{1}$$

Here h is a reference measure, given as a density relative to the underlying measure λ. Here 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

$$\Theta_{\text{full}} = \{\theta \in \mathbb{R} \,|\, A(\theta) < \infty\} \tag{2}$$

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*.

The families are extended to n outcomes by taking product distributions: $p_\theta(x^n) = \prod_{t=1}^n p_\theta(x_t)$. In this way each member of the family defines a prediction strategy p_θ such that $p_\theta\left(x_t \left| x^{t-1}\right.\right) := p_\theta(x_t)$ with $p_\theta(x_t)$ given by (1). Note that we never assume that data are i.i.d.; only the set of predictors we compare ourselves to treat it as i.i.d.

According to the standard general definition of exponential families (Barndorff-Nielsen, 1978), we can have $\theta f(x)$ instead of θx in the exponent of (1), for an arbitrary fixed function f. Families with $f(x) = x$ are called *natural* exponential families relative to random vector X (defined as $X(x) = x$).

However, as long as f is smooth and 1-to-1, 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 11.

Given a fixed horizon n and a parameter space Θ, the NML strategy (Shtarkov, 1987; Rissanen, 1996) is defined via the joint probability distribution

$$p_{nml}^{(n)}(x^n) = \frac{\sup_{\theta\in\Theta} p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, \mathrm{d}\lambda^n(y^n)}, \tag{3}$$

provided that the so-called *Shtarkov integral* in the denominator exists. To ensure that the NML-distribution exists we will assume that the parameter space is closed. For $t \le n$, the conditional probability distribution is

$$p_{nml}^{(n)}\left(x_t \left| x^{t-1}\right.\right) = \frac{p_{nml}^{(n)}(x^t)}{p_{nml}^{(n)}(x^{t-1})} \tag{4}$$

where $p_{nml}^{(n)}(x^t)$ and $p_{nml}^{(n)}(x^{t-1})$ are marginalized joint probability distributions of $p_{nml}^{(n)}(x^n)$:

$$p_{nml}^{(n)}(x^t) = \int_{\mathcal{X}^{n-t}} p_{nml}^{(n)}(x^n)\, \mathrm{d}\lambda^{n-t}(x_{t+1}^n).$$

Note that the expression for the conditional distribution of a full-length complement of a sequence x_{t+1}^n given the initial part of the sequence x^t then simplifies to:

$$p_{nml}^{(n)}\left(x_{t+1}^n \left| x^t\right.\right) = \frac{p_{nml}^{(n)}(x^n)}{p_{nml}^{(n)}(x^t)} = \frac{\sup_{\theta\in\Theta} p_\theta(x^n)}{\int_{\mathcal{X}^{n-t}} \sup_{\theta\in\Theta} p_\theta(x^t y^{n-t})\, \mathrm{d}\lambda^{n-t}(y^{n-t})}. \tag{5}$$

In many cases the NML strategy is undefined, due to the normalization factor (Shtarkov integral) being infinite. In such cases, by conditioning on a fixed initial sequence of length m the problem usually goes away. The resulting *conditional NML* (CNML) distribution achieves the minimax *conditional* regret (Grünwald, 2007, Chapter 11). CNML is defined via the conditional probability distribution in the following way

$$p_{cnml}^{(n)}\left(x_{m+1}^n \left| x^m\right.\right) = \frac{\sup_{\theta\in\Theta} p_\theta(x^n)}{\int_{\mathcal{X}^{n-m}} \sup_{\theta\in\Theta} p_\theta(x^m y^{n-m})\, \mathrm{d}\lambda^{n-m}(y^{n-m})}. \tag{6}$$

Note that (6) coincides with (5), so CNML can be considered a generalization of NML. NML and CNML are costly due to the amount of marginalization at each round. Furthermore they are horizon–dependent, i.e. the predictions to be made depend on the amount of data that will eventually be seen. Grünwald (2007) discusses in detail why this can be problematic. Two alternative strategies which avoid these issues are the Bayesian strategies with Jeffreys prior and the sequential normalized maximum likelihood strategy, SNML for short, as developed by Rissanen and Roos (2007); Roos and Rissanen (2008). SNML is defined via the conditional probability distribution in the following way

$$p_{snml}\left(x_t \left| x^{t-1}\right.\right) = p^{(t)}_{cnml}\left(x_t \left| x^{t-1}\right.\right).$$

Kotlowski and Grünwald (2011) showed that SNML provides a reasonably good approximation of CNML. At each point in time $t-1$, the SNML strategy for predicting the next outcome x_t may be viewed as the strategy that would lead to minimax optimal conditional regret if the next step was the last round of the game. Hence, it is essentially a *last-step minimax* strategy in the sense of Takimoto and Warmuth (2000b).

The other alternative, the Bayesian strategy with Jeffreys prior, is also defined via its conditional distributions as

$$p_\pi\left(x_t \left| x^{t-1}\right.\right) = \int_{\theta\in\Theta} p_\theta\left(x_t\right) \,\mathrm{d}\pi\left(\theta \left| x^{t-1}\right.\right).$$

Here $\pi\left(\theta \left| x^{t-1}\right.\right)$ is the posterior distribution based on prior $\pi(\cdot)$ and $\pi(\cdot)$ is *Jeffreys prior* defined to be proportional to $I(\theta)^{1/2}$ with I being the Fisher information. A well-known result in the literature says that if the parameter space is effectively smaller than the natural parameter space then the Bayesian strategy with Jeffreys prior is asymptotically minimax optimal (See chapters 7 and 8 in Grünwald (2007)). The nice thing about these two alternatives is that unlike CNML they are defined naturally via conditional probability distributions that are much easier to compute. In general Jeffreys prior cannot be normalized (i.e. $\int I(\theta)^{1/2}\,\mathrm{d}\theta = \infty$) but for all models used in applications its posterior after just one single observation is proper (i.e. well-defined) and can be used for predictions; see below Lemma 4 for details (note though that there exist pathological models where no finite number of observations will give a proper Jeffreys posterior (Harremoës, 2013)).

Hedayati and Bartlett (2012b) proved that these two alternatives are exactly the same as CNML and hence optimal if and only if SNML is exchangeable. Let p be any time horizon-independent sequential prediction strategy conditioned on an initial sequence x^m, which for any $n > m$ and any x^n_{m+1}, assigns a joint probability distribution $p\left(x^n_{m+1}\left|x_m\right.\right)$. We say that p is *exchangeable* if for any $n > m$, any $x_n \in \mathcal{X}^n$, the joint probability $p\left(x^n_{m+1}\left|x_m\right.\right)$ assigned to x^n_{m+1} is invariant under any permutation σ on $\{1,\ldots,n\}$ which leaves the initial part of data x^m unchanged.

Thus, exchangeability of SNML means that the joint distribution of SNML conditioned on initial data x^m is invariant under any permutation of the data sequence x^n_{m+1}. Exchangeability of SNML is usually hard to check. The natural question to ask is whether there exists an equivalent characterization that can be easily read off of the distribution or not? In this paper we show that there are only three types of exponential family distributions that have exchangeable SNML. For none of the these three families the denominator in Equation 3 is

finite. Hence, for all one-dimensional exponential families in which NML is defined it will be horizon dependent and can neither agree with SNML nor with a Bayesian strategy.

3. Main Results

We now provide a sequence of lemmas and theorems that lead up to our main result, Theorem 11. We provide a full proof of Lemma 3 and the final Theorem 11 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; full proofs of these results are in the appendix. We first provide a number of definitions that will be used repeatedly.

3.1. Definitions

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

Our analysis below involves various parameterizations of natural exponential families, in particular the natural, the mean (see below) and the geodesic (only used in the appendix) 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 (natural, mean and geodesic parameterizations are all instances of 'diffeomorphic' parameterizations (Grünwald, 2007, page 611)). We then denote parameters by γ and we let $\hat{\gamma}(x^n)$ be the maximum likelihood 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 used 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 1-to-1 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 $V(\mu)$ (Grünwald, 2007, Chapter 18). We also introduce the standard deviation σ as a function of the mean by $\sigma(\mu) = V(\mu)^{1/2}$.

Definition 1 (convex core) *Consider a natural exponential family as in (1). Let* $x_0 = \inf\{x : x \in \textit{support of } h\}$, *and* $x_1 = \sup\{x : x \in \textit{support of } h\}$. *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 in* x_0, *and with* x_1 *included if and only if* h *has a point mass in* x_1. *We denote this 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 2 (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.

3.2. Lemmas that Abstractly Characterize SNML-Exchangeability

We now present three lemmas, which give an abstract characterization of SNML exchangeability. Then in Section 3.3 we will make these concrete, leading to our main theorem.

We let m be the smallest n such that for all $x^n \in \mathcal{X}^n$, $\int p_\gamma(x^n) I(\gamma)^{1/2} \, \mathrm{d}\gamma < \infty$ and $\int_{\mathcal{X}^{k-n}} \sup_{\gamma\in\Gamma} p_\gamma\left(x^n, y^{k-n}\right) \, \mathrm{d}\lambda^{k-n}\left(y^{k-n}\right) < \infty$ for $k \geq n$, i.e. such that Jeffreys' posterior

$$\pi(\gamma \,|x^n) := \frac{p_\gamma(x^n) I(\gamma)^{1/2}}{\int p_\gamma(x^n) I(\gamma)^{1/2} \, \mathrm{d}\gamma}$$

is proper (integrates to 1) for any conditioning sequence of length equal to or longer than m, and that the conditional minimax regret is finite. In most applications $m = 1$. Note that this implies that CNML and SNML conditioned on an initial sequence of length m exist (Harremoës, 2013), so that all three prediction strategies (Bayes with Jeffreys, CNML and SNML) are well-defined. From now on, each time we mention CNML/SNML we mean "CNML/SNML conditioned on an initial sequence of length m".

We call the distribution p_γ *regular* if, for all x^n with $\hat{\gamma}(x^n) = \gamma$,

$$\mu_\gamma = \hat{\mu}(x^n) = E_{p_\gamma}[X] = n^{-1} \sum_{i=1}^{n} x_i,$$

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 (Grünwald, 2007, Chapter 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 3 *Consider a natural exponential family as in (1) where the parameter set Γ is an interval. 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:*

$$\int_\Gamma \mathrm{e}^{-nD(\gamma_0\|\gamma)} I(\gamma)^{1/2} \, \mathrm{d}\gamma = C_n. \tag{7}$$

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 $\gamma_0 \in \hat{\Gamma}_n$*

$$\int_\Gamma \mathrm{e}^{-nD(\gamma_0\|\gamma)} I(\gamma)^{1/2} \, \mathrm{d}\gamma = C_n. \tag{8}$$

The essence of Lemma 3 is that C_n remains constant as γ_0 varies. This will be key to proving our main result.

Proof Hedayati and Bartlett (2012b) showed that, if Γ is an interval, then SNML exchangeability is equivalent to that Bayes with Jeffreys prior and CNML coincide. Thus, equivalently, we must have, for *all* $x_1, \cdots, x_n \in \mathcal{X}^n$, and *all* t, such that $n > t \geq m$,

$$p_\pi\left(x_{t+1}^n \middle| x^t\right) = p_{cnml}^{(n)}\left(x_{t+1}^n \middle| x^t\right). \tag{9}$$

Since

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

and

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

in the diffeomorphic parametrization Γ, (9) is equivalent to

$$\int_\Gamma p_\gamma\left(x^n\right) I(\gamma)^{1/2} \mathrm{d}\gamma = C\left(n, x^t\right) \times p_{\hat{\gamma}(x^n)}\left(x^n\right), \tag{10}$$

where

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

We now prove that $C\left(n, x^t\right) = C_n$, i.e. it may depend on n but *it does not depend on* $x_1, \ldots, x_n$. The key observation is that (10) is satisfied for any $t \geq m$, in particular for $t = m$, so that $C\left(n, x^t\right)$ cannot depend on x_{m+1}^n. However, since $C\left(n, x^t\right)$ and all other terms in (10) are invariant under any permutation of x^t, we conclude that $C\left(n, x^t\right)$ does not depend on the whole sequence x^n.

Now we divide both sides of (10) by $p_{\hat{\gamma}(x^n)}(x^n)$ and we exponentiate inside the integral. This gives:

$$\int_\Gamma \mathrm{e}^{-\ln \frac{p_{\hat{\gamma}(x^n)}(x^n)}{p_\gamma(x^n)}} I(\gamma)^{1/2} \mathrm{d}\gamma = C_n. \tag{11}$$

We have thus shown that, assuming Γ is an interval, SNML exchangeability is equivalent to the condition that (11) holds for a fixed C_n, for all $x^n \in \mathcal{X}^n$.

Now for Part (1), Let $\gamma_0 = \hat{\gamma}\left(x^n\right)$. We now use the celebrated robustness property of exponential families (Grünwald, 2007, 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 $\hat{\gamma}(x^n) = \gamma$, we have

$$nD\left(\gamma_0 \| \gamma\right) = \ln \frac{p_{\hat{\gamma}(x^n)}\left(x^n\right)}{p_\gamma\left(x^n\right)}; \tag{12}$$

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 (Grünwald, 2007, Section 19.3, Eq. 19.10). The only-if direction follows immediately by Part (1). To see the converse,

we 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 Csiszár and Matús (2003)), and all $\gamma \in \Gamma$ are regular. Hence Equation (12) holds for all $x^n \in \mathcal{X}^n$ so that (8) implies that (11) 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 *Consider a natural exponential family as in (1). If the family is SNML 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 non-empty interior, thus we may assume that it contains an interval. We can then show by approximating the integral in (7) by a Gaussian integral using standard Laplace-approximation techniques (as in e.g. (Grünwald, 2007, Chapter 7)) that, for general 1-dimensional exponential families, the integral in (7) 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 3 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 (7) gives that the integral converges to $(1/2)\,(2\pi/n)^{1/2}$ and we have a contradiction.

3.3. Preparing and Stating the Main Theorem

In the following we will use the *Tweedie exponential families* of order $3/2$ Jørgensen (1997). These are natural exponential families characterized by a variance function of the form $V(\mu) = k\mu^{3/2}$, where μ is the mean and $V(\mu)$ 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. It is obtained as follows. Let X_i denote i.i.d. exponentially distributed random variables with mean ν and let N denote an independent Poisson distributed random variable with mean νk^{-2}. Then the elements in the exponential family are distributions of random variables of the form

$$Z = \sum_{i=1}^{N} X_i$$

for different values of the parameter ν (Jørgensen, 1997). It is interesting to note that such distributions have a point mass in 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 5 *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 (8) holds for these families, which is relatively straightforward by taking derivatives of the cumulant generating function.

Remark 6 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 7 *Suppose* $\{p_\gamma(\cdot)|\,\gamma\in\Gamma\}$ *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* $\{q_\gamma(\cdot)|\,\gamma\in\Gamma\}$ *is SNML exchangeable as well.*

Proof Sketch This is (almost) immediate from the definition of exchangeability.

Example 1 *As an example consider a random variable* X *with a Gamma distribution of the form Gamma* $(1/2, c/2)$ *with density*

$$\left(\frac{c}{2\pi}\right)^{1/2} x^{-1/2}\mathrm{e}^{-\frac{xc}{2}}.$$

Now if X *goes through the one–to–one transformation* $f(X)=1/X$ *then*

$$\frac{1}{X}\sim \textit{Inverse-Gamma}\,(1/2, c/2)$$

with density

$$\left(\frac{c}{2\pi}\right)^{1/2} x^{-3/2}\mathrm{e}^{-\frac{c}{2x}}.$$

This is the same as the density of a Levy $(0,c)$*. Hence Levy* $(0,c)$ *is also SNML exchangeable. It is indeed easy to directly verify the SNML exchangeability of Levy* $(0,c)$ *using Lemma 3.*

Theorem 8 *Consider a natural exponential family as in (1). A necessary condition for SNML exchangeability is that the standard deviation as function of the mean satisfies the differential equation*

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2+3\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}=\mathrm{const}(\mu). \tag{13}$$

Proof Sketch By Lemma 4 we may assume that the family has maximal mean-value parameter space. A fifth-order (!) Taylor expansion of (8) rewritten in the geodesic parameterization (see (25) in the appendix) gives terms of different order in n, and each term should be constant. Equation 13 corresponds to the first non-trivial term in the expansion.

Theorem 9 *Consider a natural exponential family as in (1) with maximal mean-value parameter space. A necessary condition for SNML exchangeability is that the variance function is given by*

$$V(\mu)=(k\mu+\ell)^2 \tag{14}$$

or

$$V(\mu)=(k\mu+\ell)^{3/2} \tag{15}$$

for some constants k *and* ℓ*.*

Proof Sketch The differential equation (13) can be rephrased in terms of the variance function. Two solutions are (14) or (15). Other potential solutions are ruled out by a higher-order (in fact 7th-order!!) expansion.

Now we are ready to state our main theorem. We need one more definition: we say that a full exponential family of form (1) is a *linear transformation* of another full exponential family if, for some fixed a, b, it is the set of distributions given by (1), with each occurrence of x replaced by $ax + b$.

Remark 10 By Remark 6, linear transformations preserve SNML exchangeability. In a Gaussian location family translating by b replaces a distribution by another distribution of the same exponential family and the Gaussian location families are the only families with this property. Scaling of a Gamma distribution by positive a gives another Gamma distribution in the same exponential family and the Gamma families are the only exponential families with this property.

Theorem 11 *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 it.*
- *The full Tweedie exponential family of order $3/2$ and linear transformations of it.*

Before we prove this theorem, let us briefly discuss its generality.

As we already indicated below (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 1-to-1. 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 6; the latter is the case, for, for example, the Bernoulli and Poisson distributions.

Proof Lemma 5 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), 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 9 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 14, where the variance function is quadratic. Exponential families with quadratic variance functions have been classified by Morris (1982). 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 15. 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

$$V(\mu) = 2\mu^{3/2}. \tag{16}$$

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 Morris (1982), the Tweedie family of order $3/2$ is the only family satisfying (16). ■

4. Discussion

The present paper has focused on 1-dimensional exponential families with non-empty interior parameter spaces. 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, but the limited space did not allow us to go into this problem here.

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.

One of the consequences of this paper is that for 1-dimensional exponential families, NML (if it is defined without conditioning) will always be horizon dependent. We conjecture that this conclusion will hold for arbitrary models. Only conditional versions of NML allow the kind of consistency that we call SNML exchangeability, and even after conditioning, SNML exchangeability is restricted to a few but very important models.

Acknowledgements

Wojciech Kotłowski has been supported by the Foundation of Polish Science under the Homing Plus programme. We gratefully acknowledge the support of the NSF through grant CCF-1115788 and of the Australian Research Council through Australian Laureate Fellowship FL110100281.

References

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

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998. Special Commemorative Issue: Information Theory: 1948-1998.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

I. Csiszár and F. Matús. Information projections revisited. *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1474–1490, 2003.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

P. Harremoës. Extendable MDL. Accepted for presentation at *International Symposium for Information Theory (ISIT 2013)*, ArXiv: 1301.6465, Jan. 2013.

F. Hedayati and P. Bartlett. Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS '12)*, 2012a.

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

A. Hurwitz. Über die angenäherte Darstellung der Irrationalzahlen durch rationale Brüche (On the approximation of irrational numbers by rational numbers, in German). Mathematische Annalen 39 (2): 279–284, 1891.

B. Jørgensen. *The Theory of Dispersion Models*. Chapman & Hall, 1997.

S. Kakade, M. Seeger, and D. Foster. Worst-case bounds for Gaussian process models. In *Proceedings of the 2005 Neural Information Processing Systems Conference (NIPS 2005)*, 2006.

W. Kotłowski and P. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the Twenty-Fourth Conference on Learning Theory (COLT' 11)*, 761–779, Budapest, 2011.

F. Liang and A. R. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50: 2708–2726, 2004.

C. Morris. Natural exponential families with quadratic variance functions. *Ann. Statist.*, 10:65–80, 1982.

J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

J. Rissanen and T. Roos. Conditional NML universal models. In *Information Theory and Applications Workshop (ITA-07)*, 337–341, 2007.

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

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

E. Takimoto and M. Warmuth. The last-step minimax algorithm. In *Conference on Algorithmic Learning Theory (ALT '00)*, 2000a.

E. Takimoto and M. Warmuth. The minimax strategy for Gaussian density estimation. In *Conference on Learning Theory (COLT '00)*, 100–106, 2000b.

Appendix

In the proofs we introduce τ as short for 2π.

Proof [Proof of Lemma 4] Without loss of generality consider the mean-value parameter space. Assume that 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 3 we have for all n, and all regular points in $x \in \hat{M}_n \cap [\mu_0, \mu_1]$ that

$$\int_{[\mu_0,\mu_1]} \frac{\mathrm{e}^{-nD(x\|\mu)}}{\sigma(\mu)} \, \mathrm{d}\mu = C_n \tag{17}$$

is independent of x. Note that all points in the interior of $[\mu_0, \mu_1]$ must be regular (Grünwald, 2007, Section 19.3, Eq. 19.10).

By a standard Laplace approximation of the integral in (17) (done by a Taylor approximation of the KL divergence, $D(x\|\mu) \approx \frac{1}{2}(x-\mu)^2 V(x)^{-1}$, 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, that for each x in the interior of M_c,

$$\frac{\int_{M_c} \frac{\mathrm{e}^{-nD(x\|\mu)}}{\sigma(\mu)} \, \mathrm{d}\mu}{\left(\frac{\tau}{n}\right)^{1/2}} \to 1 \tag{18}$$

and

$$\frac{\int_{\{\mu \in M_c : \mu \geq x\}} \frac{\mathrm{e}^{-nD(x\|\mu)}}{\sigma(\mu)} \, \mathrm{d}\mu}{\left(\frac{\tau}{n}\right)^{1/2}} \to \frac{1}{2}. \tag{19}$$

For a precise statement and proof of these results, see e.g. (Grünwald, 2007, Theorem 8.1 combined with Eq. (8.14)). Combining (18) with (17), taking $M_c = [\mu_0, \mu_1]$, it follows that $C_n \to (\tau/n)^{1/2}$. 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 (17) also holds for $x = \mu_0$, so we get

$$\frac{\int_{[\mu_0,\mu_1]} \frac{\mathrm{e}^{-nD(\mu_0\|\mu)}}{\sigma(\mu)} \, \mathrm{d}\mu}{\left(\frac{\tau}{n}\right)^{1/2}} \to 1. \tag{20}$$

Now assume by means of 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 $M_c = M'$, (19) with $x = \mu_0$ gives that the same integral as in (20) 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. ■

Proof [Lemma 5] For each of the families it is sufficient to prove that

$$\int_{cc} \frac{\mathrm{e}^{-nD(\gamma_0\|\gamma)}}{\sigma(\gamma)}\,\mathrm{d}\gamma$$

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 $D(\gamma_0\|\gamma) = D(0\|\gamma - \gamma_0)$, and $V(\gamma) = \sigma^2$, so the integral is invariant because of the invariance of the Lebesgue integral.

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

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

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

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

Therefore we have to prove that the following integral is constant

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

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. ■

Proof [Lemma 7] Since the family $p_\gamma(\,\cdot\,)$ is SNML exchangeable, for any $n > m$ the following joint distribution is invariant under permutations of x^n that leaves x^m invariant:

$$p_{snml}\left(x_{m+1}^n \middle| \, x^m\right) = \prod_{t=m+1}^{n} \frac{\sup_\gamma p_\gamma(x^t)}{\int_{\mathcal{X}} \sup_\gamma p_\gamma(x^{t-1}, x) \, \mathrm{d}x} \tag{21}$$

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

$$q_\gamma(y) = p_\gamma(f^{-1}(y)) \left| \frac{\mathrm{d}\, f^{-1}(y)}{\mathrm{d}\, y} \right| . \tag{22}$$

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

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

Hence $p_{snml}\left(y_{m+1}^n \middle| \, y^m\right)$ 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 $v(y)\mathrm{d}y = \mathrm{d}x$. ■

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:

$$\begin{aligned} \beta &= \int_{\theta_0}^{\theta} I(s)^{1/2}\,\mathrm{d}s && (23)\\ &= \int_{\mu_0}^{\mu} \frac{1}{\sigma(t)}\,\mathrm{d}t, && (24) \end{aligned}$$

so that $\mathrm{d}\beta = I(\theta)^{1/2}\,\mathrm{d}\theta = \mathrm{d}\mu/\sigma(\mu)$. Note that this is a bijection. This allows us to replace the integration measure in the condition of Lemma 3 and we get a condition equivalent to (8): for any $n > m$ the following is independent of $\beta_0 \in \hat{B}^n$

$$\int_B \mathrm{e}^{-nD(\beta_0\|\beta)}\,\mathrm{d}\beta. \tag{25}$$

Proof [Theorem 8] We denote the integral in Equation 25 by $s(\beta_0, n)$. We may assume the family has maximal mean-value parameter space, so that (25) must hold for all $\beta_0 \in \hat{B}_n$, all n. First we will establish the following relation between the geodesic parametrization and the mean value parameterization

$$\begin{aligned} \frac{\partial}{\partial\beta}(\ldots) &= \frac{\mathrm{d}\mu}{\mathrm{d}\beta}\frac{\partial}{\partial\mu}(\ldots)\\ &= \sigma(\mu)\frac{\partial}{\partial\mu}(\ldots), \end{aligned}$$

because $\frac{\mathrm{d}\beta}{\mathrm{d}\mu} = \sigma^{-1}(\mu)$. We use the fact that $D(\beta_0\|\beta) = D(\mu_0\|\mu)$, where $\mu = \mu(\beta)$ and $\mu_0 = \mu(\beta_0)$ are corresponding parameters in different parametrizations.

$$\begin{aligned} D(\beta_0\|\beta) &= \mu_0\cdot(\theta_0-\theta) + A(\theta) - A(\theta_0)\\ \frac{\partial D(\beta_0\|\beta)}{\partial\beta} &= (\mu-\mu_0)\cdot\sigma^{-1},\\ \frac{\partial^2 D(\beta_0\|\beta)}{\partial\beta^2} &= 1-(\mu-\mu_0)\cdot\sigma^{-1}\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}. \end{aligned} \tag{26}$$

Hence $D_2 = 1$, where D_n denotes $\frac{\partial^n D(\beta_0\|\beta)}{\partial\beta^n}\Big|_{\beta=\beta_0}$ throughout this section.

A Taylor expansion of Equation 25 as function of n gives that certain Taylor coefficients must equal zero and an elaborate calculation of the Taylor coefficient leads to Equation 13.

Using a fifth-order Taylor expansion we will show the following:

$$s(\beta_0, n) = \Phi + n^{-3/2}\cdot 3\tau^{1/2}\cdot u(\beta_0) + O\left(n^{-2}\right) \tag{27}$$

where

$$u(\beta_0) = \frac{5}{2}\cdot\left(\frac{D_3}{3!}\right)^2 - \frac{D_4}{4!}, \tag{28}$$

$\Phi = \frac{\tau^{1/2}}{n^{1/2}}$ is a Gaussian integral (scaled by n), and the n^{-2} remainder term may be negative or positive. Condition 13 easily follows from Equation 27 as follows: take β_0, β_1 in $\hat{B}^\circ$. By

Equation 25 we must have that $s(\beta_0, n) - s(\beta_1, n) = 0$ for all large n. But by Equation 27 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 $u(\beta_0) = u(\beta_1)$. Since we can do this for any β_0 and β_1, Condition 13 follows.

Now we proceed to prove the claim in Equation 27. 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

$$s(\beta_0, n) = f(\beta_0, n) + g(\beta_0, n) + h(\beta_0, n) \tag{29}$$

where we define:

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

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

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

We have

$$g \leq \sup_{\beta' > \beta_0 + c} e^{-(n-m)D(\beta_0 \| \beta')} \int_{\beta > \beta_0 + c} e^{-mD(\beta_0 \| \beta)} \, d\beta \leq c_2 e^{-c_3 n^{c_4}} \tag{30}$$

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

$$f(\beta_0, n) = \int_A e^{-n\frac{1}{2}(\beta_0 - \beta)^2} \left(e^{-n\frac{D_3}{3!}(\beta_0 - \beta)^3} e^{-n\frac{D_4}{4!}(\beta_0 - \beta)^4} e^{-n \cdot O(\beta_0 - \beta)^5} \right) \, d\beta \tag{31}$$

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- and fifth-order terms in the integral can itself be well approximated by a first-order Taylor approximation of e^x and we can rewrite f as

$$\int_A e^{-n\frac{1}{2}(\beta_0 - \beta)^2} \left(e^{-n\frac{D_3}{3!}(\beta_0 - \beta)^3} (1 + V)(1 + W) \right) \, d\beta$$

where $V = -n\frac{D_4}{4!}(\beta_0 - \beta)^4 + O\left(n^2 (\beta_0 - \beta)^8\right)$ and $W = O\left(n(\beta_0 - \beta)^5\right)$. Similarly, the second factor in the integral can be well-approximated by a second order Taylor approximation of $e^x = 1 + x + (1/2)x^2 + O(x^3)$ so that we can further rewrite f as

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

where

$$U = -n\frac{D_3}{3!}(\beta_0-\beta)^3 + \frac{1}{2}n^2\left(\frac{D_3}{3!}\right)^2(\beta_0-\beta)^6 + O\left(n^3(\beta_0-\beta)^9\right).$$

Writing $\Phi_A := \int_A e^{-n\frac{1}{2}(\beta_0-\beta)^2}\,d\beta$ we can thus further rewrite f as

$$f = \Phi_A + \int_A e^{-n\frac{1}{2}(\beta_0-\beta)^2}(U+V+R_1+R_2)\,d\beta$$

where R_1 and R_2 are remainder terms,

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

and

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

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

$$f = \Phi_A + \int_A e^{-n\frac{1}{2}(\beta_0-\beta)^2}(U+V)\,d\beta + O\left(n^{-2}\right).$$

Now, using the fact that $\int_{-a}^{a} x^3 e^{-nx^2}\,dx = 0$ 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

$$f = \Phi_A + \int_A e^{-n\frac{1}{2}(\beta_0-\beta)^2}\left(\frac{1}{2}n^2\left(\frac{D_3}{3!}\right)^2(\beta_0-\beta)^6 - n\frac{D_4}{4!}(\beta_0-\beta)^4\right)d\beta + O\left(n^{-2}\right).$$

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

$$\begin{aligned} f &= \Phi + \frac{n^2}{2}\left(\frac{D_3}{3!}\right)^2 \cdot \left(15\frac{\tau^{1/2}}{n^{7/2}}\right) - n\frac{D_4}{4!}\cdot\left(3\frac{\tau^{1/2}}{n}\right) + O(n^{-2}) \\ &= \Phi + n^{-3/2}\cdot\tau^{1/2}\cdot\left(15\left(\frac{D_3}{3!}\right)^2 - 6\frac{D_4}{4!}\right) + O\left(n^{-2}\right). \end{aligned} \tag{32}$$

Combining with (29) and (30) that there exists a constant, such that for all $n \ge m$, all $\beta_0 \in \hat{B}^\circ$,

$$5(D_3)^2 - 3D_4 = \text{const}(\beta_0). \tag{33}$$

We rephrase condition (33) in terms of the mean value parameterization, we calculate higher derivatives of the divergence based on (26)

$$\begin{aligned}
\frac{\partial^3 D(\beta_0\|\beta)}{\partial\beta^3} &= -\frac{\mathrm{d}\sigma}{\mathrm{d}\mu} + (\mu-\mu_0)\cdot\left(\sigma^{-1}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - \frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}\right),\\
\frac{\partial^4 D(\beta_0\|\beta)}{\partial\beta^4} &= \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - 2\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} + (\mu-\mu_0)\cdot\left(-\sigma^{-1}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^3 + 2\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} - \sigma\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3}\right),\\
\frac{\partial^5 D(\beta_0\|\beta)}{\partial\beta^5} &= -\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^3 + 2\sigma\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} - 3\sigma^2\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3}\\
&\quad + (\mu-\mu_0)\cdot\left(\sigma^{-1}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - 3\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} + 2\sigma\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}\right)^2 + \sigma\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3} - \sigma^2\frac{\mathrm{d}^4\sigma}{\mathrm{d}\mu^4}\right),\\
\frac{\partial^6 D(\beta_0\|\beta)}{\partial\beta^6} &= \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - 4\sigma\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} - 3\sigma^2\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3} + 4\sigma^2\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}\right)^2 - 4\sigma^3\frac{\mathrm{d}^4\sigma}{\mathrm{d}\mu^4}\\
&\quad + (\mu-\mu_0)\cdot(\dots).
\end{aligned}$$

Then

$$\begin{aligned}
D_3 &= -\frac{\mathrm{d}\sigma}{\mathrm{d}\mu},\\
D_4 &= \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - 2\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2},\\
D_5 &= -\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^3 + 2\sigma\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} - 3\sigma^2\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3},\\
D_6 &= \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - 4\sigma\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} - 3\sigma^2\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3} + 4\sigma^2\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}\right)^2 - 4\sigma^3\frac{\mathrm{d}^4\sigma}{\mathrm{d}\mu^4},
\end{aligned}$$

where, as before, $\mu_0 = \mu(\beta_0)$. Further we get

$$\begin{aligned}
5(D_3)^2 - 3D_4 &= 5\left(-\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - 3\left(\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - 2\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}\right)\\
&= 2\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 + 6\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2}.
\end{aligned}$$

Plugging the above into (33) and rearranging the terms gives the following differential equation for σ

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 + 3\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} = \mathrm{const}(\mu). \tag{34}$$

This is a necessary condition for exchangability. ■

Proof [Theorem 9]

We also need to take a closer look at higher-order terms in the Taylor expansion of the integral (25) and obtain a stronger necessary condition for exchangeability. As in the proof Theorem 8, we expand the integral over $A = [\beta_0 - c, \beta_0 + c]$:

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

where

$$\begin{aligned} X_3 &= -n\frac{D_3}{3!}(\beta_0-\beta)^3 + \frac{1}{2}n^2\left(\frac{D_3}{3!}\right)^2(\beta_0-\beta)^6 + \frac{1}{3!}n^3\left(\frac{D_3}{3!}\right)^3(\beta_0-\beta)^9 \\ &\quad + \frac{1}{4!}n^4\left(\frac{D_3}{3!}\right)^4(\beta_0-\beta)^{12} + O\left(n^5(\beta_0-\beta)^{15}\right) \ , \\ X_4 &= -n\frac{D_4}{4!}(\beta_0-\beta)^4 + \frac{1}{2}n^2\left(\frac{D_4}{4!}\right)^2(\beta_0-\beta)^8 + O\left(n^3(\beta_0-\beta)^{12}\right) \ , \\ X_5 &= -n\frac{D_5}{5!}(\beta_0-\beta)^5 + O\left(n^2(\beta_0-\beta)^{10}\right) \ , \\ X_6 &= -n\frac{D_6}{6!}(\beta_0-\beta)^6 + O\left(n^2(\beta_0-\beta)^{12}\right) \ , \\ X_7 &= -O\left(n(\beta_0-\beta)^7\right) \ . \end{aligned}$$

We assume that condition (33) is satisfied, so that $O\left(n^{-3/2}\right)$ term in the expansion (cf. Equation 27) 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 $O\left(\mathrm{e}^{-cn}\right)$, and $(\beta_0-\beta)^m$ under Gaussian integral over the full real line results in $O\left(n^{-(m+1)/2}\right)$ if m is even, and 0 if m is odd, there will be no terms of order $O\left(n^{-2}\right)$. Therefore, we need to look for terms of order $O\left(n^{-5/2}\right)$. There are five of them and their sum must be independent of β_0 (using similar argument as for the $O\left(n^{-3/2}\right)$ term in the proof of Theorem 8):

$$\begin{aligned} &\frac{1}{4!}n^4\left(\frac{D_3}{3!}\right)^4(\beta_0-\beta)^{12} \ + \ \frac{1}{2}n^2\left(\frac{D_4}{4!}\right)^2(\beta_0-\beta)^8 \ - \ n\frac{D_6}{6!}(\beta_0-\beta)^6 \\ &- \ \frac{1}{2}n^3\left(\frac{D_3}{3!}\right)^2\frac{D_4}{4!}(\beta_0-\beta)^{10} \ + \ n^2\frac{D_3}{3!}\frac{D_5}{5!}(\beta_0-\beta)^8 = \mathrm{const}\,(\beta_0) \ . \end{aligned}$$

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

$$\int \mathrm{e}^{-n\frac{1}{2}(\beta_0-\beta)^2}(\beta_0-\beta)^m \, \mathrm{d}\beta = (m-1)!!\tau^{1/2}n^{-\frac{m+1}{2}} ,$$

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

$$\frac{11!!}{4!(3!)^4}D_3^4 + \frac{7!!}{2(4!)^2}D_4^2 - \frac{5!!}{6!}D_6 - \frac{9!!}{2(3!)^2 4!}D_3^2 D_4 + \frac{7!!}{3!5!}D_3 D_5 = \mathrm{const}\,(\beta_0) ,$$

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:

$$385D_3^4 + 105D_4^2 - 24D_6 - 630D_3^2D_4 + 168D_3D_5 = \text{const}(\beta_0). \tag{35}$$

In order to evaluate D_3, D_4, D_5, and D_6 we calculate the derivatives of σ under the condition that the differential equation (34) is satisfied for some constant c

$$\begin{aligned}
2\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 + 6\sigma\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} &= 3c, \\
\frac{\mathrm{d}^2\sigma}{\mathrm{d}\mu^2} &= \frac{1}{2}\sigma^{-1}c - \frac{1}{3}\sigma^{-1}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2, \\
\frac{\mathrm{d}^3\sigma}{\mathrm{d}\mu^3} &= -\frac{5}{6}\sigma^{-2}\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}c + \frac{5}{9}\sigma^{-2}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^3, \\
\frac{\mathrm{d}^4\sigma}{\mathrm{d}\mu^4} &= \frac{25}{9}\sigma^{-3}c\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - \frac{5}{3}\sigma^{-3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - \frac{5}{12}\sigma^{-3}c^2.
\end{aligned}$$

We plug this into (35), and get:

$$\begin{aligned}
&D_2 = 1, \\
&D_3 = -\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}, \\
&D_4 = \frac{5}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - c, \\
&D_5 = -\frac{10}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^3 + \frac{7c}{2}\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}, \\
&D_6 = \frac{2\cdot 35}{9}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - \frac{215c}{2\cdot 9}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 + \frac{8}{3}c^2.
\end{aligned}$$

We plug this into (35), and get:

$$\begin{aligned}
&385D_3^4 + 105D_4^2 - 24D_6 - 630D_3^2D_4 + 168D_3D_5 \\
= \;& 385\left(-\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 + 105\left(\frac{5}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - c\right)^2 - 24\left(\frac{2\cdot 35}{9}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - \frac{215c}{2\cdot 9}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 + \frac{8}{3}c^2\right) \\
& -630\left(-\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2\left(\frac{5}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - c\right) + 168\left(-\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)\left(-\frac{10}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^3 + \frac{7c}{2}\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right) \\
= \;& 385\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 + 105\cdot\left(\frac{5}{3}\right)^2\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 + 105\cdot c^2 - 35\cdot 2\cdot 5\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 c - \frac{8\cdot 2\cdot 35}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 \\
& + \frac{4\cdot 215c}{3}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 - 8^2c^2 \\
& -210\cdot 5\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 + 630\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2 c + 56\cdot 10\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 - 84\cdot 7c\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2.
\end{aligned}$$

Collecting the terms gives

$$\begin{aligned} & \left(385+105\cdot\left(\frac{5}{3}\right)^2-\frac{8\cdot 2\cdot 35}{3}-210\cdot 5+56\cdot 10\right)\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^4 \\ & -\left(35\cdot 2\cdot 5-\frac{4\cdot 215}{3}-630+84\cdot 7\right)c\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2+\left(105-8^2\right)c^2 \\ = \quad & -\frac{64}{3}c\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2+41c^2 \end{aligned}$$

Interestingly all term with $(\mathrm{d}\sigma/\mathrm{d}\mu)^4$ have disappeared and we get:

$$-\frac{64}{3}c\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}\right)^2+41c^2=\mathrm{const}(\mu). \tag{36}$$

Assume that $c\neq 0$. Equation (36) is satisfied only when

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\mu}=\mathrm{const}(\mu),$$

which has a general solution of the form:

$$\sigma\left(\mu\right)=k\mu+\ell$$

for some constants c_1 and k and we get

$$V\left(\mu\right)=\left(k\mu+\ell\right)^2.$$

Assume $c=0$. We now solve (34). The differential equation can be rewritten as

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\left(\sigma^{4/3}\right)=0.$$

Hence there exists constants k and ℓ such that

$$\sigma^{4/3}=k\mu+\ell$$

or equivalently

$$V\left(\mu\right)=\left(k\mu+\ell\right)^{3/2}.$$

■