Semi-parametric Bayesian Partially Identified Models based on Support Function

Liao, Yuan and Simoni, Anna

University of Maryland, CNRS (Paris, France), THEMA Université de Cergy-Pontoise

December 2012

Semi-parametric Bayesian Partially Identified Models based on Support Function*

Yuan Liao† — University of Maryland

Anna Simoni‡ — CNRS and THEMA

This Version: December 2012

Abstract

Bayesian partially identified models have received a growing attention in recent years in the econometric literature, due to their broad applications in empirical studies. Classical Bayesian approach in this literature has been assuming a parametric model, by specifying an ad-hoc parametric likelihood function. However, econometric models usually only identify a set of moment inequalities, and therefore assuming a known likelihood function suffers from the risk of misspecification, and may result in inconsistent estimations of the identified set. On the other hand, moment-condition based likelihoods such as the limited information and exponential tilted empirical likelihood, though guarantee the consistency, lack of probabilistic interpretations. We propose a semi-parametric Bayesian partially identified model, by placing a nonparametric prior on the unknown likelihood function. Our approach thus only requires a set of moment conditions but still possesses a pure Bayesian interpretation. We study the posterior of the support function, which is essential when the object of interest is the identified set. The support function also enables us to construct two-sided Bayesian credible sets (BCS) for the identified set. It is found that, while the BCS of the partially identified parameter is too narrow from the frequentist point of view, that of the identified set has asymptotically correct coverage probability in the frequentist sense. Moreover, we establish the posterior consistency for both the structural parameter and its identified set. We also develop the posterior concentration theory for the support function, and prove the semi-parametric Bernstein von Mises theorem. Finally, the proposed method is applied to analyze a financial asset pricing problem.

*The authors are grateful to the seminar and conference participants at CREST (LS and LMI), University of Luxembourg, University of Mannheim, THEMA and 4^{th} French Econometrics Conference for useful comments. Anna Simoni gratefully acknowledges financial support from the University of Mannheim through the DFG-SNF Research Group FOR916 and hospitality from CREST. Yuan Liao gratefully acknowledges financial support from the University of Maryland at College Park.

†Department of Mathematics, University of Maryland at College Park, College Park, MD 20742 (USA). Email: yuanliao@umd.edu

‡CNRS and THEMA, Université de Cergy-Pontoise - 33, boulevard du Port, 95011 Cergy-Pontoise (France). Email: simoni.anna@gmail.com

Key words: partial identification, posterior consistency, concentration rate, support function, two-sided Bayesian credible sets, identified set, coverage probability, moment inequality models

JEL code: C110, C140, C58

1 Introduction

1.1 Bayesian inference for partially identified models

Partially identified models have been receiving extensive attentions in recent years, due to their broad applications in econometrics. Partial identification of a structural parameter arises when the data available and the constraints coming from economic theory only allow to place the parameter inside a proper subset of the parameter space. Due to the limitation of the data generating process, the data cannot provide any information within the set where the structural parameter is partially identified (called *identified set*).

This paper aims at developing Bayesian inference for partially identified models. A Bayesian approach may be appealing for several reasons. First, while frequentist approaches cannot tell anything inside the identified set, the Bayesian approach can. When informative (subjective) priors are available for the structural parameter, the shape of the posterior density may be not flat even inside the identified set, providing more information about the parameter that cannot be told by the data. When no a priori information is available, using a uniform prior helps us estimate the true identified set. The Bayesian analysis for partially identified models produces a posterior distribution whose support will asymptotically concentrate around the true identified set. Therefore, the asymptotic behavior for the posterior distribution is different from that of the traditional point identified case, the latter being usually (asymptotically) normally distributed due to the Bernstein-von Mises theorem. Hence, the information from the prior is washed away by the data when the structural parameter is identifiable. A second appealing feature of Bayesian methods arises in situations where we are interested only in a projection of the identified region, that is, a subset of the structural parameter. It turns out that projecting a high dimensional identified region to a low dimensional space using a Bayesian approach is easier than with frequentist approaches because this simply requires the marginalization of a joint distribution. Third, when the model incorporates a multidimensional parameter with some components that are identified and some others that are not, we can learn from the data something also about the non-identified parameters through the information brought by the identified parameters. Moreover, when (asymptotic) equivalence between Bayesian credible sets (BCS) and frequentist confidence sets (FCS) is established, we can take advantage of the fact that BCS are often easier to construct than FCS thanks to the use of Markov Chain Monte Carlo (MCMC) methods. Finally, sometimes frequentist inference relies heavily on point identification, and in some cases achieving the point identification requires stringent assumptions that are hard to verify. In contrast, a Bayesian procedure nevertheless makes inference based on the posterior, whose construction does not require point identification. We illustrate this point further in the following two examples.

Example 1.1 (Functional of nonparametric instrumental regression)**.** In a nonparametric IV regression model $E(y|W) = E(g(X)|W)$ with instrument W (e.g., Hall and Horowitz 2005, Florens and Simoni 2012), suppose we are interested in a functional $h(g)$ of g. The current literature makes inference about $h(g)$ assuming its point identification. However, the identification of $h(g)$ relies on a stringent assumption that is hard to verify (see e.g., Severini and Tripathi 2006 and Santos 2012). Using a Bayesian partial identification approach, in contrast, we can put a prior on $\theta = h(g)$ directly without requiring point identification. The deduced posterior of θ nevertheless enables statistical inference. In particular, if point identification is indeed guaranteed, it can be inferred from the shape of the posterior distribution. □

Example 1.2 (Quantile regression with endogenous censoring)**.** In the model $y = X^T\theta + \epsilon$ and $\mathrm{Med}(\epsilon|X) = 0.5$, only $(I(y < c), \min(y, c), X)$ is observed for some censoring variable c. In particular, the censoring may arbitrarily depend on y and thus is endogenous. Though a sufficient condition for the point identification of θ has been given in the literature (e.g., Khan and Tamer 2009), it is stringent and also hard to verify. In contrast, a Bayesian procedure nevertheless imposes a prior on θ and makes inference via the posterior distribution. On the other hand, the posterior can help to check if point identification is indeed guaranteed. □

There are in general two Bayesian approaches for partially identified models currently developed in the literature. The first one is based on a parametric likelihood function, which is assumed to be known by econometricians up to a finite dimensional parameter. This approach has been used frequently in the literature, see e.g., Moon and Schorfheide (2012), Poirier (1998), Gustafson (2012), Bollinger and Hasselt (2009), Norets and Tang (2012) among many others. However, econometric models usually only identify a set of moment inequalities instead of the full likelihood function. Examples are: interval-censored data, interval instrumental regression, asset pricing (Chernozhukov et al. 2008), incomplete structural models (Menzel 2011) etc. Assuming a parametric form of the likelihood function is therefore ad-hoc. Once the likelihood is mis-specified, the posterior can be misleading. The second approach starts from a set of moment inequalities, and uses a moment-condition-based likelihood such as the limited information likelihood (Kim 2002) and the exponential tilted empirical likelihood (Schennach 2005). Further references may be found in Liao and Jiang (2010), Kitagawa (2012) and Wan (2011). This approach avoids assuming the knowledge of the true likelihood function. However, it does not have a probabilistic interpretation. The use of moment-condition-based likelihoods makes this approach quasi-Bayesian, which uses the Bayesian machinery for inference, see Chernozhukov and Hong (2003).

We propose a pure Bayesian procedure without assuming a parametric form of the true likelihood function. Instead, we place nonparametric priors on the likelihood and obtain the marginal posterior distribution for the partially identified parameter as well as the posterior for the identified set. A similar Bayesian procedure was recently used in Florens and Simoni (2011). As a result, our procedure is semi-parametric Bayesian that involves both finite and infinite dimensional parameters. Our approach thus only requires a set of moment conditions but still possesses a probabilistic interpretation.

Let θ denote the partially identified structural parameter. In addition, we assume that there is a finite dimensional nuisance parameter ϕ that is point identified by the data generating process and that characterizes the identified set. In general, there are two ways in the literature to specify a prior on θ. In the moment-condition-based model, as Kim (2002) and Liao and Jiang (2010), the prior $\pi(\theta)$ is placed marginally, and does not need to take into account the partial identification. Hence $\pi(\theta)$ can be supported on the entire parameter space. In contrast, in the likelihood-based model as considered by Moon and Schorfheide (2012) and Gustafson (2012), the prior $\pi(\theta|\phi)$ is placed conditionally on ϕ, and needs to incorporate the partial identification structure by assuming it is supported only on the identified set, the latter being parametrized by ϕ. We further illustrate this difference in a simple example of interval censoring.

Example 1.3 (Interval censored data)**.** Suppose $Y \in \mathbb{R}$ is censored between Y_1 and Y_2. We are interested in the structural parameter $\theta = EY$, but only Y_1 and Y_2 are observable. Let $\phi = (\phi_1, \phi_2)' = (E(Y_1), E(Y_2))'$, then θ is partially identified on $\Theta(\phi) = [\phi_1, \phi_2]$. The moment-condition-based approach starts from a moment inequality model $E(Y_1 - \theta) \leq 0$ and $E(\theta - Y_2) \leq 0$, and places a prior $\pi(\theta)$ that is supported on the entire parameter space. In contrast, the likelihood-based approach places a prior $\pi(\theta|\phi)$ that is only supported on $[\phi_1, \phi_2]$. Therefore, the latter's prior specification takes into account the fact that Y is censored in $[Y_1, Y_2]$, while the first approach does not need so. $\square$

In this paper we specify a conditional prior $\pi(\theta|\phi)$ for θ given ϕ which incorporates the partial identification structure as in the likelihood-based approach. Examples of such priors include the uniform prior, truncated normal prior, and many priors that have bounded support. Then, the unknown likelihood function is defined for ϕ only, where ϕ is a point-identified nuisance parameter.

We provide a frequentist validation of our procedure. This means that we admit the existence of a true value of the structural parameter and the identified set, and prove that the posterior distribution concentrates asymptotically in a neighborhood of this true value. This property is known as *posterior consistency* and is important because it guarantees that, with a sufficiently large amount of data, we can almost surely recover the truth accurately. Lack of consistency is particularly undesirable and a Bayesian procedure should not be used if the corresponding posterior is inconsistent.

1.2 Highlights of our contributions

We highlight three distinguished features of our approach, which also illustrate our main contributions.

Semi-parametric Bayesian partial identification

We endow the point identified nuisance parameter ϕ with a prior $\pi(\phi)$. The true likelihood function $l_n(\phi)$ is defined on the support of ϕ. Without assuming any parametric form for $l_n(\cdot)$, we place a nonparametric prior $\pi(l_n)$ on the space of probability density (or distribution) functions l_n. The prior specification is completed by a conditional prior $\pi(\theta|\phi)$ which takes into account the partial identification structure. Therefore, the model contains finite dimensional parameters (θ, ϕ) and an infinite dimensional parameter l_n, where (ϕ, l_n) are point-identified nuisance parameters. The marginal posterior of θ is then given by

$$p(\theta|Data) \propto \int \pi(\theta|\phi)\pi(\phi)l_n(\phi)\pi(l_n)d\phi dl_n.$$

Such a semi-parametric posterior requires only a set of moment inequalities, and therefore is a robust (and pure) Bayesian procedure.

In partially identified models, inference may be carried out both for the structural parameter θ and for the identified set. The prior specification $\pi(\theta|\phi)$ on θ plays a role only for inference on θ.

We propose two types of priors on the point-identified parameter (ϕ, l_n). The first consists of a nonparametric prior on the distribution function of the data generating process, with the Dirichlet process prior as an important example. Using this prior, the prior $\pi(\phi)$ of the parameter ϕ can be recovered by viewing ϕ as a function of the distribution function. This prior is appealing when we have no prior information for ϕ. On the contrary, if there is informative prior information for ϕ, it is more convenient to place an alternative semi-parametric prior specified as the product of a prior on ϕ and a prior on the underlying likelihood function l_n. This type of prior on l_n is usually specified on the space of probability density functions, and includes the finite mixture of normals and Dirichlet mixture of normals as examples.

For these prior schemes, we show that asymptotically $p(\theta|Data)$ will be supported within an arbitrarily small neighborhood of the true identified set, which is the notion of *posterior consistency* under partial identification. Moreover, we construct the posterior for the identified set, and show that asymptotically it concentrates within a $\sqrt{\frac{\log n}{n}}$ Hausdorff-neighborhood around the true identified set.

Support function

Our setup is similar to that of Moon and Schorfheide (2012) in that the identified set is completely determined by the identified nuisance parameter ϕ, and hence can be written as $\Theta(\phi)$. Once the posterior of ϕ is determined, so is that of $\Theta(\phi)$. For a definition of the prior and posterior of $\Theta(\phi)$ we refer to Florens and Simoni (2011) who define them in terms of capacity functionals. To make inference on $\Theta(\phi)$ we can take advantage of the fact that when $\Theta(\phi)$ is closed and convex it is completely characterized by its *support function* $S_\phi(\cdot)$ defined as:

$$S_\phi(p) = \sup_{\theta \in \Theta(\phi)} \theta^T p$$

where $p \in \mathbb{S}^{\dim(\theta)}$, the unit sphere. Therefore, inference on $\Theta(\phi)$ may be conveniently carried out through inference on its support function. The posterior distribution of $S_\phi(\cdot)$ is also determined by that of ϕ. We show that in a general moment inequality model, the support function has an asymptotic linear representation in a neighborhood of the true value for ϕ. The posterior of $S_\phi(\cdot)$ is shown to asymptotically concentrate within a $\sqrt{\frac{\log n}{n}}$ sup-norm-neighborhood around the support function of the true identified set. Moreover, we prove the Bernstein-von Mises theorem, that is, the posterior distribution of the support function is shown to be asymptotically normal. We also calculate the support function for a number of interesting examples, including interval censored data, missing data, interval instrumental regression and asset pricing model.

Two-sided Bayesian credible sets for the identified set

We construct two types of Bayesian credible sets (BCS), one for the partially identified parameter θ and the other for the identified set $\Theta(\phi)$. In particular, the BCS for the *identified set* is constructed based on the support function and is two-sided, that is, we find sets $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}}$, where $\hat{\phi}_M$ is any consistent estimator of ϕ (*e.g.* the posterior mode of ϕ, see Section 6 for definitions) such that with probability one, $P(\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}|Data) = 1-\tau$ for credible level $1-\tau$. It is found that the two-sided BCS for the identified set have asymptotically correct coverage probability, in the sense that

$$P_{D_n}(\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}) \geq 1-\tau+o_p(1)$$

where P_{D_n} denotes the sampling probability. Therefore, $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}}$ can also be used as frequentist confidence sets for the identified set. The notation of $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}}$, $\Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$ and q_τ are to be formally defined in Section 6. On the other side, we find that also in the semi-parametric Bayesian model, Moon and Schorfheide (2012)'s conclusion about the BCS for the partially identified parameter θ still holds. Indeed, the BCS for the partially identified parameter tends to be smaller than frequentist confidence sets in large samples.

Note that we consider a fixed data generating process (DGP). The constructed BCS has asymptotically correct coverage probability for any specific DGP, and the uniformity issue as in Andrews and Soares (2010) is not considered. In addition, all the results on the identified set, support function and posterior consistency for θ are valid even when point identification is actually achieved, that is, when $\Theta(\phi)$ is a singleton.

1.3 Literature review

There is a growing literature on Bayesian partially identified models. Besides those mentioned above, the list also includes Gelfand and Sahu (1999), Neath and Samaniego (1997), Epstein and Seo (2011), Stoye (2012), Kline (2011), etc. There is also an extensive literature from a frequentist point of view. A partial list includes Andrews and Guggenberger (2009), Andrews and Soares (2010), Beresteanu, Molchanov and Molinari (2010), Bugni (2010), Canay (2010), Chernozhukov, Hong and Tamer (2007), Chiburis (2009), Imbens and Manski (2004), Romano and Shaikh (2010), Rosen (2008), Stoye (2010), among others. See Tamer (2010) for a review.

When the identified set is closed and convex, the support function becomes one of the useful tools to characterize its properties. Therefore the support function has been recently introduced to study partially identified models, and the literature on this perspective has been growing rapidly, see for example, Bontemps, Magnac and Maurin (2012), Beresteanu and Molinari (2008), Beresteanu et al. (2012), Kaido and Santos (2012), Kaido (2012) and Chandrasekhar et al. (2012).

1.4 Organization

The paper is organized as follows. Section 2 sets up the model and proposes two types of prior specification on the underlying likelihood function. Section 3 achieves the posterior consistency for the (marginal) posterior distribution of the structural parameter. Section 4 derives the posterior consistency for the identified set and provide the concentration rate. Section 5 studies the posterior of the support function in moment inequality models. In particular, the Bernstein von Mises theorem for the support function is proved. Section 6 constructs the Bayesian credible sets for both the structural parameter and its identified set and looks in detail at the missing data example. Section 7 discusses the case when point identification is actually achieved. In this case, all the derived results on the identified set and the support function are still valid. Section 8 applies the support function approach to a financial asset pricing study. Finally, Section 9 concludes with further discussions. All the proofs are given in the appendix.

2 General Setup of Bayesian Partially Identified Model

2.1 The Model

Econometric models often involve a finite dimensional structural parameter θ. In many cases such a structural parameter is only partially identified by the data generating process on a non-singleton set, which we call *identified set*. The goal of an econometrician is to make inference on the partially identified parameter as well as the identified set based on the data.

Along with θ, the model also includes a finite dimensional nuisance parameter $\phi \in \Phi$ that is point identified by the data generating process. Here Φ denotes the parameter space for ϕ. The point identified parameter often arises naturally as it characterizes the data distribution. In most of partially identified models, the identified set is also characterized by ϕ, hence we denote it by $\Theta(\phi)$ to indicate that once ϕ is determined, so is the identified set. Let Θ denote the parameter space for θ; we assume $\Theta(\phi) \subseteq \Theta$.

We put a prior on (θ, ϕ), which induces a prior on the identified set $\Theta(\phi)$ via ϕ. Due to the identification feature, for any given $\phi \in \Phi$, the conditional prior $\pi(\theta|\phi)$ is specified such that

$$\pi(\theta \in \Theta(\phi)|\phi) = 1.$$

Our analysis focuses on the situation where $\Theta(\phi)$ is a closed and convex set for each ϕ. Therefore $\Theta(\phi)$ can be uniquely characterized by its *support function*. Let $d = \dim(\theta)$. For any fixed ϕ, the support function for $\Theta(\phi)$ is a function $S_\phi(\cdot) : \mathbb{S}^d \to \mathbb{R}$ such that

$$S_\phi(p) = \sup_{\theta \in \Theta(\phi)} \theta^T p.$$

where $\mathbb{S}^d$ denotes the unit sphere in $\mathbb{R}^d$. The support function plays a central role in convex analysis since it determines all the characteristics of a convex set. For example, if $\theta \in \Theta(\phi)$, then its kth component has bounds $\theta_k \in [-S_\phi(-e_k), S_\phi(e_k)]$, where e_k is the kth standard basis vector (a vector of all zeros, except for a one in the kth position). Also, $\theta \in \Theta(\phi)$ if and only if $p^T\theta \leq S_\phi(p)$ for all $p \in \mathbb{S}^d$.

Characterization and frequentist estimation of the identified set through its support function has been previously proposed by Bontemps *et al.* (2011) and Beresteanu *et al.* (2012) and also used by Kaido and Santos (2011) among others. It is also one of the essential objects for our Bayesian inference. At the best of our knowledge a Bayesian estimation of the support function of the identified set has not been proposed in the literature so far.

Similar to $\Theta(\phi)$, we put a prior on $S_\phi(\cdot)$ via the prior on ϕ. In this paper we investigate the asymptotic frequentist properties of the posterior distribution of the support function, as well as those of θ and of $\Theta(\phi)$, including the posterior concentration rates and the Bernstein von Mises theorem as in Bickel and Kleijn (2012). In addition, we carry out Bayesian inferences by constructing two-sided Bayesian credible sets for the identified set $\Theta(\phi)$ based on the support function.

Before formalizing our Bayesian setup, let us present a few examples that have received much attention in partially identified econometric models literature.

Example 2.1 (Interval censored data - *continued*)**.** Let (Y, Y_1, Y_2) be a 3-dimensional random vector such that $Y \in [Y_1, Y_2]$ with probability 1. The random variables Y_1 and Y_2 are observed while Y is unobservable (see, e.g., Moon and Schorfheide 2012). We denote: $\theta = E(Y)$ and $\phi = (\phi_1, \phi_2)' := (E(Y_1), E(Y_2))'$. Therefore, we have the following identified set for θ: $\Theta(\phi) = [\phi_1, \phi_2]$. The support function for $\Theta(\phi)$ is easy to derive:

$$S_\phi(1) = \phi_2, \quad S_\phi(-1) = -\phi_1.$$

□

Example 2.2 (Interval regression model)**.** The regression model with interval censoring has been studied by, for example, Haile and Tamer (2003), etc. Let (Y, Y_1, Y_2) be a 3-dimensional random vector such that $Y \in [Y_1, Y_2]$ with probability 1. The random variables Y_1 and Y_2 are observed while Y is unobservable. Assume that

$$Y = X^T\theta + \epsilon$$

where X is a vector of observable regressors. In addition, assume there is a d-dimensional vector of nonnegative exogenous variables Z such that $E(Z\epsilon) = 0$. Here Z can be either a vector of instrumental variables when X is endogenous, or a nonnegative transformation of X when X is exogenous. It follows that

$$E(ZY_1) \leq E(ZY) = E(ZX^T)\theta \leq E(ZY_2). \tag{2.1}$$

We denote $\phi = (\phi_1, \phi_2, \phi_3)$ with $(\phi_1^T, \phi_3^T) = (E(ZY_1)^T, E(ZY_2)^T)$ and $\phi_2 = E(ZX^T)$. Then the identified set for θ is given by $\Theta(\phi) = \{\theta \in \Theta : \phi_1 \leq \phi_2\theta \leq \phi_3\}$. Suppose ϕ_2^{-1} exists. The support function for $\Theta(\phi)$ is given by (for $\text{sgn}(x) = I(x > 0) - I(x < 0)$)[1]:

$$S_\phi(p) = p^T\phi_2^{-1}\left(\frac{\phi_1 + \phi_3}{2}\right) + \alpha_p^T\left(\frac{\phi_3 - \phi_1}{2}\right), \quad p \in \mathbb{S}^d$$

where $\alpha_p = ((p^T\phi_2^{-1})_1\text{sgn}(p^T\phi_2^{-1})_1, ..., (p^T\phi_2^{-1})_d\text{sgn}(p^T\phi_2^{-1})_d)^T$.

□

[1]See Appendix A for detailed derivations of the support function in this example. Similar results but in a slightly different form are presented in Bontemps et al. (2012).

Example 2.3 (missing data)**.** Consider a bivariate random vector (Y, M) where M is a binary random variable which takes the value $M = 1$ when Y is missing and 0 otherwise. The parameter of interest is the marginal distribution F_Y of Y: $\theta = F_Y(y)$. This problem without the missing-at-random assumption has been extensively studied in the literature, see for example, Manski and Tamer (2002), Manski (2003), etc. Let F and F_M denote respectively the joint distribution of (Y, M) and marginal distribution of M. Moreover, $F_{Y|M}$ denotes the conditional distribution of Y given M. By the Law of Total Probability: $\theta = F_{Y|M}(y|M = 0)F_M(M = 0) + F_{Y|M}(y|M = 1)F_M(M = 1)$. Since $F_{Y|M}(y|M = 1)$ cannot be recovered from the data then the empirical evidence partially identifies θ and θ is characterized by the following moment restrictions:

$$F(y, M = 0) \leq \theta \leq F(y, M = 0) + F_M(M = 1).$$

Here $\phi = (F(y, M = 0), F_M(M = 1)) = (\phi_1, \phi_2)$. The identified set is $\Theta(\phi) = [\phi_1, \phi_1 + \phi_2]$. and it support function is: $S_\phi(1) = \phi_1 + \phi_2, \quad S_\phi(-1) = -\phi_1$. □

2.2 Semi-parametric Bayesian Setup

Let F denote the distribution function for the observed data, which is point identified by the data generating process. In a parametric Bayesian partially identified model as in Poirier (1998), Gustafson (2005, 2012) and Moon and Schorfheide (2012), F is linked with a known likelihood function for ϕ. Therefore the model is parametric and one does not put priors on F. For example, in the interval censored data Example 2.1, if we know that Y_1 and Y_2 are jointly normal with mean (ϕ_1, ϕ_2) and a known covariance matrix Σ. Then the likelihood function is given by

$$l(\phi) = (2\pi \det(\Sigma))^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_{1i} - \phi_1, Y_{2i} - \phi_2)\Sigma^{-1}(Y_{1i} - \phi_1, Y_{2i} - \phi_2)^T\right)$$

where $\{(Y_{1i}, Y_{2i})\}_{i=1}^n$ is a set of *i.i.d.* realizations of (Y_1, Y_2). Then F is the cdf of a bivariate normal distribution with mean vector ϕ and covariance Σ. The standard Bayesian approach for a partially identified model (parametric) proceeds by specifying a joint prior distribution $\pi(\theta, \phi)$ and obtains the marginal posterior for θ:

$$p(\theta|\{(Y_{1i}, Y_{2i})\}_{i=1}^n) \propto \int_\Phi \pi(\theta, \phi)l(\phi)d\phi.$$

However, like for usual point identified models, assuming a known likelihood function may suffer from a model specification problem, and may lead to very misleading results. Instead, econometric applications often involve only a set of moment conditions as (2.1). This gives rise to the so-called *moment inequality models*, e.g., Chernozhukov, Hong and Tamer (2007), Bugni (2010), Liao and Jiang (2010), Andrews and Soares (2010), Kaido and Santos (2011), and many other references therein. A parametric form of the likelihood function and of F is unavailable in these models, and ad-hoc assumptions that make the model parametric can result to severe misleading conclusions.

A much more robust approach is to proceed without assuming a parametric form for the likelihood function, but put a prior on F instead. This yields to the semi-parametric Bayesian setup. The statistical model therefore contains three parameters: the structural parameter of interest θ, a finite dimensional nuisance parameter ϕ which can be point identified by the DGP, and a nuisance infinite dimensional parameter F which characterizes the distribution of the data.

A further distinction among the parameters may be done on the basis of identification: the *identified parameters* (ϕ, F), which characterize the sampling distribution of the observable random variables, and the *partially identified parameter* θ, which is linked to the sampling distribution through ϕ. We have to take this difference into account when we construct the prior distribution for the model parameters. Therefore, the prior distribution is naturally decomposed into a marginal prior for the identified parameter and a conditional prior for θ given the identified parameter such that

$$\pi(\theta \in \Theta(\phi)|\phi) = 1.$$

We specify a conditional prior distribution for θ given ϕ taking the form

$$\pi(\theta|\phi) \propto I_{\theta\in\Theta(\phi)}g(\theta)$$

where $g(\cdot)$ is some probability density function with respect to the Lebesgue measure and $I_{\theta\in\Theta(\phi)}$ is the indicator function of $\Theta(\phi)$ which takes the value 1 if $\theta \in \Theta(\phi)$. By construction this prior puts all its mass on $\Theta(\phi)$, $\forall\phi \in \Phi$.

Below we describe two possible ways to specify the prior on (ϕ, F): a fully nonparametric prior and a semi-parametric prior. The first prior scheme consists in placing a fully nonparametric prior on F which induces a prior on ϕ through a transformation $\phi = \phi(F)$. When there is more informative prior information for ϕ directly, it is more convenient to place a prior on (ϕ, η) where η is an infinity-dimensional nuisance parameter (often a density function) that is independent of ϕ a priori and that characterizes F. The prior on (ϕ, F) is then deduced from the prior on (ϕ, η).

Below we formally define these two prior specifications. An illustrative example is given in Section 2.3. Let X denote the observable random variable for which we have n i.i.d. observations $D_n = \{X_i\}_{i=1}^n$. Let $(\mathcal{X}, \mathfrak{B}_x, F)$ denote a probability space in which X takes values. Let $\mathcal{F}$ denote the set of probability measures on $(\mathcal{X}, \mathfrak{B}_X)$, which is also the parameter space of F.

2.2.1 Nonparametric prior

Since ϕ is point identified, we assume it can be rewritten as a measurable function of F as $\phi = \phi(F)$, for instance $\phi = E(X) = \int xF(dx)$. A possible way to construct the prior distribution consists of specifying a nonparametric prior distribution for F and then deduce from it the prior distribution for ϕ via $\phi(F)$. The Bayesian experiment is

$$X|F \sim F, \qquad F \sim \pi(F), \qquad \theta|\phi = \phi(F) \sim \pi(\theta|\phi(F))$$

The prior distribution $\pi(F)$ is a distribution on $\mathcal{F}$. Examples of such a prior include

Dirichlet process priors (Ferguson (1973)) and Polya tree (Lavine (1992)). The case where $\pi(F)$ is a Dirichlet process prior in partially identified models has been proposed by Florens and Simoni (2011).

Conditionally on ϕ, the data are completely uninformative about θ: the prior distribution of θ is revised by the data only through the information brought by the identified parameter $\phi(F)$. Indeed, since $\phi(F)$ is identified, it is straightforward to show that the posterior of θ conditional on $\phi(F)$ satisfies

$$p(\theta|\phi(F), D_n) = \pi(\theta|\phi(F)).$$

(see Poirier (1998), who called the data to be conditionally uninformative for θ given ϕ). Let $p(F|D_n)$ denote the marginal posterior of F which, by abuse of notation, can be written $p(F|D_n) \propto \pi(F)\prod_{i=1}^n F(X_i)$. The posterior distribution of ϕ, $\Theta(\phi)$, $S_\phi(\cdot)$ are deduced from the posterior of F. Then, for any measurable set $B \subset \Theta$, the marginal posterior probability of θ is given by, averaging over F:

$$\begin{aligned} P(\theta \in B|D_n) &= \int_{\mathcal{F}} P(\theta \in B|\phi(F), D_n)p(F|D_n)dF \\ &= \int_{\mathcal{F}} \pi(\theta \in B|\phi(F))p(F|D_n)dF = E[\pi(\theta \in B|\phi(F))|D_n] \end{aligned} \quad (2.2)$$

where the conditional expectation is taken with respect to the posterior of F. The corresponding marginal posterior density function of θ will be denoted by $p(\theta|D_n)$.

2.2.2 Semi-parametric prior

Alternatively, instead of modeling F nonparametrically, we could reformulate the model and parameterize the sampling distribution F in terms of a finite-dimensional parameter $\phi \in \Phi$ and a nuisance parameter $\eta \in \mathcal{P}$, where $\mathcal{P}$ is an infinite-dimensional measurable space. Therefore, $\mathcal{F} = \{F_{\phi,\eta}; \phi \in \Phi, \eta \in \mathcal{P}\}$. When we consider the frequentist properties of the posterior distribution, we assume there is a fixed true value for F, denoted by F_0. Since both ϕ and F are identified then there exist unique $\phi_0 \in \Phi$ and $\eta_0 \in \mathcal{P}$ such that $F_0 = F_{\phi_0,\eta_0}$, where ϕ_0 and η_0 denote the true values of ϕ and η. Denote by $l_n(\phi, \eta)$ the model's likelihood function.

One of the appealing features of this semi-parametric approach is that it allows us to impose a prior $\pi(\phi)$ directly on the identified parameter ϕ, which is convenient whenever we have good prior information regarding ϕ. In contrast, a nonparametric prior specification may be inconvenient to incorporate subjective prior information.

For instance, in the interval censored data example, we can write

$$Y_1 = \phi_1 + u, \quad Y_2 = \phi_2 + v$$

$$u \sim f_1, \quad v \sim f_2,$$

where both u and v are random errors with zero mean and unknown density functions f_1 and f_2 such that $u \perp\!\!\!\perp v | f_1, f_2$ and the supports of the corresponding distributions of

$Y_1, Y_2|\phi_1, \phi_2, f_1, f_2$ are disjoint[2]. Then $\eta = (f_1, f_2)$, and the likelihood function is

$$l_n(\phi, \eta) = \prod_{i=1}^{n} f_1(Y_{1i} - \phi_1) f_2(Y_{2i} - \phi_2).$$

We put priors on (ϕ, f_1, f_2). This is a location-model studied for instance by Ghosal et al. (1999) and Amewou-Atisso et al. (2003). Examples of priors on density functions f_1 and f_2 include mixture of Dirichlet process priors, Gaussian process priors, etc.

The joint prior distribution $\pi(\theta, \phi, \eta)$ is naturally decomposed as

$$\pi(\theta, \phi, \eta) = \pi(\theta|\phi) \times \pi(\phi, \eta). \tag{2.3}$$

We place an independent prior on (ϕ, η) as $\pi(\phi, \eta) = \pi(\phi)\pi(\eta)$. Therefore, the Bayesian experiment is

$$X|\phi, \eta \sim F_{\phi,\eta}, \qquad (\phi, \eta) \sim \pi(\phi, \eta) = \pi(\phi) \times \pi(\eta), \qquad \theta|\phi, \eta \sim \pi(\theta|\phi).$$

The posterior distribution of ϕ has a density function given by

$$p(\phi|D_n) \propto \int_{\mathcal{P}} \pi(\phi, \eta) l_n(\phi, \eta) d\eta. \tag{2.4}$$

Then the marginal posterior of θ is, for any measurable set $B \in \Theta$:

$$P(\theta \in B|D_n) \propto \int_{\Phi} \int_{\mathcal{P}} \pi(\theta \in B|\phi) \pi(\phi, \eta) l_n(\phi, \eta) d\eta d\phi. \tag{2.5}$$

Moreover, the corresponding posterior density function is: $p(\theta|D_n) = \int_{\Phi} \pi(\theta|\phi) p(\phi|D_n) d\phi$. where $p(\phi|D_n)$ has been defined in (2.4). The posteriors of $\Theta(\phi)$ and $S_\phi(\cdot)$ are deduced from that of ϕ. Suppose for example we are interested in whether $\Theta(\phi) \cap A$ is an empty set for some $A \subset \Theta$, we then look at the posterior probability

$$P(\Theta(\phi) \cap A|D_n) = \frac{\int_{\{\phi: \Theta(\phi) \cap A = \emptyset\}} \int_{\mathcal{P}} \pi(\phi)\pi(\eta) l_n(\phi, \eta) d\eta d\phi}{\int_{\Phi} \int_{\mathcal{P}} \pi(\phi)\pi(\eta) l_n(\phi, \eta) d\eta d\phi}.$$

The finite-dimensional posterior distribution of the support function $S_\phi(\cdot)$ is the distribution $P(S_\phi(p_i) \in A_i, \text{ for } 1 \leq i \leq k)$, $k \in \mathbb{N}$, for every $(p_1, \ldots, p_k)$ such that $p_i \in \mathbb{S}^d$, $i = 1, \ldots, k$, and for every product of measurable sets A_i in $\mathbb{R}$.

Example 2.4 (Interval regression model - *continued*)**.** Consider Example 2.2, where $\phi =$

[2] In order to implement this we have two possibilities. Let $[\underline{u}, \overline{u}]$ and $[\underline{v}, \overline{v}]$ denote the supports of f_1 and f_2 and $[\underline{\phi}_1, \overline{\phi}_1]$ and $[\underline{\phi}_2, \overline{\phi}_2]$ denote the supports of ϕ_1 and ϕ_2, respectively. First, we can specify a conditional prior $\pi(f_1, f_2|\phi_1, \phi_2)$ such that $\overline{u} + \overline{\phi}_1 \leq \underline{v} + \underline{\phi}_2$. A second possibilities is to specify a independent prior on (f_1, f_2) and on (ϕ_1, ϕ_2) such that $\overline{u} \leq \underline{v}$ and $\overline{\phi}_1 \leq \underline{\phi}_2$.

$(\phi_1, \phi_2, \phi_3) = (E(ZY_1), E(ZX^T), E(ZY_2))$. Write

$$ZY_1 = \phi_1 + u_1, \quad ZY_2 = \phi_3 + u_3, \quad \text{vec}(ZX^T) = \text{vec}(\phi_2) + u_2,$$

where u_1, u_2 and u_3 are correlated and their joint unknown probability density function is $\eta(u_1, u_2, u_3)$. The likelihood function is then

$$l_n(\phi, \eta) = \prod_{i=1}^{n} \eta(Z_i Y_{1i} - \phi_1, Z_i Y_{2i} - \phi_3, \text{vec}(Z_i X_i^T) - \text{vec}(\phi_2)).$$

□

Many nonparametric priors can be used for $\pi(\eta)$ in the location-model of the type of Example 2.4, where $l_n(\phi, \eta) = \prod_{i=1}^{n} \eta(X_i - \phi)$, or of the type of the interval data example. The next examples show three possible ways for constructing priors $\pi(\eta)$ on probability density functions.

Example 2.5. The finite mixture of normals (e.g., Lindsay and Basak (1993), Ray and Lindsay (2005)) assumes η to take the form

$$\eta(x) = \sum_{i=1}^{k} w_i h(x; \mu_i, \Sigma_i)$$

where $h(x; \mu_i, \Sigma_i)$ is the density of a multivariate normal distribution with mean μ_i and variance Σ_i and $\{w_i\}_{i=1}^{k}$ are unknown weights such that $\sum_{i=1}^{k} w_i \mu_i = 0$. Then $\int \eta(x) x dx = \sum_{i=1}^{k} w_i \int h(x; \mu_i, \Sigma_i) x dx = 0$. We impose prior $\pi(\eta) = \pi(\{\mu_l, \Sigma_l, w_l\}_{l=1}^{k})$, then

$$\begin{aligned} p(\phi|D_n) &\propto \int_{\mathcal{P}} \pi(\phi)\pi(\eta) \prod_{i=1}^{n} \eta(X_i - \phi) d\eta \\ &= \int \pi(\phi) \prod_{i=1}^{n} \sum_{j=1}^{k} w_j h(X_i - \phi; \mu_j, \Sigma_j) \pi(\{\mu_l, \Sigma_l, w_l\}_{l=1}^{k}) dw_j d\mu_j d\Sigma_j. \end{aligned}$$

□

Example 2.6. Dirichlet mixture of normals (e.g., Ghosal, Ghosh and Ramamoorthi (1999) Ghosal and van der Vaart (2001), Amewou-Atisso, et al. (2003)) assumes

$$\eta(x) = \int h(x - z; 0, \Sigma) dH(z)$$

where $h(x; 0, \Sigma)$ is the density of a multivariate normal distribution with mean zero and variance Σ and H is a probability distribution such that $\int z H(z) dz = 0$. Then $\int x \eta(x) dx = 0$. To place a prior on η, we let H have the Dirichlet process prior distribution $D_\alpha \equiv \mathcal{D}(\nu_0, Q_0)$ where α is a finite positive measure, $\nu_0 = \alpha(\mathcal{X}) \in \mathbb{R}_+$ and $Q_0 = \alpha/\alpha(\mathcal{X})$ is a base probability on $(\mathcal{X}, \mathfrak{B}_x)$ such that $Q_0(x) = 0$, $\forall x \in (\mathcal{X}, \mathfrak{B}_x)$. In addition, we place a prior on

Σ independent of the prior on H. Then

$$p(\phi|D_n) \propto \int \pi(\phi)\pi(\Sigma)D_\alpha(H)\prod_{i=1}^{n}\int h(X_i-\phi-z;0,\Sigma)dH(z)d\Sigma dH.$$

□

Example 2.7. Random Bernstein polynomials (e.g., Walker et al. (2007) and Ghosal (2001)) admits a density function

$$\eta(x)=\sum_{j=1}^{k}[H(j/k)-H((j-1)/k)]\mathcal{B}e(x;j,k-j+1),$$

where $\mathcal{B}e(x;a,b)$ stands for the beta density with parameters $a,b>0$ and H is a random distribution function with prior distribution assumed to be a Dirichlet process. Moreover, the parameter k is also random with a prior distribution independent of the prior on H. Then $p(\phi|D_n)\propto\int\pi(\phi)\prod_{i=1}^{n}\eta(X_i-\phi)\pi(H)\pi(k)dHdk$. □

Besides, other commonly used priors are wavelet expansions (Rivoirard and Rousseau (2012)), Polya tree priors (Lavine (1992)), Gaussian process priors (van der Vaart and van Zanten (2008), Castillo (2008)), etc.

2.3 Interval censored data: an example

For illustration purposes, we consider a simple version of the interval censored data example 2.1 where $Y_2=Y_1+1$ and only Y_1 is observable, *i.e.* $Y_1\equiv X$ in our general notation. Let $\phi=EY_1$ and $\theta=EY$, then the identified set is $\Theta(\phi)=[\phi,\phi+1]$. Let F denote the marginal distribution of Y_1. Then a more formal way to write ϕ should be $\phi=\phi(F)=E(Y_1|F)$.

Let us specify a Dirichlet process prior for F: $\pi(F)=\mathcal{D}ir(\nu_0,Q_0)$, where $\nu_0\in\mathbb{R}_+$ and Q_0 is a base probability on $(\mathcal{X},\mathfrak{B}_x)$ such that $Q_0(x)=0$, $\forall x\in(\mathcal{X},\mathfrak{B}_x)$. By using the stick-breaking representation (see Sethuraman (1994)), the deduced prior distribution of the transformation $\phi(F)$ is

$$\pi(\phi\in A)=P\left(\sum_{j=1}^{\infty}\alpha_j\xi_j\in A\right),\quad\forall A\subset\Phi$$

where $\{\xi_j\}_{j\geq1}$ denote independent drawings from Q_0, $\alpha_j=v_j\prod_{l=1}^{j}(1-v_l)$ with $\{v_l\}_{l\geq1}$ independent drawings from a Beta distribution $\mathcal{B}e(1,\nu_0)$ and $\{v_l\}_{l\geq1}$ are independent of $\{\xi_j\}_{j\geq1}$. The posterior distribution of F is still a Dirichlet process: $p(F|D_n)=\mathcal{D}ir(\nu_n,Q_n)$, with $\nu_n=\nu_0+n$, and $Q_n=\frac{\nu_0}{\nu_0+n}Q_0+\frac{n}{\nu_0+n}\hat{F}$, where $\hat{F}$ is the empirical distribution of the sample $(Y_{11},\ldots,Y_{1n})$. The posterior distribution of the transformation $\phi(F)$ is

$$P(\phi\in A|D_n)=P\left(\rho\sum_{j=1}^{n}\beta_jY_{1j}+(1-\rho)\sum_{j=1}^{\infty}\alpha_j\xi_j\in A\middle|D_n\right),\quad\forall A\subset\Phi,$$

where ρ is drawn from a Beta distribution $\mathcal{B}e(n, \nu_0)$ independently of the other quantities and $(\beta_1, \ldots, \beta_n)$ are drawn from a Dirichlet distribution with parameters $(1, \ldots, 1)$ on the simplex S_{n-1} of dimension $(n-1)$. With the prior $\pi(\theta|\phi) \propto I_{\theta \in \Theta(\phi)} g(\theta)$, the marginal posterior density function of θ evaluated at some fixed $\tilde{\theta}$ is

$$
\begin{aligned}
p\left(\tilde{\theta}|D_n\right) &\propto \int_{\mathcal{F}} I\left(\tilde{\theta} \in [\phi, \phi+1]\right) g\left(\tilde{\theta}\right) p(F|D_n) dF \\
&= g\left(\tilde{\theta}\right) P\left(\phi(F) \leq \tilde{\theta} \leq \phi(F) + 1 \middle| D_n\right) \\
&= g\left(\tilde{\theta}\right) P\left(\theta - 1 \leq \rho \sum_{j=1}^{n} \beta_j Y_{1j} + (1-\rho) \sum_{j=1}^{\infty} \alpha_j \xi_j \leq \theta \middle| D_n\right).
\end{aligned}
$$

The alternative semi-parametric prior can be formulated as follows. Define $u = Y_1 - \phi_1$, and assume u has a continuous density f. The likelihood is thus given by $l_n(\phi, f) = \prod_{i=1}^n f(Y_i - \phi)$. We place any parametric prior $\pi(\phi)$ on ϕ and a Dirichlet mixture of normals prior on f, which assumes $f(u) = \int h(u - z; 0, \sigma^2) dH(z)$ where H is a probability measure that has a Dirichlet process prior D_α and σ^2 is a variance parameter for the normal mixtures that has an inverse Gamma prior (see Example 2.6 for details). We then obtain the posterior

$$
p(\phi|D_n) \propto \int \pi(\phi)\pi(\sigma^2) D_\alpha(H) \prod_{i=1}^{n} \int h(Y_i - \phi - z; 0, \sigma^2) dH(z) d\sigma^2 dH.
$$

The marginal posterior density function of θ evaluated at some fixed $\tilde{\theta}$ is

$$
p(\tilde{\theta}|D_n) \propto g(\tilde{\theta}) P(\tilde{\theta} - 1 < \phi < \tilde{\theta}|D_n).
$$

3 Posterior Consistency for θ

In Bayesian analysis, one starts with a prior knowledge (sometimes uninformative) on the parameter and updates it according to the marginal posterior given the data. In classical point identified parametric and semi-parametric models, under mild assumptions the posterior is asymptotically normal due to the Bernstein von Mises theorem and hence its shape is not affected anymore asymptotically by the prior specification. In contrast, the shape of the posterior of a partially identified parameter still relies upon its prior distribution (see Poirier (1998)) even asymptotically. Only the support of the prior distribution of θ (given ϕ) is revised after data are observed and eventually converges towards the true identified set asymptotically. This corresponds to frequentist consistency of the posterior for partially identified parameters and is due to the fact that the point-identified parameter ϕ completely characterizes the support.

We assume there is a true value of ϕ, denoted by ϕ_0, which induces a true identified set $\Theta(\phi_0)$ and a true F, denoted by F_0. Our goal is to achieve the frequentist *posterior consistency* for the partially identified parameter: that is, for any $\epsilon > 0$ there exists a $\tau \in (0, 1]$ such that

$$P(\theta \in \Theta(\phi_0)^{\epsilon}|D_n) \to^p 1 \quad \text{and} \quad P(\theta \in \Theta(\phi_0)^{-\epsilon}|D_n) \to^p (1-\tau)$$

where

$$\Theta(\phi)^{\epsilon} = \{\theta \in \Theta : d(\theta, \Theta(\phi)) \le \epsilon\} \tag{3.1}$$

is the ϵ-envelope of $\Theta(\phi)$ and

$$\Theta(\phi)^{-\epsilon} = \{\theta \in \Theta(\phi) : d(\theta, \Theta \backslash \Theta(\phi)) \ge \epsilon\} \tag{3.2}$$

is the ϵ-contraction of $\Theta(\phi)$ with $\Theta\backslash\Theta(\phi) = \{\theta \in \Theta; \theta \notin \Theta(\phi)\}$ and $d(\theta, \Theta(\phi)) = \inf_{x\in\Theta(\phi)} \|\theta - x\|$, see *e.g.* Molchanov (2005) and Chernozhukov, Hong and Tamer (2007). Thus, *posterior* (or *frequentist*) *consistency* for a partially identified parameter means that the posterior distribution of θ puts all its mass on a set whose boundaries belongs to the set $\{\theta \in \Theta; d(\theta; \partial\Theta(\phi_0)) \le \epsilon\}$ where $\partial\Theta(\phi_0)$ denotes the boundary of $\Theta(\phi_0)$. Posterior consistency is one of the benchmarks of a Bayesian procedure under consideration, which ensures that with a sufficiently large amount of data, it is nearly possible to discover the truth identified set. Therefore lack of consistency is extremely undesirable. Liao and Jiang (2010, 2011) studies the posterior consistency for partially identified models, however, with a pseudo likelihood function whose probabilistic interpretation is still in question[3]. More recently, Kitagawa (2011) considered the posterior consistency for $\Theta(\phi)$ in terms of the *posterior lower probability* when the parametric form of the likelihood is known.

We recall that the conditional prior on θ (given ϕ) is specified as

$$\pi(\theta|\phi) \propto g(\theta) I_{\theta\in\Theta(\phi)} \tag{3.3}$$

for some $g(\theta)$. In the special case where θ is point identified, then $\{\theta\} = \Theta(\phi)$ becomes a function of ϕ, whose prior is completely determined by that of ϕ instead of by (3.3).

In this section we focus on the frequentist consistency of the marginal posterior of θ (marginalized with respect to the posterior of ϕ). We will investigate the posterior concentration rate of $\Theta(\phi)$ and $S_{\phi}(\cdot)$ in subsequent sections. For a measurable set $B \subset \Theta$, the marginal posterior probability is given by (2.2):

$$p(\theta \in B|D_n) = \int_{\mathcal{F}} \pi(\theta \in B|\phi(F)) p(F|D_n) dF$$

when the prior on ϕ is induced by the nonparametric prior specified on F, and by (2.5):

$$p(\theta \in B|D_n) = \int_{\Phi} \pi(\theta \in B|\phi) p(\phi|D_n) d\phi$$

when the prior on ϕ is specified through a semi-parametric prior as described in section 2.2.2. Recall that F and ϕ are point-identified and frequentist asymptotic properties of the marginal posterior of θ rely on frequentist asymptotic properties of the posterior of F and ϕ. Therefore, we assume that the priors $\pi(F)$ and $\pi(\phi)$ specified for F and ϕ are such that the corresponding posterior distributions are consistent:

[3]See Schennach (2005) for discussions of probabilistic interpretations of pseudo likelihood functions.

Assumption 3.1. *At least one of the following holds:*

(i). The measurable function $\phi : \mathcal{F} \to \Phi$ is continuous and the prior $\pi(F)$ is such that the posterior $p(F|D_n)$ satisfies:

$$\int_{\mathcal{F}} m(F)p(F|D_n)dF \to^p \int_{\mathcal{F}} m(F)\delta_{F_0}(dF)$$

for any bounded and continuous function $m(\cdot)$ on $\mathcal{F}$ where δ is the Dirac function, and F_0 is the true distribution function of X;

(ii). the prior $\pi(\phi)$ is such that the posterior $p(\phi|D_n)$ satisfies:

$$\int_{\Phi} m(\phi)p(\phi|D_n)d\phi \to^p \int_{\Phi} m(\phi)\delta_{\phi_0}(d\phi)$$

for any bounded and continuous function $m(\cdot)$ on Φ.

Assumptions 3.1 *(i)* and *(ii)* correspond to the nonparametric and semi-parametric prior, respectively and are verified by many nonparametric and semi-parametric priors. Examples are: Dirichlet process priors, Polya Tree process priors, Gaussian process priors, etc. We refer to Ghosh and Ramamoorthi (2003) for examples and sufficient conditions for posterior consistency. For instance, when $\pi(F)$ is the Dirichlet process prior, the second part of Assumption 3.1 *(i)* was proved in Ghosh and Ramamoorthi (2003, Theorem 3.2.7). The condition that $\phi(F)$ is continuous in F is verified in many examples relevant for applications. For instance, in example 2.1, $\phi(F) = E(Y|F)$ and in example 2.2, $\phi(F) = (E(ZY_1|F), E(ZX^T|F), E(ZY_2|F))$, which are all linear functionals of F.

Assumption 3.2. *For any $\epsilon > 0$ there are measurable sets $A_1, A_2 \subset \Phi$ such that $0 < \pi(\phi \in A_i) \leq 1$, $i = 1, 2$ and*
(i) for all $\phi \in A_1$, $\Theta(\phi_0)^\epsilon \cap \Theta(\phi) \neq \emptyset$; for all $\phi \notin A_1$, $\Theta(\phi_0)^\epsilon \cap \Theta(\phi) = \emptyset$,
(ii) for all $\phi \in A_2$, $\Theta(\phi_0)^{-\epsilon} \cap \Theta(\phi) \neq \emptyset$; for all $\phi \notin A_2$, $\Theta(\phi_0)^{-\epsilon} \cap \Theta(\phi) = \emptyset$.

This assumption allows us to prove the posterior consistency without assuming the prior $\pi(\theta|\phi)$ to be a continuous function of ϕ, and therefore priors like $I_{\phi_1<\theta<\phi_2}$ in the interval censoring data example are allowed. Under this assumption and if the conditional prior $\pi(\theta|\phi)$ is a regular conditional distribution, the conditional prior probability of the ϵ-envelope (and of the ϵ-contraction) of the identified set can be approximated by a continuous function, i.e., there is a sequence of bounded and continuous functions $h_m(\phi)$ such that (see lemma C.1 in the appendix) almost surely in ϕ:

$$\pi(\theta \in \Theta(\phi_0)^\epsilon|\phi) = \lim_{m\to\infty} h_m(\phi).$$

A similar approximation holds for the conditional prior of the ϵ-contraction $\pi(\theta \in \Theta(\phi_0)^{-\epsilon}|\phi)$. Assumption 3.2 is satisfied as long as the identified set $\Theta(\phi)$ is compact and the prior of ϕ is spread over a large support of the parameter space.

Assumption 3.3. *For any $\epsilon > 0$, and $\phi \in \Phi$, $\pi(\theta \in \Theta(\phi)^{-\epsilon}|\phi) < 1$.*

This is an assumption on the prior for θ, which means the identified set should be *sharp* with respect to the prior information. Roughly speaking, the support of the prior should not be a proper subset of any ϵ-contraction of the identified set $\Theta(\phi)$. If otherwise the prior information restricts θ to be inside a strict subset of $\Theta(\phi)$ so that Assumption 3.3 is violated, then that prior information should be taken into account and we should shrink $\Theta(\phi)$ to a sharper set. In the special case when θ is point identified ($\Theta(\phi)$ is a singleton), the ϵ-contraction is empty and thus $\pi(\theta \in \Theta(\phi)^{-\epsilon}|\phi) = 0$.

The following theorem gives the posterior consistency for partially identified parameters.

Theorem 3.1. *Let $\pi(\theta|\phi)$ be a regular conditional distribution. Under assumptions 3.1-3.3, for any $\epsilon > 0$, there is $\tau \in (0, 1]$ such that*

$$P(\theta \in \Theta(\phi_0)^{\epsilon}|D_n) \to^p 1 \quad \textit{and} \quad P(\theta \in \Theta(\phi_0)^{-\epsilon}|D_n) \to^p (1 - \tau).$$

4 Posterior consistency for $\Theta(\phi)$

Let ϕ_0 be the true value of ϕ, which corresponds to the true identified set $\Theta(\phi_0)$. The estimation accuracy of the identified set is often measured, in the literature, by the Hausdorff distance. Specifically, for a point a and a set A, let $d(a, A) = \inf_{x \in A} \|a - x\|$, where $\|\cdot\|$ denotes the Euclidean norm. The Hausdorff distance between sets A and B is defined as

$$d_H(A, B) = \max\left\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A)\right\} = \max\left\{\sup_{a \in A} \inf_{b \in B} \|a - b\|, \sup_{b \in B} \inf_{a \in A} \|b - a\|\right\}.$$

It follows immediately that $d_H(A, B) = d_H(B, A)$ and when both A and B are compact, $d_H(A, B) = 0$ if and only if $A = B$. This section aims at deriving the rate $r_n = o(1)$ such that for some constant $C > 0$,

$$P(d_H(\Theta(\phi), \Theta(\phi_0)) < Cr_n|D_n) \to^p 1.$$

The above result is based upon the posterior concentration rate for ϕ – in the sense that r_n is the same as the concentration rate for ϕ – as well as some continuity condition on $d_H(\Theta(\phi), \Theta(\phi_0))$ with respect to ϕ.

In a semi-parametric Bayesian model where ϕ is point identified and either a nonparametric or a semi-parametric prior is placed, the posterior of ϕ achieves a near-parametric concentration rate under proper conditions on the prior. Since our goal is to study the posterior of $\Theta(\phi)$ and θ instead of ϕ, we state a high level assumption on the posterior of ϕ as follows instead of deriving it from more general conditions. More formal derivations of this assumption will be presented in appendix B.

Assumption 4.1. *The marginal posterior of ϕ is such that*

$$P(\|\phi - \phi_0\| \leq Cn^{-1/2}(\log n)^{1/2}|D_n) \to^p 1.$$

This assumption is imposed for both kinds of priors described in Section 2, and is a standard result in semi-nonparametric Bayesian literature. If we place a nonparametric prior on F as described in Section 2.2.1, this assumption becomes

$$P(\|\phi(F) - \phi(F_0)\| \leq Cn^{-1/2}(\log n)^{1/2}|D_n) \to^p 1.$$

Primitive conditions for the validity of this case can be found in a recent work by Rivoirard and Rousseau (2012). On the other hand, if we parametrize the model in $\mathcal{F} = \{F_{\phi,\eta} : \phi \in \Phi, \eta \in \mathcal{P}\}$ as described in Section 2.2.2, with η being an infinite-dimensional nuisance parameter, a sufficient condition for Assumption 4.1 is found in both Bickel and Kleijn (2012) and the appendix of this paper.

Instead of assuming continuity of $d_H(\Theta(\phi), \Theta(\phi_0))$ with respect to ϕ, which is sufficient in order to get the concentration rate of $\Theta(\phi)$, we place less demanding assumption that still allow us to get the concentration rate. With this aim, we consider a more specific partially identified model: the *moment inequality model*, which assumes that θ satisfies k moment restrictions:

$$\Psi(\theta, \phi) \leq 0, \quad \Psi(\theta, \phi) = (\Psi_1(\theta, \phi), ..., \Psi_k(\theta, \phi))^T \tag{4.1}$$

where $\Psi : \Theta \times \Phi \to \mathbb{R}^k$ is a known function of (θ, ϕ). The model depends on the data X via the point identified parameter ϕ. In the moment inequality model, the identified set can be characterized as:

$$\Theta(\phi) = \{\theta \in \Theta : \Psi(\theta, \phi) \leq 0\}. \tag{4.2}$$

Since most of the partially identified models can be characterized as moment inequality models, model (4.1)-(4.3) has received extensive attention in the partially identified literature.

Assumption 4.2. *The parameter space* $\Theta \times \Phi$ *is compact.*

Assumption 4.3. $\{\Psi(\theta, \cdot) : \theta \in \Theta\}$ *is Lipschitz equi-continuous on* Φ*, that is, for some* $K > 0$*,* $\forall \phi_1, \phi_2 \in \Phi$*,*

$$\sup_{\theta \in \Theta} \|\Psi(\theta, \phi_1) - \Psi(\theta, \phi_2)\| \leq K\|\phi_1 - \phi_2\|.$$

Given the compactness of Θ, this assumption is satisfied by many interesting examples of moment inequality models.

Assumption 4.4. *There exists a closed neighborhood* $U(\phi_0)$ *of* ϕ_0*, such that for any* $a_n = O(1)$*, and any* $\phi \in U(\phi_0)$*, there exists* $C_\phi > 0$ *that might depend on* ϕ*,*

$$\inf_{\theta : d(\theta, \Theta(\phi)) \geq C_\phi a_n} \max_{i \leq k} \Psi_i(\theta, \phi) > a_n.$$

Intuitively, when θ is bounded away from $\Theta(\phi)$ (up to a rate a_n), at least one of the moment inequalities is violated, which means $\max_{i \leq k} \Psi_i(\theta, \phi) > 0$. This assumption quantifies how much $\max_{i \leq k} \Psi_i(\theta, \phi)$ will depart from zero. This is a sufficient condition for the partial identification condition in Chernozhukov, Hong and Tamer (2007). If we define

$$Q(\theta, \phi) = \|\max(\Psi(\theta, \phi), 0)\| = \left[\sum_{i=1}^{k} (\max(\Psi_i(\theta, \phi), 0))^2\right]^{1/2}$$

then $Q(\theta,\phi)=0$ if and only if $\theta\in\Theta(\phi)$. The partial identification condition in Chernozhukov et al. (2007, Condition (4.5)) assumes that there exists $K>0$ so that for all θ,

$$Q(\theta,\phi)\geq Kd(\theta,\Theta(\phi)), \tag{4.3}$$

which says Q should be bounded below by a number proportional to the distance from the identified set if θ is bounded away from the identified set. Assumption 4.4 is a sufficient condition for (4.3).

Example 4.1 (Interval censored data - *continued*)**.** In the interval censoring data example, $\Psi(\theta,\phi)=(\theta-\phi_2,\phi_1-\theta)^T$, for any $\phi=(\phi_1,\phi_2)$ and $\tilde{\phi}=(\tilde{\phi}_1,\tilde{\phi}_2)$, $\|\Psi(\theta,\phi)-\Psi(\theta,\tilde{\phi})\|=\|\phi-\tilde{\phi}\|$. This verifies Assumption 4.3. Moreover, for any θ such that $d(\theta,\Theta(\phi))\geq a_n$, either $\theta\leq\phi_1-a_n$ or $\theta\geq\phi_2+a_n$. If $\theta\leq\phi_1-a_n$, then $\Psi_2(\theta,\phi)=\phi_1-\theta\geq a_n$; if $\theta\geq\phi_2+a_n$, then $\Psi_1(\theta,\phi)=\theta-\phi_2\geq a_n$. This verifies Assumption 4.4. □

The following theorem shows the concentration rate for the identified set.

Theorem 4.1. *Under Assumptions 4.1-4.4, for some $C>0$,*

$$P(d_H(\Theta(\phi),\Theta(\phi_0))>Cn^{-1/2}(\log n)^{1/2}|D_n)\to^p 0.$$

Remark 4.1. The above result holds for both nonparametric prior $\phi(F)$ and semi-parametric prior (ϕ,η) as described in Section 2. The concentration rate is nearly parametric: $n^{-1/2}(\log n)^{1/2}$. The term $\sqrt{\log n}$ arises commonly in the posterior concentration rate literature. The posterior probability in the theorem is now converging to zero, instead of only being smaller than an arbitrarily small constant. Same rate of convergence in the frequentist perspective has been achieved by Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Kaido and Santos (2011), among others.

Remark 4.2. Recently Kitagawa (2012) obtained the posterior consistency for $\Theta(\phi)$: for any $\epsilon>0$,

$$\lim_{n\to\infty}P(d_H(\Theta(\phi),\Theta(\phi_0))>\epsilon|D_n)=0$$

for almost every sampling sequence of D_n. This result was obtained for the case where θ is a scalar whose identified set $\Theta(\phi)$ is a connected interval and $d_H(\Theta(\phi),\Theta(\phi_0))$ is assumed to be a continuous map of ϕ. In multi-dimensional cases where $\Theta(\phi)$ is a more general convex set, however, verifying the continuity of $d_H(\Theta(\phi),\Theta(\phi_0))$ is much more technically involved, due to the challenge of computing the Hausdorff distance in multi-dimensional minifolds. In contrast, our Lipschitz equi-continuity condition in Assumption 4.3 is much easier to verify in specific examples, as it depends on the moment conditions directly.

5 Bayesian Inference of Support Function

This section develops Bayesian inference for the support function $S_\phi(p)$ of the identified set $\Theta(\phi)$ in the moment inequality model (4.1)-(4.3). Bayesian inference for the support function has two main interests. First, it provides an alternative way to perform estimation of the identified set $\Theta(\phi)$. Second, it allows us to construct a two-sided BCS for $\Theta(\phi)$ in the next section. In this section, we first develop an asymptotically valid linearization in ϕ of the support function. Based on this result we show that posterior consistency can be achieved and prove the Bernstein von Mises theorem for the support function.

5.1 Moment Inequality Model

Our analysis focuses on identified sets which are closed and convex. These sets are completely determined by their support functions, and efficient estimation of support functions may lead to optimality of estimation and inference of the identified set. As a result, much of the new development in the partially identified literature focuses on the support function, e.g., Kaido and Santos (2011), Kaido (2012), Beresteanu and Molinari (2008), Bontemps, Magnac and Maurin (2012).

In the moment inequality model, $\Theta(\phi) := \{\theta \in \Theta;\ \Psi(\theta,\phi) \leq 0\}$, where $\Psi(\theta,\phi)$ is given in (4.1) and each component of $\Psi(\theta,\phi)$ is a convex function of θ for every $\phi \in \Phi$ as stated in the next assumption.

Assumption 5.1. *$\Psi(\theta,\phi)$ is continuous in (θ,ϕ) and convex in θ for every $\phi \in \Phi$.*

Let us consider the support function $S_\phi(\cdot) : \mathbb{S}^d \to \mathbb{R}$ of the identified set $\Theta(\phi)$. We restrict its domain to the unit sphere $\mathbb{S}^d$ in $\mathbb{R}^d$ since $S_\phi(p)$ is positively homogeneous in p. Under assumption 5.1 the support function is the optimal value of an ordinary convex program:

$$S_\phi(p) = \sup_{\theta\in\Theta}\{p^T\theta;\ \Psi(\theta,\phi) \leq 0\}$$

and therefore it also admits a Lagrangian representation (see Rockafellar, chapter 28):

$$S_\phi(p) = \sup_{\theta\in\Theta}\{p^T\theta - \lambda(p,\phi)^T\Psi(\theta,\phi)\}, \tag{5.1}$$

where $\lambda(p,\phi) : \mathbb{S}^d \times \mathbb{R}^{d_\phi} \to \mathbb{R}^k_+$ is a k-vector of Lagrange multipliers. Note that d_ϕ is the dimension of ϕ.

We denote by $\Psi_S(\theta,\phi_0)$ the k_S-subvector containing the constraints that are strictly convex functions of θ and by $\Psi_L(\theta,\phi_0)$ the k_L constraints that are linear in θ. Obviously, $k_S + k_L = k$. The corresponding Lagrange multipliers are denoted by $\lambda_S(p,\phi_0)$ and $\lambda_L(p,\phi_0)$, respectively, for $p \in \mathbb{S}^d$. Moreover, define $\Xi(p,\phi) = \arg\max_{\theta\in\Theta}\{p^T\theta;\ \Psi(\theta,\phi) \leq 0\}$ as the *support set* of $\Theta(\phi)$. Then, by definition,

$$p^T\theta = S_\phi(p), \quad \forall\theta \in \Xi(p,\phi).$$

We also denote by $\nabla_\phi\Psi(\theta,\phi)$ the $k \times d_\phi$ matrix of partial derivatives of Ψ with respect to ϕ. Let $B(\phi_0,\delta) = \{\phi \in \Phi;\ \|\phi - \phi_0\| \leq \delta\}$ denote a closed ball centered at ϕ_0 with radius δ. For every $\phi \in B(\phi_0, r_n)$ and $\theta \in \Theta(\phi)$, we denote by $Act(\theta,\phi) := \{i;\ \Psi_i(\theta,\phi) = 0\}$ the set of the inequality active constraint indices and by $d_A(\theta,\phi)$ the number of its elements. For every $i \in Act(\theta,\phi)$, $\nabla_\theta\Psi_i(\theta,\phi)$ denotes the d-vector of partial derivatives of Ψ_i with respect to θ. We assume the following:

Assumption 5.2. *The true value ϕ_0 is in the interior of Φ, and Θ is convex and compact.*

Assumption 5.3. *There is some $\delta > 0$ such that for all $\phi \in B(\phi_0, \delta)$, we have:*
(i) the $k \times d_\phi$ matrix $\nabla_\phi \Psi(\theta, \phi)$ exists and is continuous in (θ, ϕ);
(ii) the set $\Theta(\phi)$ is non empty;
(iii) there exists a $\theta \in \Theta$ such that $\Psi(\theta, \phi) < 0$;
(iv) $\Theta(\phi) \subset int(\Theta)$ where $int(\Theta)$ denotes the interior of Θ;
(v) for every $i \in Act(\theta, \phi_0)$, with $\theta \in \Theta(\phi_0)$, the vector $\nabla_\theta \Psi_i(\theta, \phi)$ exists and is continuous in $(\theta, \phi) \in \Theta \times B(\phi_0, \delta)$.

Assumption 5.3 (iii) implies assumption 5.3 (ii). However, we prefer to keep both conditions since in order to establish some technical results we only need condition (ii) which is weaker.

The next assumption concerns the inequality active constraints. In particular, assumption 5.4 (i) may be restrictive in the one dimensional case (*i.e.* $d = 1$) but is easily verified in the cases with $d > 1$. For instance, in example 2.1 this assumption is not verified in the degenerate case where $\phi_1 = \phi_2$. Assumption 5.4 (ii) says that the active inequality constraint gradients $\nabla_\theta \Psi_i(\theta, \phi_0)$ are linear independent. This assumption guarantees that a θ which solves the optimization problem (5.1) with $\phi = \phi_0$ satisfies the Kuhn-Tucker conditions. Alternative assumptions that are weaker than assumption 5.4 (ii) could be used, but the advantage of assumption 5.4 (ii) is that it is easy to check.

Assumption 5.4. *(i) $d_A(\theta, \phi_0) \leq d$ for every $\theta \in \Theta(\phi_0)$ where $d_A(\theta, \phi_0)$ is the number of active constraints;*

(ii) the gradient vectors $\{\nabla_\theta \Psi_i(\theta, \phi)\}_{i \in Act(\theta, \phi_0)}$, are linearly independent $\forall \theta \in \Theta(\phi_0)$.

The following assumption is sufficient for the differentiability of the support function at ϕ_0:

Assumption 5.5. *At least one of the following holds:*

(i) For the ball $B(\phi_0, \delta)$ in Assumption 5.3, for every $(p, \phi) \in \mathbb{S}^d \times B(\phi_0, \delta)$, $\Xi(p, \phi)$ is a singleton;

(ii) There are linear constraints in $\Psi(\theta, \phi_0)$, which are also separable in θ, that is, $\Psi_L(\theta, \phi_0) = A_1 \theta + A_2(\phi_0)$ for some function $A_2 : \Phi \to \mathbb{R}^{k_L}$ (not necessarily linear) and some $(k_L \times d)$-matrix A_1.

Assumption 5.5 is particularly important for the linearization of the support function that we develop in section 5.2. In fact, if one of the two parts of Assumption 5.5 holds then the support function is differentiable for every $(p, \phi) \in \mathbb{S}^d \times B(\phi_0, \delta)$ and we have a closed form for its derivative.

The last set of assumptions that we introduce will be used to prove the Bernstein von Mises theorem for $S_\phi(\cdot)$ and allows to strengthen the result of theorem 5.1 below. The first three assumptions are (local) Lipschitz equi-continuity assumptions.

Assumption 5.6. *For the ball $B(\phi_0, \delta)$ in assumption 5.3, for some $K_1, K_2, K_3 > 0$ and $\forall \phi_1, \phi_2 \in B(\phi_0, \delta)$:*

(i) $\sup_{p\in\mathbb{S}^d}\|\lambda(p,\phi_1)-\lambda(p,\phi_2)\|\le K_1\|\phi_1-\phi_2\|$*;*

(ii) $\sup_{\theta\in\Theta}\|\nabla_\phi\Psi(\theta,\phi_1)-\nabla_\phi\Psi(\theta,\phi_2)\|\le K_2\|\phi_1-\phi_2\|$*;*

(iii) $\|\nabla_\phi\Psi(\theta_1,\phi_0)-\nabla_\phi\Psi(\theta_2,\phi_0)\|\le K_3\|\theta_1-\theta_2\|$*, for every* $\theta_1,\theta_2\in\Theta$*;*

(iv) If $\Xi(p,\phi_0)$ *is a singleton* $\forall p\in W$ *for some compact subset* $W\subseteq\mathbb{S}^d$ *then there exists a* $\varepsilon=O(\delta)$ *such that* $\Xi(p,\phi_1)\subseteq\Xi^\varepsilon(p,\phi_0)$*.*

We show in the following example that Assumptions 5.1-5.6 are easily satisfied

Example 5.1 (Interval censored data - *continued*)**.** The setup is the same as in Example 2.1. Assumption 5.2 is verified if Y_1 and Y_2 are two random variables with finite first moments $\phi_{0,1}$ and $\phi_{0,2}$, respectively. Moreover, $\Psi(\theta,\phi)=(\phi_1-\theta,\theta-\phi_2)^T$, $\phi=(\phi_1,\phi_2)^T$,

$$\nabla_\phi\Psi(\theta,\phi)=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

so that Assumptions 5.1, 5.2 and 5.3 *(i)*-*(ii)* are trivially satisfied. Assumption 5.3 *(iii)* holds for every θ inside (ϕ_1,ϕ_2); Assumption 5.3 *(iv)* is satisfied if ϕ_1 and ϕ_2 are bounded. To see that Assumptions 5.3 *(v)* and 5.4 are satisfied note that $\forall\theta<\phi_{0,1}$ we have $Act(\theta,\phi_0)=\{1\}$, $\forall\theta>\phi_{0,2}$ we have $Act(\theta,\phi_0)=\{2\}$ while $\forall\theta\in[\phi_{0,1},\phi_{0,2}]$ we have $Act(\theta,\phi_0)=\varnothing$. Assumption 5.5 *(i)* and *(ii)* are both satisfied since the support set takes the values $\Xi(1,\phi)=\phi_2$ and $\Xi(-1,\phi)=-\phi_1$ and the constraints in $\Psi(\theta,\phi_0)$ are both linear with $A_1=(-1,1)^T$ and $A_2(\phi_0)=\nabla_\phi\Psi(\theta,\phi_0)\phi_0$.

In order to verify assumption 5.6, we use the largest eigenvalue as the matrix norm. The eigenvalues of $\nabla_\phi\Psi(\theta,\phi)$ are $\{1,-1\}$ for every θ and ϕ. Hence, assumptions 5.6 *(ii)*-*(iii)* are verified. The lagrange multiplier is $\lambda(p,\phi)=(-pI(p<0),pI(p\ge 0))^T$ so that assumption 5.6 *(i)* is satisfied since the norm is equal to 0. Finally, the support set $\Xi(p,\phi)=\phi_1 I(p<0)+\phi_2 I(p\ge 0)$ is a singleton for every $\phi\in B(\phi_0,\delta)$ and $\Xi(p,\phi_0)^\varepsilon=\{\theta\in\Theta;\|\theta-\theta_*\|\le\varepsilon\}$ where $\theta_*=\Xi(p,\phi_0)=\phi_{0,1}I(p<0)+\phi_{0,2}I(p\ge 0)$. Therefore, $\|\Xi(p,\phi)-\theta_*\|\le\delta$ and assumption 5.6 *(iv)* holds with $\varepsilon=\delta$. □

5.2 Asymptotic Analysis

The support function of a closed and convex set is in general non-differentiable in p but it admits directional derivatives, see e.g. Milgrom and Segal (2002). Luckily, when assumption 5.5 holds the derivative of the support function exists. We exploit this fact to derive an expansion in ϕ of the support function. This allows us to establish a Bernstein-von Mises type result for the posterior distribution of the support function.

The next theorem states that the support function can be locally approximated (asymptotically) by a linear function of $\phi_1,\phi_2\in B(\phi_0,r_n)$ for $r_n=o(1)$ a bounded sequence depending on the sample size n. The expansion is stochastic when ϕ is interpreted as a random variable associated with the posterior distribution $P(\cdot|D_n)$.

Theorem 5.1. *Let* $\theta_*(p):\mathbb{S}^d\to\Theta$ *be a Borel measurable mapping satisfying* $\theta_*(p)\in\Xi(p,\phi_0)$ *for all* $p\in\mathbb{S}^d$*. If assumptions 5.1-5.5 hold with* $\delta=r_n$ *for some* $r_n=o(1)$*, then there* *is a N such that for every $n \geq N$ there exist:* (i) *a real function $f(\phi_1, \phi_2)$ defined for every $\phi_1, \phi_2 \in B(\phi_0, r_n)$ and* (ii) *a function $\lambda(p, \phi_0) : \mathbb{S}^d \times \mathbb{R}^{d_\phi} \to \mathbb{R}^k_+$ such that for every $\phi_1, \phi_2 \in B(\phi_0, r_n)$:*

$$\sup_{p \in \mathbb{S}^d} \left| (S_{\phi_1}(p) - S_{\phi_2}(p)) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)[\phi_1 - \phi_2] \right| = f(\phi_1, \phi_2)$$

and $\frac{f(\phi_1,\phi_2)}{\|\phi_1 - \phi_2\|} \to 0$ uniformly in $\phi_1, \phi_2 \in B(\phi_0, r_n)$ as $n \to \infty$.

We remark that the functions λ and θ_* do not depend on the specific choice of ϕ_1 and ϕ_2 inside $B(\phi_0, r_n)$, but only on p and the true value ϕ_0. With the approximation given in the theorem we are now ready to state posterior consistency (with concentration rate) and asymptotic normality of the posterior distribution of $S_\phi(p)$. In the next theorems we will set $r_n = (\log n)^{1/2} n^{-1/2}$ when n is sufficiently large.

Theorem 5.2. *Under assumption 4.1 and the assumptions of Theorem 5.1 with $r_n = \sqrt{(\log n)/n}$, for some $C > 0$,*

$$P(\sup_{p \in \mathbb{S}^d} |S_\phi(p) - S_{\phi_0}(p)| < C(\log n)^{1/2} n^{-1/2} | D_n) \to^p 1. \quad (5.2)$$

Remark 5.1. Notice that $d_H(\Theta(\phi), \Theta(\phi_0)) = \sup_{p \in \mathbb{S}^d} |S_\phi(p) - S_{\phi_0}(p)|$. Therefore, (5.2) is another statement of Theorem 4.1. However, they are obtained by different proof strategies. In particular, Theorem 5.2 is obtained as a byproduct of the Bernstein-von Mises theorem stated in theorem 5.3 below and is based on the asymptotic local expansion of the support function as in theorem 5.1. As will be shown below, this expansion also yields the Bernstein von Mises theorem of the support function, that is, the posterior of the support function is asymptotically normal.

We now state a Bernstein-von Mises (BvM) theorem for the support function. This theorem is valid under the assumption that a Bernstein-von Mises (BvM) theorem holds for the posterior distribution of the finite-dimensional identified parameter ϕ. We denote by $\| \cdot \|_{TV}$ the total variation distance, that is, for two probability measures P and Q,

$$\|P - Q\|_{TV} := \sup_B |P(B) - Q(B)|$$

where B is an element of the σ-algebra on which P and Q are defined.

Assumption 5.7. *Let $P_{\sqrt{n}(\phi - \phi_0)|D_n}$ denote the posterior distribution of $\sqrt{n}(\phi - \phi_0)$. We assume*

$$\|P_{\sqrt{n}(\phi - \phi_0)|D_n} - \mathcal{N}_{d_\phi}(\tilde{\Delta}_{n,\phi_0}, \tilde{I}^{-1}_{\phi_0})\|_{TV} \to^p 0$$

where $\mathcal{N}_{d_\phi}$ denotes the d_ϕ-dimensional normal distribution, $\tilde{\Delta}_{n,\phi_0} := n^{-1/2} \sum_{i=1}^n \tilde{I}^{-1}_{\phi_0} \tilde{l}_{\phi_0}(X_i)$, $\tilde{l}_{\phi_0}$ is the semi-parametric efficient score function of the model and $\tilde{I}_{\phi_0}$ denotes the semi-parametric efficient information matrix.

For primitive conditions for the validity of this assumption in semi-parametric models we refer to Bickel and Kleijn (2012) and Rivoirard and Rousseau (2012). Despite of the notation, remark that $\tilde{l}_{\phi_0}$ and $\tilde{I}_{\phi_0}$ also depend either on the true η_0 or on the true F_0 – depending whether the model has been re-parameterized or not. The *semi-parametric efficient score function* and the *semi-parametric efficient information* contribute to the stochastic local asymptotic normality (LAN) expansion of the integrated likelihood, which is necessary in order to get the BvM result in assumption 5.7. A precise definition of $\tilde{l}_{\phi_0}$ and $\tilde{I}_{\phi_0}$ may be found in van der Vaart (2002) (Definition 2.15).

Theorem 5.3. *If the assumptions of Theorem 5.1 and assumption 5.6 hold with* $\delta = r_n = \sqrt{(\log n)/n}$*, under assumption 5.7:*

$$\|P_{\sqrt{n}\sup_{p\in\mathbb{S}^d}(S_\phi(p)-S_{\phi_0}(p))|D_n} - \mathcal{N}(\bar{\Delta}_{n,\phi_0}, \bar{I}_{\phi_0}^{-1})\|_{TV} \to^p 0$$

where $\bar{\Delta}_{n,\phi_0} = \lambda(p,\phi_0)^T \nabla_\phi \Psi(\theta_*(p),\phi_0)\tilde{\Delta}_{n,\phi_0}$ *and*

$$\bar{I}_{\phi_0}^{-1} = \lambda(p,\phi_0)^T \nabla_\phi \Psi(\theta_*(p),\phi_0)\tilde{I}_{\phi_0}^{-1}\nabla_\phi \Psi(\theta_*(p),\phi_0)^T \lambda(p,\phi_0).$$

The asymptotic mean and covariance matrix may be easily estimated by replacing ϕ_0 by any consistent estimator $\hat{\phi}$ of ϕ_0. So that $\theta_*(p)$ will be replaced by an element $\hat{\theta}_*(p) \in \Xi(p,\hat{\phi})$ and an estimate of $\lambda(p,\phi_0)$ will be obtained by solving – eventually numerically – the ordinary convex program in (5.1) with ϕ_0 replaced by $\hat{\phi}$.

Remark 5.2. The posterior asymptotic variance of the support function $\bar{I}_{\phi_0}^{-1}$ is the same as that of the frequentist estimator obtained by Kaido and Santos (2012, Theorem 3.2), and both are derived based on a linear expansion of the support function. On one hand, the linear expansion of Theorem 5.1 is obtained from expanding $\Psi(\theta_*(p),\phi) - \Psi(\theta_*(\theta),\phi_0)$ in a neighborhood of ϕ_0. This gives the asymptotic variance $\nabla_\phi \Psi(\theta_*(p),\phi_0)\tilde{I}_{\phi_0}^{-1}\nabla_\phi \Psi(\theta_*(p),\phi_0)^T$, which is semi-parametric efficient for estimating $\Psi(\theta_*(p),\phi_0)$ as guaranteed by the Bernstein von Mises theorem proved by Bickel and Kleijn (2012). On the other hand, Kaido and Santos (2012)'s frequentist estimator of the support function has a linear representation in terms of $\widehat{\Psi}(\theta_*(p)) - \Psi(\theta_*(\theta),\phi_0)$, where $\widehat{\Psi}(\theta_*(p))$ is a sample analog of $\Psi(\theta_*(\theta),\phi_0)$ and is therefore semi-parametric efficient. This implies that the asymptotic variances of the support function from both Bayesian and frequentist approaches are the same.

The asymptotic normality of the posterior of $S_\phi(p)$ also implies that the posterior coverage and the coverage under the limiting normal of our two-sided BCS for the identified set – that we construct in the next section – are the same.

6 Bayesian Credible Sets

In this section we focus on two kinds of credible sets: credible sets for θ and credible sets for the identified set $\Theta(\phi)$.

6.1 Credible set for θ

Bayesian inference on θ can be carried out through finite-sample Bayesian credible sets (BCS). A BCS is a set BCS(τ) such that

$$P(\theta \in \mathrm{BCS}(\tau)|D_n) = 1 - \tau \tag{6.1}$$

at level $1-\tau$, for $\tau \in (0,1)$. Apparently such a definition is not unique. One of the popular choice of the credible set is the highest-probability-density (HPD) set, which has been widely used in empirical studies and also used in the Bayesian partially identified literature e.g., Moon and Schorfheide (2012) and Norets and Tang (2012).

The BCS then can be compared with the frequentist confidence set (FCS). Let $P_{D_n}(.)$ denote the probability measure based on the sampling distribution, where $(\theta, \phi, \eta) = (\theta_0, \phi_0, \eta_0)$ or $(\theta, F) = (\theta_0, F_0)$. A frequentist confidence set FCS(τ) for θ_0 satisfies

$$\lim_{n\to\infty} \inf_{\phi\in\Phi} \inf_{\theta_0\in\Theta(\phi)} P_{D_n}(\theta_0 \in \mathrm{FCS}(\tau)) \geq 1 - \tau.$$

There have been various procedures proposed in the frequentist literature to construct FCS(τ) that satisfies the above inequality. One of the key properties of these proposed FCS is that they are based on some consistent estimator $\hat{\phi}$ of ϕ_0, and $\Theta(\hat{\phi}) \subset \mathrm{FCS}(\tau)$. Moon and Schorfheide (2012) compared HPD with FCS and showed that in a parametric Bayesian model with known likelihood, for any $\tau > 0$, $P(\theta \in \mathrm{HPD}(\tau), \theta \notin \mathrm{FCS}(\tau)|D_n) = o_p(1)$, that is, the FCS is too large to do Bayesian inference. Under the more robust semi-parametric Bayesian setup, the frequntist confidence set is also "too big" from the Bayesian point of view, shown by Theorem 6.1 below.

The following assumption is needed.

Assumption 6.1. *(i) The frequentist FCS(τ) is such that, there is $\hat{\phi}$ with $\|\hat{\phi} - \phi_0\| = o_p(1)$ satisfying $\Theta(\hat{\phi}) \subset FCS(\tau)$.*
(ii) $\sup_{(\theta,\phi)\in\Theta\times\Phi} \pi(\theta|\phi) < \infty$.

Many frequentist FCS's satisfy condition (i), see, e.g., Imbens and Manski (2004), Chernozhukov, Hong and Tamer (2007), Rosen (2008), Andrews and Soares (2010), etc. Condition (ii) is easy to verify since $\Theta \times \Phi$ is compact. Examples of $\pi(\theta|\phi)$ include: the uniform prior with density

$$\pi(\theta|\phi) = \mu(\Theta(\phi))^{-1} I_{\theta\in\Theta(\phi)},$$

where $\mu(\cdot)$ denotes the Lebesgue measure; and the truncated normal prior with density

$$\pi(\theta|\phi) = \left[\int_{\Theta(\phi)} h(x; \lambda, \Sigma)dx\right]^{-1} h(\theta; \lambda, \Sigma) I_{\theta\in\Theta(\phi)},$$

where $h(x; \lambda, \Sigma)$ is the density function of multinormal $N(\lambda, \Sigma)$.

Theorem 6.1. *Under Assumptions 4.1, the assumptions of theorem 5.1 with $r_n = \sqrt{(\log n)/n}$, and 6.1, for any $\tau > 0$,*
(i)

$$P(\theta \notin FCS(\tau)|D_n) = o_p(1),$$

(ii)

$$P(\theta \in FCS(\tau), \theta \notin BCS(\tau)|D_n) \to^p \tau.$$

Remark 6.1. Theorem 6.1 (i) shows that the posterior probability that θ lies inside the frequentist confidence set is arbitrarily close to one, as $n \to \infty$. This indicates that the FCS is too big to do insightful statistical inference from the Bayesian point of view. On the other hand, (ii) demonstrates that with a nonnegligible probability, FCS is strictly larger than BCS. Therefore, FCS is conservative from a Bayesian perspective.

Remark 6.2. Similar results have been shown by Moon and Schorfheide (2012) when HPD is used as the Bayesian credible set. The result presented here, besides allowing a semi-parametric likelihood function, is more general. Our proof for part (i) is slightly different from the proof in Moon and Schorfheide (2012, Corollary 1), in that we rely on the continuity of $d(\Theta(\phi), \Theta(\phi_0))$ with respect to ϕ, and is achieved through an asymptotic expansion of the support function. The proof for part (ii) follows the same argument of Moon and Schorfheide (2012)'s.

6.2 Two-sided credible set for $\Theta(\phi)$

We now construct an asymptotic valid BCS for $\Theta(\phi)$. We are aiming at constructing two-sided credible sets A_1 and A_2 such that

$$P(A_1 \subset \Theta(\phi) \subset A_2|D_n) \geq 1 - \tau$$

with probability approaching one. The one-sided set A_2 is easy to obtain. As suggested in an earlier circulated version of Moon and Schorfheide (2012) and Norets and Tang (2012), suppose $BCS_\phi(\tau)$ is a $1-\tau$ Bayesian credible set of ϕ, then it is easy to show that

$$P\left(\Theta(\phi) \subset \bigcup_{x \in BCS_\phi(\tau)} \Theta(x) \middle| D_n\right) = 1 - \tau$$

for every sampling sequence D_n. However, it is difficult to extend the idea of using the BCS of ϕ to construct the two-sided sets, more specifically, to construct the lower set A_1. In this section, we apply a new idea, with the help of the support function for such a task. To our best knowledge, this is the first in the literature that constructs the two-sided BCS for $\Theta(\phi)$.

To illustrate why support function can help, for $\Theta(\phi)$, recall its ϵ-envelope as $\Theta(\phi)^\epsilon = \{\theta \in \Theta : d(\theta, \Theta(\phi)) \leq \epsilon\}$, and ϵ-contraction as $\Theta(\phi)^{-\epsilon} = \{\theta \in \Theta(\phi) : d(\theta, \Theta \backslash \Theta(\phi)) \geq \epsilon\}$ where $\epsilon \geq 0$ and $\Theta \backslash \Theta(\phi) = \{\theta \in \Theta : \theta \notin \Theta(\phi)\}$ as in (3.1) and (3.2). If $\Theta(\phi_1)$, $\Theta(\phi_2)^\epsilon$ and $\Theta(\phi_3)^{-\epsilon}$ are convex, for some ϕ_1, ϕ_2 and $\phi_3 \in \Phi$, then we have:

$$\Theta(\phi_1) \subset \Theta(\phi_2)^\epsilon \text{ if and only if } \sup_{\|p\|=1} (S_{\phi_1}(p) - S_{\phi_2}(p)) \leq \epsilon.$$

and

$$\Theta(\phi_3)^{-\epsilon} \subset \Theta(\phi_1) \text{ if and only if } \sup_{\|p\|=1} (S_{\phi_3}(p) - S_{\phi_1}(p)) \leq \epsilon.$$

Let $\hat{\phi}_M$ be the posterior mode, that is, $\hat{\phi}_M = \arg\max p(\phi|D_n)$. Then for any $c_n \geq 0$,

$$P(\Theta(\hat{\phi}_M)^{-c_n} \subset \Theta(\phi) \subset \Theta(\hat{\phi}_M)^{c_n}|D_n) = P(\sup_{\|p\|=1} |S_{\Theta(\phi)}(p) - S_{\Theta(\hat{\phi}_M)}(p)| \leq c_n|D_n).$$

Note that the right hand side of the above equation depends on the posterior of the support function. The posterior mode is only an example, we point out that any consistent estimator could be used to construct the two-sided credible region. Let q_τ be the $1-\tau$ quantile of the posterior of

$$J(\phi) = \sqrt{n} \sup_{\|p\|=1} |S_\phi(p) - S_{\hat{\phi}_M}(p)|$$

so that

$$P\left(J(\phi) \leq q_\tau \middle| D_n \right) = 1 - \tau. \tag{6.2}$$

The posterior of $J(\phi)$ is determined by that of ϕ. Hence q_τ can be simulated from the MCMC draws of $p(\theta|D_n)$. Immediately, we have the following theorem:

Theorem 6.2. *Suppose for any $\tau \in (0,1)$, q_τ is defined as in (6.2), then for every sampling sequence D_n,*

$$P(\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}|D_n) = 1 - \tau.$$

Remark 6.3. It is straightforward to construct the one-sided BCS for $\Theta(\phi)$ using the described procedure. For example, let $\tilde{q}_\tau$ and $\hat{q}_\tau$ be such that $P(\sup_{\|p\|=1}(S_\phi(p) - S_{\hat{\phi}_M}(p)) \leq \tilde{q}_\tau|D_n) = 1 - \tau$, and $P(\sup_{\|p\|=1}(S_{\hat{\phi}_M}(p) - S_\phi(p)) \leq \hat{q}_\tau|D_n) = 1 - \tau$, then $P(\Theta(\phi) \subset \Theta(\hat{\phi}_M)^{\tilde{q}_\tau/\sqrt{n}}|D_n) = 1 - \tau$ and $P(\Theta(\hat{\phi}_M)^{-\hat{q}_\tau/\sqrt{n}} \subset \Theta(\phi)|D_n) = 1 - \tau$ for every sampling sequence D_n.

6.3 Frequentist coverage probability of BCS for $\Theta(\phi)$

As we have shown in theorem 6.1, the BCS for θ does not have a correct frequentist coverage when θ is partially identified, since the BCS tends to be a subset of the interior of FCS. Gustafson (2012) showed that from a frequentist point of view, there is always a region of the identified set which Bayesian credible interval fails to cover.

In contrast, the constructed two-sided BCS for the identified set has desired frequentist properties. Recently, Kitagawa (2012) constructed a one-sided credible set that also has a correct frequentist probability when $\Theta(\phi)$ is an one dimensional interval for a scalar. The frequentist coverage probability for a more general multi-dimensional BCS have been largely unknown in the literature before. Our two-sided BCS is constructed based on the support function, for which the Bernstein von Mises Theorem holds (see Theorem 5.3) in the moment inequality model, which implies that the frequentist coverage probability is asymptotically correct. We show this in theorem 6.3 below.

The analysis relies on the following assumption, which requires the asymptotic normality of the posterior mode of ϕ (or of the consistent estimation used to construct the BCS). The asymptotic normality of posteriors modes has been long realized, and holds under mild conditions.

Assumption 6.2. *The posterior mode $\hat{\phi}_M$ is such that*

$$\sqrt{n}(\hat{\phi}_M - \phi_0) \to^d N(0, \tilde{I}_{\phi_0}^{-1})$$

where $\tilde{I}_{\phi_0}$ denotes the semi-parametric efficient information matrix as in Assumption 5.7.

Theorem 6.3. *Consider the moment inequality model in (4.1). If assumptions 5.1-5.6 hold with $\delta = r_n = \sqrt{(\log n)/n}$ then the constructed two-sided Bayesian credible set has asymptotically correct frequentist coverage probability, that is,*

$$P_{D_n}(\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}) \geq 1 - \tau + o_p(1).$$ [4]

Similarly we can show that the one-sided BCS's as constructed in Remark 6.3 also have asymptotically correct coverage probabilities. For example, for $\tilde{q}_\tau$ such that $P(\sup_{\|p\|=1}(S_\phi(p) - S_{\hat{\phi}_M}(p)) \leq \tilde{q}_\tau | D_n) = 1 - \tau$, then

$$P_{D_n}(\Theta(\phi_0) \subset \Theta(\hat{\phi}_M)^{\tilde{q}_\tau/\sqrt{n}}) \geq 1 - \tau + o_p(1). \tag{6.3}$$

Remark 6.4. Our BCS is constructed based on the support function, whose frequentist coverage probability is guaranteed by the Bernsten von Mises theorem of the support function, proved in Theorem 5.3. Since the normal distribution is also the limiting distribution for efficient frequentist inference about the support function (see Kaido and Santos 2011), our two-sided BCS can be interpreted as asymptotically efficient confidence region for the identified set.[5]

We can also use $\Theta(\hat{\phi}_M)^{\tilde{q}_\tau/\sqrt{n}}$ as the frequentist confidence set for θ, which then will have asymptotically correct frequentist coverage probability. The result is stated as following:

Corollary 6.1. *Under the assumptions of Theorem 6.3,*

$$\inf_{\theta \in \Theta(\phi_0)} P_{D_n}(\theta \in \Theta(\hat{\phi}_M)^{\tilde{q}_\tau/\sqrt{n}}) \geq 1 - \tau + o_p(1).$$

6.4 Missing data: an example

We illustrate our method using a missing data example, which was discussed thoroughly by Manski (2004). For simplicity of exposition, we present the simplest version. Let Y be a binary variable, indicating whether a treatment is succesful ($Y = 1$) or not ($Y = 0$). However, Y is observed subject to missing. We write $M = 0$ if Y is missing, and $M = 1$ otherwise. Hence we in fact observe (M, MY). The parameter of interest is $\theta = P(Y = 1)$, the probability of success. Moreover, we denote the identified parameters

$$\phi_1 = P(M = 1), \quad \phi_2 = P(Y = 1 | M = 1).$$

[4]The result presented here is understood as: There is a random sequence $\Delta(D_n)$ that depends on D_n such that $\Delta(D_n) = o_p(1)$, and for any sampling sequence D_n, we have $P_{D_n}(\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}) \geq 1 - \tau + \Delta(D_n)$. Similar interpretation applies to (6.3) and Corollary 6.1.

[5]The asymptotic efficiency based on the support function is achieved by Kaido and Santos (2011, Theorem 5.4).

Let $\phi_0 = (\phi_{10}, \phi_{20})$ be the true values of $\phi = (\phi_1, \phi_2)$ respectively. Then without further assumption on $P(Y = 1|M = 0)$, θ_0 is only partially identified on $\Theta(\phi_0)$ where $\Theta(\phi) = [\phi_1\phi_2, \phi_1\phi_2 + 1 - \phi_1]$. The support function is easy to calculate, which is

$$S_\phi(1) = \phi_1\phi_2 + 1 - \phi_1 \quad S_\phi(-1) = -\phi_1\phi_2.$$

Suppose we observe i.i.d. data $\{(M_i, Y_iM_i)\}_{i\leq n}$, and find that $\sum_{i=1}^n M_i = n_1$ and $\sum_{i=1}^n Y_iM_i = n_2$, the number of nonmissing observations and observed success respective. In this example, the true likelihood function $L(\phi) \propto \phi_1^{n_1}(1-\phi_1)^{n-n_1}\phi_2^{n_2}(1-\phi_2)^{n_1-n_2}$ is known.

We place independent Beta priors Beta(α_1, β_1) and Beta(α_2, β_2) on (ϕ_1, ϕ_2). The uniform distribution is a special case of Beta prior. Then the posterior of (ϕ_1, ϕ_2) is a product of Beta$(\alpha_1 + n_1, \beta_1 + n - n_1)$ and Beta$(\alpha_2 + n_2, \beta_2 + n_1 - n_2)$. If in addition, we have subjective prior information on θ and place a prior $\pi(\theta|\phi)$ supported on $\Theta(\phi)$, then by integrating out ϕ, we immediately obtain the marginal posterior of θ.

We now present the BCS for $\Theta(\phi)$ obtained by using the support function of $\Theta(\phi)$. First, by taking the derivative of $p(\phi|D_n)$, we obtain the posterior mode: $\hat{\phi}_{1M} = (n_1+\alpha_1-1)/(n+\alpha_1+\beta_1-2)$, and $\hat{\phi}_{2M} = (n_2+\alpha_2-1)/(n_1+\alpha_2+\beta_2-2)$. Then

$$J(\phi) = \sqrt{n}\max\left\{|\phi_1\phi_2 - \phi_1 - \hat{\phi}_{1M}\hat{\phi}_{2M} + \hat{\phi}_{1M}|, |\phi_1\phi_2 - \hat{\phi}_{1M}\hat{\phi}_{2M}|\right\}.$$

Let q_τ be the $1-\tau$ quantile of the posterior of $J(\phi)$, which can be obtained by simulating from the Beta distributions. The lower and upper $1-\tau$ level BCS's for $\Theta(\phi)$ are $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$ where

$$\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} = [\hat{\phi}_{1M}\hat{\phi}_{2M} + q_\tau/\sqrt{n}, \hat{\phi}_{1M}\hat{\phi}_{2M} + 1 - \hat{\phi}_{1M} - q_\tau/\sqrt{n}],$$
$$\Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}} = [\hat{\phi}_{1M}\hat{\phi}_{2M} - q_\tau/\sqrt{n}, \hat{\phi}_{1M}\hat{\phi}_{2M} + 1 - \hat{\phi}_{1M} + q_\tau/\sqrt{n}],$$

which are also two-sided asymptotic $1-\tau$ frequentist confidence intervals of the true $\Theta(\phi_0)$.

Here we present a simple simulated example, where the true $\phi_0 = (0.7, 0.5)$. This implies the true identified interval to be $[0.35, 0.65]$ and about thirty percent of the simulated data are "missing". Suppose we had no prior knowlege about the true ϕ_0, and place a uniform prior on it. Thus $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$. In addition, $B = 1,000$ posterior draws $\{\phi^i\}_{i=1}^B$ are sampled from $p(\phi_1, \phi_2|D_n) \sim$Beta$(n_1+1, n-n_1+1)\times$Beta$(n_2+1, 1+n_1-n_2)$. Then, for each of them compute $J(\phi^i)$ and set $q_{0.05}$ as the 95% upper quantile of $\{J(\phi^i)\}_{i=1}^B$ to obtain the critical value of BCS and construct the two-sided BCS for the identified set. Each simulation is repeated for 1000 times to calculate the coverage frequency of the true identified interval. See Table 1 for the results.

7 From Partial Identification to Point Identification

We have been focusing on partially identified models. However, results derived for the identified set and the support function are still valid when point identification is achieved. This is important because in many cases it is possible that we actually have point identification and, in that event, $\Theta(\phi)$ degenerates to a singleton. For example, in the interval

Table 1: Frequentist coverage probability of the true identified interval

n	Lower	Upper	TwoSided
50	0.967	0.954	0.927
100	0.975	0.971	0.954
500	0.976	0.974	0.953

Lower, Upper and Two sided represent the frequencies of the events $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0)$, $\Theta(\phi_0) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$, *and* $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$ *over 1000 replicates.*

censored model, it is possible that $EY_1 = EY_2$, in which case $\theta = EY$ is point identified.

When point identification is indeed achieved, the one-sided coverage $\Theta(\phi) \subset \Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0)$ in Theorems 6.2 and 6.3, and the asymptotic normality for the posterior of the support function of Theorem 5.3 still hold because they are generally guaranteed by the semi-parametric Bernstein-von Mises theorem for ϕ when $\Theta(\phi)$ is a singleton (e.g., Rivoirard and Rousseau 2012, Bickel and Kleijn 2011). Theorem 4.1 is also guaranteed by the concentration theory for the posterior of ϕ (Assumption 4.1), which then implies the posterior consistency of the support function, and the same concentration rate as Theorem 5.2.

When θ is identified, $\{\theta\} = \Theta(\phi) = f(\phi)$, which is a function of ϕ, and $S_\phi(p) = p^T\theta$. Thus the posterior of θ is the same as the posterior of $\Theta(\phi)$, which is completely determined by that of ϕ under "smoothness" conditions on Ψ. As a result, Theorem 3.1 is still valid because it is implied by Theorem 4.1, which also comes straightforward from the posterior consistency of ϕ if $f(\cdot)$ is continuous at ϕ_0. Theorem 6.1, however, does not hold anymore because when θ is point identified, its BCS and FCS are asymptotically identical due to the Bernstein-von Mises theorem. As a result, the BCS for θ will have a correct frequentist coverage probability asymptotically.

8 Financial Asset Pricing

8.1 The model

Asset pricing models state that the equilibrium price P_t^i of a financial asset i is equal to

$$P_t^i = E[M_{t+1}P_{t+1}^i|\mathcal{I}_t], \qquad i = 1, \dots, N$$

where P_{t+1}^i denotes the price of asset i at the period $(t+1)$, M_{t+1} is the stochastic discount factor (SDF hereafter) and $\mathcal{I}_t$ denotes the information set at time t. In vectorial form this rewrites as

$$\iota = E[M_{t+1}R_{t+1}|\mathcal{I}_t]$$

where ι is the N-dimensional vector of ones and R_{t+1} is the N-dimensional vector of gross asset returns at time $(t+1)$: $R_{t+1} = (r_{1,t+1}, \dots, r_{N,t+1})'$ with $r_{i,t+1} = P_{t+1}^i/P_t^i$. This model can be reinterpreted as a model of the SDF and may be used to detect the SDFs that are compatible with asset return data. Hansen and Jagannathan (1991) have obtained a lower bound on the volatility of SDFs that could be compatible with a given SDF-mean value and a given set of asset return data. Therefore, the set of SDFs M_{t+1} that can price existing assets generally form a proper set.

Let m and Σ denote, respectively, the vector of unconditional mean returns and covariance matrix of returns of the N risky assets, that is, $m = E(R_{t+1})'$ and $\Sigma = E(R_{t+1} - m)(R_{t+1} - m)'$. Denote $\mu = E(M_{t+1})$ and $\sigma^2 = Var(M_{t+1})$. We assume that m, Σ, μ and σ^2 do not vary with t. Hansen and Jagannathan (1991) show that the minimum variance $\sigma^2_\phi(\mu)$ achievable by a SDF with mean μ and compatible with the observed (m, Σ) is given by

$$\begin{aligned} \sigma^2_\phi(\mu) &= (\iota - \mu m)'\Sigma^{-1}(\iota - \mu m) =: \phi_1\mu^2 - 2\phi_2\mu + \phi_3 \\ \text{with } \phi_1 &= m'\Sigma^{-1}m, \qquad \phi_2 = m'\Sigma^{-1}\iota, \qquad \phi_3 = \iota'\Sigma^{-1}\iota. \end{aligned} \tag{8.1}$$

Therefore, an SDF correctly prices an asset only if, for given (m, Σ), its mean μ and variance σ^2 are such that $\sigma^2 \geq \sigma^2_\phi(\mu)$. An SDF's mean and variance (μ, σ^2) are said to be *admissible* if they satisfy this inequality and we define the set of admissible SDF's means and standard deviations as

$$\Theta(\phi) = \left\{(\mu, \sigma^2) \in \Theta;\ \sigma^2_\phi(\mu) - \sigma^2 \leq 0\right\} \tag{8.2}$$

where $\phi = (\phi_1, \phi_2, \phi_3)'$ and $\Theta \subset \mathbb{R}_+ \times \mathbb{R}_+$ is a compact set that we can choose based for instance on some prior knowledge. Usually, we can fix upper bounds $\bar{\mu} > 0$ and $\bar{\sigma} > 0$ as big as we want and take $\Theta = [0, \bar{\mu}] \times [0, \bar{\sigma}^2]$. In practice, $\bar{\mu}$ and $\bar{\sigma}$ must be chosen sufficiently large such that $\Theta(\phi)$ is non-empty. Making inference on $\Theta(\phi)$ allows to check whether a family of SDF (and then a given utility function) prices a financial asset correctly or not. Frequentist inference for this set is carried on in Chernozhukov, Kocatulum and Menzel (2012).

We develop a Bayesian approach. By using our previous notation we define $\theta = (\mu, \sigma^2)$ and

$$\Psi(\theta, \phi) = \phi_1\mu^2 - 2\phi_2\mu + \phi_3 - \sigma^2.$$

8.2 Support function

In this case $k = 1$ and $\Psi(\theta, \phi)$ is convex in θ. More precisely, $\Psi(\theta, \phi)$ is linear in σ^2 and strictly convex in μ (because Σ positive definite implies that $\phi_1 > 0$). Thus assumption 5.1 is verified. Assumption 5.3 (i) is also trivially satisfied. Moreover, $\Theta(\phi)$ is empty when $\bar{\sigma}^2 < \phi_1\bar{\mu}^2 - 2\phi_2\bar{\mu} + \phi_3$ for $\mu \in [0, \bar{\mu}]$. This happens in three cases: either (I) for $\phi_2^2 - \phi_1\phi_3 - \phi_1\bar{\sigma}^2 < 0$ or (II) for $\phi_2^2 - \phi_1\phi_3 - \phi_1\bar{\sigma}^2 > 0$ such that $\bar{\mu} < \frac{\phi_2}{\phi_1} \pm \frac{\phi_2^2 - \phi_1\phi_3 - \phi_1\bar{\sigma}^2}{\phi_1}$ or (III) for $\phi_2^2 - \phi_1\phi_3 - \phi_1\bar{\sigma}^2 > 0$ such that $\frac{\phi_2}{\phi_1} \pm \frac{\phi_2^2 - \phi_1\phi_3 - \phi_1\bar{\sigma}^2}{\phi_1} < 0$. Therefore, assumption 5.3 (ii) is verified for every ϕ such that (I), (II) and (III) do not hold. This is easily possible by taking $\bar{\mu}$ sufficiently large and $\bar{\sigma}^2$ not too large. Assumptions 5.3 (iii)-(v) and 5.4 are also satisfied.

In this example we can make inference on the support function of $\Theta(\phi)$ without requiring that assumption 5.5 (ii) hold. In fact, assumption 5.5 (i) holds for every $\phi \in \Phi$ and for every $p \in \mathbb{S}^2$ except for $p = (1, 0)$, $p = (-1, 0)$ and $p = (0, 1)$. For these values of p, however, it is easy to show that the support function is differentiable at ϕ_0 without assumption 5.5, see appendix A.2. Assumption 5.6 *(ii)* is trivially satisfied since $||\nabla_\phi\Psi(\theta, \phi_1) - \nabla_\phi\Psi(\theta, \phi_2)|| = 0$, assumption 5.6 *(iii)* is satisfied with $K = 1$ and Assumption 5.6 *(iv)* is true due to the continuity of $\Psi(\theta, \cdot)$ in ϕ. Assumption 5.6 *(i)* must be checked case by case (that is, for every region of values of p) since $\lambda(p, \phi)$ takes a different expression in each case, see appendix A.2.

Under assumption 5.1 we can rewrite

$$\begin{aligned}
\Xi(p,\phi) &= \arg\max_{\theta\in\Theta}\left\{p^T\theta;\ \Psi(\theta,\phi)\leq 0\right\} \\
&= \arg\max_{\theta\in\Theta}\left\{p_1\mu + p_2\sigma^2 - \lambda(p,\phi)(\phi_1\mu^2 - 2\phi_2\mu + \phi_3 - \sigma^2)\right\} \\
&= \arg\max_{0\leq\mu<\bar{\mu},\, 0<\sigma^2<\bar{\sigma}^2}\left\{p_1\mu + p_2\sigma^2 - \lambda(p,\phi)(\phi_1\mu^2 - 2\phi_2\mu + \phi_3 - \sigma^2)\right\}
\end{aligned}$$

where $p = (p_1, p_2)$, $\lambda_2 > 0$ and $\lambda_3 > 0$. The support function and $\Xi(p, \phi)$ have an explicit expression, but is very long and complicated. We present it in Appendix A.2.

8.3 Dirichlet process prior

Let F denote a probability distribution. The Bayesian model is $R_t|F \sim F$ and $\psi = (m, \Sigma) = \psi(F)$, where

$$\psi_1(F) = \int rF(dr), \quad \psi_2(F) = \int rr^T F(dr) - \int rF(dr)\int rF(dr)^T.$$

Let us impose a Dirichlet process prior for F, with parameter v_0 and base probability measure F_0 on $\mathbb{R}^N$. By Sethuraman (1994)'s decomposition, the Dirichet process prior induces a prior for ψ as: $m = \sum_{j=1}^{\infty}\alpha_j\xi_j$, and $\Sigma = \sum_{j=1}^{\infty}\alpha_j\xi_j\xi_j^T - \sum_{i=1}^{\infty}\alpha_i\xi_i\sum_{j=1}^{\infty}\alpha_j\xi_j^T$ where ξ_j are independently sampled from F_0; $\alpha_j = u_j\prod_{l=1}^{j}(1-u_l)$ with $\{u_i\}_{i=1}^n$ drawn from $\text{Beta}(1, v_0)$. These priors then induce a prior for ϕ. The posterior distribution for (m, Σ) can be calculated explicitly:

$$\begin{aligned}
\Sigma|D_n \sim\ & (1-\gamma)\sum_{j=1}^{\infty}\alpha_j\xi_j\xi_j^T + \gamma\sum_{t=1}^{n}\beta_t R_t R_t^n \\
& - \left((1-\gamma)\sum_{j=1}^{\infty}\alpha_j\xi_j + \gamma\sum_{t=1}^{n}\beta_t R_t\right)\left((1-\gamma)\sum_{j=1}^{\infty}\alpha_j\xi_j + \gamma\sum_{t=1}^{n}\beta_t R_t\right)^T,
\end{aligned}$$

$$m|D_n \sim (1-\gamma)\sum_{j=1}^{\infty}\alpha_j\xi_j + \gamma\sum_{t=1}^{n}\beta_t R_t, \quad \gamma\sim\text{Beta}(T, v_0), \quad \{\beta_j\}_{j=1}^n \sim Dir(1,...,1).$$

We can then simulate the posterior for ϕ based on the distributions of $\Sigma|D_n$, $m|D_n$ and (8.1).

8.4 Simulation

We present a simple simulated example. The returns R_t are assumed to follow a 2-factor model: $R_t = \Lambda f_t + u_t + 2\iota$, where Λ is a $N \times 2$ matrix of factor loadings. The error terms

$\{u_{it}\}_{i\leq N, t\leq n}$ are both cross sectionally and serially independent, and are uniform U$[-2,2]$. Besides, the components of Λ are standard normal, and the factors are also uniform U$[-2,2]$. The true $m = ER_t = 2\iota$, $\Sigma = \Lambda\Lambda' + I_N$. It is noted that in our DGP, the true likelihood is not Gaussian.

We set $N = 5, n = 200$. When calculate the posteriors, the DGP's distributions and the factor model structure are treated unknown, and we apply the nonparametric Dirichlet Process prior on the CDF of $R_t - m$, with parameter $v_0 = 3$, and based measure $F_0 = N(0,1)$. We use a uniform prior for (σ^2, μ), and obtain the posterior distributions for $(m, \Sigma, \phi_1, \phi_2, \phi_3, \sigma^2, \mu)$. More concretely, the prior is assumed to be:

$$\pi(\sigma^2, \mu|\phi) = \pi(\sigma^2|\phi, \mu)\pi(\mu); \quad \sigma^2|\phi, \mu \sim U[\sigma^2_\phi(\mu), \bar{\sigma}^2], \mu \sim U[0, \bar{\mu}],$$

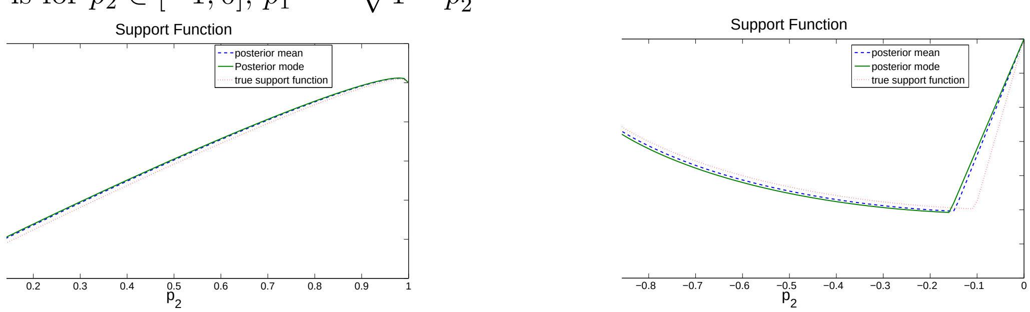
where μ and ϕ are a priori independent. We draw 1,000 times from the posterior of (ϕ, σ^2, μ). Each time we first draw (m, Σ) from their marginal posterior distributions, based on which obtain the posterior draw of ϕ from (8.1). In addition, draw μ uniformly from $[0, \bar{\mu}]$, and finally σ^2 uniformly from $[\sigma^2_\phi(\mu), \bar{\sigma}^2]$, where $\sigma^2_\phi(\mu)$ is calculated based on the drawn ϕ and μ.

The posterior mean $(\bar{\phi}_1, \bar{\phi}_2, \bar{\phi}_3)$ of ϕ is calculated, based on which we calculate an estimate of the boundary of the identified set (we set $\bar{\mu} = 1.4$ and $\bar{\sigma}^2 = 6$):

$$A_1 = \{\mu \in [0, \bar{\mu}], \sigma^2 \in [0, \bar{\sigma}^2] : \sigma^2 = \bar{\phi}_1\mu^2 - 2\bar{\phi}_2\mu + \bar{\phi}_3\}.$$

In addition, we estimate the support function $S_\phi(p)$ using either the posterior mean $\phi = \bar{\phi}$ or the posterior mode $\phi = \hat{\phi}_M$. The theoretical marginal posterior for ϕ is hard to compute. Thus to calculate the posterior mode, we first estimate the marginal posterior density for ϕ_i using kernel smoothing based on the draws $\{\phi_i\}_{i=1}^{1,000}$. The posterior mode $\hat{\phi}_M$ is then given by the values that maximizes the estimated marginal density. The support function $S_\phi(p)$ takes value for $p_1^2 + p_2^2 = 1$. In Figure 1, we plot the posterior estimates of the support function for two cases: $p_2 \in [0,1]$, $p_1 = \sqrt{1-p_2^2}$, and $p_2 \in [-1, 0]$, $p_1 = -\sqrt{1-p_2^2}$.

Figure 1: Posterior estimates of support function. Left panel is for $p_2 \in [0,1]$, $p_1 = \sqrt{1-p_2^2}$; right panel is for $p_2 \in [-1,0]$, $p_1 = -\sqrt{1-p_2^2}$

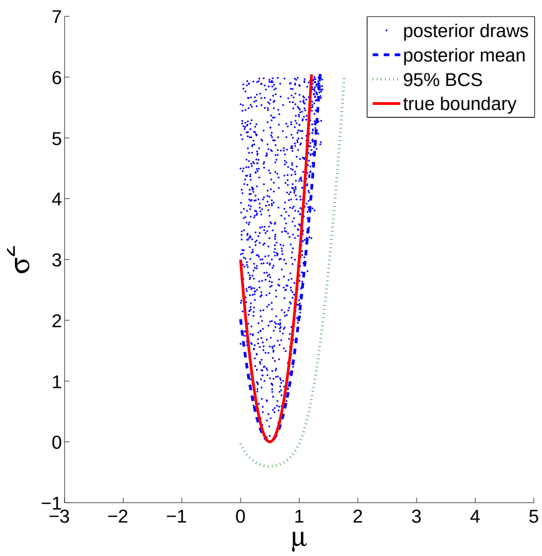


Using the approximate posterior model, we calculate the 95% posterior quantile q_τ for $J(\phi)$, based on which we construct the BCS $\Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$ for the identified set. The boundary of $\Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}$ is given by

$$A_2 = \left\{ \mu \in [0, \bar{\mu}], \sigma^2 \in [0, \bar{\sigma}^2] : \inf_z \sqrt{|z-\mu|^2 + |\sigma^2_{\hat{\phi}_M}(z) - \sigma^2|^2} = q_\tau/\sqrt{n} \right\}.$$

In Figure 2, we plot the posterior draws of (μ, σ^2), A_1, A_2 and the boundary of the true identified set.

Figure 2: 1,000 posterior draws of (μ, σ^2). Solid line is the boundary of the true identified set; dashed line represents the estimated boundary using the posterior mean; dotted line gives the 95% BCS.



9 Conclusion

We propose a semi-parametric Bayesian procedure for inference about partially identified models. Bayesian approaches are appealing in many aspects. Classical Bayesian approach in this literature has been assuming a parametric model, by specifying an ad-hoc parametric likelihood function. However, econometric models usually only identify a set of moment inequalities, and therefore assuming a known likelihood function suffers from the risk of misspecification, and may result in inconsistent estimations of the identified set. On the other hand, moment-condition based likelihoods such as the limited information and exponential tilted empirical likelihood, though guarantee the consistency, lack of probabilistic interpretations. Our approach thus only requires a set of moment conditions but still possesses a pure Bayesian interpretation.

Our analysis focuses on identified sets which are closed and convex. These sets are completely characterized by their support function, and efficient estimation of support function may lead to optimality of estimation and inference of the identified set. By imposing a prior on the support function, we construct its posterior distribution. It is shown that the support function for a very general moment inequality model admits a linear expansion, and the posterior is consistent. The Bernstein-von Mises theorem is proven.

Note that in this paper we consider a fixed data generating process (DGP). The constructed BCS has asymptotically correct coverage probability for any specific DGP, and the uniformity issue as in Andrews and Soares (2010) is not considered. The semi-parametric posterior concentration theory has been often developed for a specific DGP even when we have point identification, which relies on the existence of certain exponential tests and Schwartz' s theorem (see e.g., Wu and Ghosal 2008, Ghosh and Ramamoorthi 2003, Ghosal and van der Vaart 2001, Shen and Wasserman 2001). Besides, deriving the asymptotic representation of the support function for a fixed DGP is already technically involved. Extending these results uniformly in a class of DGP would be a challenging problem. We plan to address this issue in the future research.

A Support functions for two examples

A.1 Support function for the interval regression model

Consider Example 2.2. We now derive the support function for the identified set.

Lemma A.1. *Suppose ϕ_2^{-1} exists, then*

$$\Theta(\phi)=\left\{\theta\in\Theta:\theta=\phi_2^{-1}(\frac{\phi_1+\phi_3}{2}+u),u\in(-\frac{\phi_3-\phi_1}{2},\frac{\phi_3-\phi_1}{2})\right\}.$$

Proof. Defline $\xi=\phi_2\theta$. Then $\theta=\phi_2^{-1}\xi$. Let $u=\xi-\frac{\phi_1+\phi_3}{2}$ Then the identified set can be written as: $\Theta(\phi)=\{\phi_2^{-1}\xi:\phi_1\leq\xi\leq\phi_3\}=\left\{\phi_2^{-1}(\frac{\phi_1+\phi_3}{2}+u):\phi_1\leq u+\frac{\phi_1+\phi_3}{2}\leq\phi_3\right\}$. This then gives the result.

□

Now we are ready to calculate the support function for $\Theta(\phi)$.

Theorem A.1. *Suppose ϕ_2^{-1} exists. The support function for $\Theta(\phi)$ is given by:*

$$S_{\Theta(\phi)}(p)=p^T\phi_2^{-1}(\frac{\phi_1+\phi_3}{2})+\alpha_p^T(\frac{\phi_3-\phi_1}{2}),$$

where $d=\dim(\theta)$, $sgn(x)=I(x>0)-I(x<0)$,

$$\alpha_p=\begin{pmatrix}(p^T\phi_2^{-1})_1 sgn(p^T\phi_2^{-1})_1\\ \vdots\\ (p^T\phi_2^{-1})_d sgn(p^T\phi_2^{-1})_d\end{pmatrix}.$$

Proof. The proof is based on straightfoward calculations. Let $\Delta=(\phi_3-\phi_1)/2$, then

$$S_{\Theta(\phi)}(p)=\sup_{\theta\in\Theta(\phi)}p^T\theta=p^T\phi_2^{-1}(\frac{\phi_1+\phi_3}{2})+\sup_{-\Delta\leq u\leq\Delta}p^T\phi_2^{-1}u.$$

In addition,

$$\begin{aligned}\sup_{-\Delta\le u\le\Delta} p^T\phi_2^{-1}u &= \sup_{-\Delta\le u\le\Delta}\sum_{(p^T\phi_2^{-1})_i>0}(p^T\phi_2^{-1})_i u_i + \sum_{(p^T\phi_2^{-1})_i<0}(p^T\phi_2^{-1})_i u_i\\ &= \sum_{(p^T\phi_2^{-1})_i>0}(p^T\phi_2^{-1})_i\Delta_i - \sum_{(p^T\phi_2^{-1})_i<0}(p^T\phi_2^{-1})_i\Delta_i,\end{aligned} \tag{A.1}$$

which proves the theorem. □

A.2 Support function for the financial asset pricing model

Let $p \in \mathbb{S}^d$ and $\phi \in \Phi$, denote $p = (p_1, p_2)$ and $D = \phi_2^2 - \phi_1\phi_3$. Given the particular form that the parameters ϕ_1, ϕ_2 and ϕ_3 take in our example and by the Cauchy-Schwarz inequality we have: $\phi_2^2 \le \phi_1\phi_3$. This implies:

$$D = \phi_2^2 - \phi_1\phi_3 \le 0.$$

The support function is given by

$$S_\phi(p) = p^T\Xi(p,\phi), \quad \|p\| = 1.$$

Here $\Xi(p,\phi)$ is determined by:

1. for $p_2 > 0$, $p_1 > 0$:

$$\Xi(p,\phi) = \begin{cases} (\bar\mu, \bar\sigma^2) & \text{if } \bar\sigma^2 \ge \phi_1\bar\mu^2 - 2\phi_2\bar\mu + \phi_3 \\ \left(\frac{\phi_2+\sqrt{D+\phi_1\bar\sigma^2}}{\phi_1}, \bar\sigma^2\right) & \text{if } \bar\sigma^2 < \phi_1\bar\mu^2 - 2\phi_2\bar\mu + \phi_3. \end{cases}$$

2. for $p_2 < 0$, $p_1 < 0$:

$$\Xi(p,\phi) = \begin{cases} \left(\frac{\phi_2}{\phi_1}, \phi_3 - \frac{\phi_2^2}{\phi_1}\right), \text{ if } \phi_2 > 0, \text{ and } \frac{p_1}{p_2} \le \phi_2 I(\phi_3 \le \bar\sigma^2) + \sqrt{D+\phi_1\bar\sigma^2}I(\phi_3 > \bar\sigma^2) \\ \left(\frac{\phi_2-\sqrt{D+\phi_1\bar\sigma^2}}{\phi_1}I(\phi_3 > \bar\sigma^2), \phi_3 I(\phi_3 \le \bar\sigma^2) + \bar\sigma^2 I(\phi_3 > \bar\sigma^2)\right) \quad \text{otherwise.} \end{cases}$$

3. for $p_2 < 0$, $p_1 > 0$:

   3.1. $\Xi(p,\phi) = \left(\frac{\phi_2}{\phi_1} - \frac{p_1}{2p_2\phi_1}, \frac{p_1^2}{4p_2\phi_1} - \frac{\phi_2^2}{\phi_1} + \phi_3\right)$, if I and II below are satisfied:

   I. $2\phi_2 - 2\phi_1\bar\mu \le \frac{p_1}{p_2} < 2\phi_2$ and

   II. for $\mu = \frac{\phi_2}{\phi_1} - \frac{p_1}{2p_2\phi_1}$ one of the following two conditions is verified:

   II.a $D < 0$, $D + \phi_1\bar\sigma^2 \ge 0$ and $\frac{\phi_2-\sqrt{D+\phi_1\bar\sigma^2}}{\phi_1} \le \mu \le \frac{\phi_2+\sqrt{D+\phi_1\bar\sigma^2}}{\phi_1}$

   II.b $D = 0$ and $\frac{\phi_2-\sqrt{D+\phi_1\bar\sigma^2}}{\phi_1} \le \mu \le \frac{\phi_2+\sqrt{D+\phi_1\bar\sigma^2}}{\phi_1}$

   3.2. $\Xi(p,\phi) = \left(0, \phi_3 I(\phi_3 \le \bar\sigma^2) + \bar\sigma^2 I(\phi_3 > \bar\sigma^2)\right)$, if $\frac{p_1}{p_2} \ge 2\phi_2$;

3.3. $\Xi(p,\phi) = \left(\bar{\mu}, \phi_1\bar{\mu}^2 - 2\phi_2\bar{\mu} + \phi_3\right)$ if $2\phi_2 - 2\phi_1\bar{\mu} > \frac{p_1}{p_2}$ and either II.a or II.b above is satisfied for $\mu = \bar{\mu}$;

3.4. (imaginary solution) $\Xi(p,\phi) = \left(\frac{\phi_2+\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}, \bar{\sigma}^2\right)$ if $2\phi_2 - 2\phi_1\bar{\mu} > \frac{p_1}{p_2}$, $D < 0$ and $D + \phi_1\bar{\sigma}^2 < 0$;

3.6. $\Xi(p,\phi) = \left(\bar{\mu}, \bar{\sigma}^2\right)$ if $2\phi_2 - 2\phi_1\bar{\mu} > \frac{p_1}{p_2}$, $D + \phi_1\bar{\sigma}^2 \geq 0$ and either $\bar{\mu} < \frac{\phi_2-\sqrt{D+\bar{\sigma}^2\phi_1}}{\phi_1}$ or $\bar{\mu} > \frac{\phi_2+\sqrt{D+\bar{\sigma}^2\phi_1}}{\phi_1}$;

3.7. (imaginary solution) $\Xi(p,\phi) = \left(\frac{\phi_2+\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}, \bar{\sigma}^2\right)$ if I above is satisfied, $D < 0$ and $D + \phi_1\bar{\sigma}^2 < 0$;

3.8. $\Xi(p,\phi) = \left(\frac{\phi_2}{\phi_1} - \frac{p_1}{2p_2\phi_1}, \bar{\sigma}^2\right)$ if I above is satisfied, $D < 0$, $D + \phi_1\bar{\sigma}^2 \geq 0$ and either $\mu < \frac{\phi_2-\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}$ or $\mu > \frac{\phi_2+\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}$ for $\mu = \frac{\phi_2}{\phi_1} - \frac{p_1}{2p_2\phi_1}$;

3.9. $\Xi(p,\phi) = \left(\frac{\phi_2}{\phi_1} - \frac{p_1}{2p_2\phi_1}, \bar{\sigma}^2\right)$ if I above is satisfied, $D = 0$ and either $\mu < \frac{\phi_2-\sqrt{\phi_1\bar{\sigma}^2}}{\phi_1}$ or $\mu > \frac{\phi_2+\sqrt{\phi_1\bar{\sigma}^2}}{\phi_1}$ for $\mu = \frac{\phi_2}{\phi_1} - \frac{p_1}{2p_2\phi_1}$;

4. $p_2 > 0$, $p_1 < 0$: $\Xi(p,\phi) = \left(\frac{\phi_2-\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1} I(\phi_3 > \bar{\sigma}^2), \bar{\sigma}^2\right)$.

5. $p_2 = 0$, $p_1 = 1$:

$$\Xi(p,\phi) = \begin{cases} (\bar{\mu}, \sigma^2) & \forall \sigma^2 \in [\phi_1\bar{\mu}^2 - 2\phi_2\bar{\mu} + \phi_3, \bar{\sigma}^2] \text{ if } \bar{\sigma}^2 \geq \phi_1\bar{\mu}^2 - 2\phi_2\bar{\mu} + \phi_3 \\ \left(\frac{\phi_2+\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}, \bar{\sigma}^2\right) & \text{if } \bar{\sigma}^2 < \phi_1\bar{\mu}^2 - 2\phi_2\bar{\mu} + \phi_3. \end{cases}$$

6. $p_2 = 0$, $p_1 = -1$:

$$\Xi(p,\phi) = \begin{cases} (0, \sigma^2) & \forall \sigma^2 \in [\phi_3, \bar{\sigma}^2] \text{ if } \phi_3 < \bar{\sigma}^2 \\ \left(\frac{\phi_2-\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}, \bar{\sigma}^2\right) & \text{if } \phi_3 \geq \bar{\sigma}^2. \end{cases}$$

7. $p_2 = 1$, $p_1 = 0$: $\Xi(p,\phi) = (\mu, \bar{\sigma}^2)$, $\forall \mu \in [\max\left(0, \frac{\phi_2-\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}\right), \min\left(\bar{\mu}, \frac{\phi_2+\sqrt{D+\phi_1\bar{\sigma}^2}}{\phi_1}\right)]$.

8. $p_2 = -1$, $p_1 = 0$:

$$\Xi(p,\phi) = \begin{cases} \left(\frac{\phi_2}{\phi_1}, \phi_3 - \frac{\phi_2^2}{\phi_1}\right) & \text{if } \phi_2 \geq 0 \text{ and } \phi_1\phi_3 - \phi_2^2 > 0 \\ (0, \phi_3) & \text{if } \phi_2 < 0 \end{cases}$$

The linearization of the support function given in theorem 5.1 remains valid despite the fact that Assumption 5.5 is not satisfied for three values of $p \in \mathbb{S}^2$, that is, $p = (\pm 1, 0)$ and $p = (0, 1)$. Denote $\mathbb{S}_{ns} : \{(1,0), (-1,0), (0,1)\} \subset \mathbb{S}^2$. For $p \in \mathbb{S}^2 \backslash \mathbb{S}_{ns}$ the proof of the result in theorem 5.1 remains unchanged.

For $p \in \mathbb{S}_{ns}$ the proof is the same as the proof of theorem 5.1 except for the proof of some intermediate results which we now detail.

Proof of Lemma G.14: by using the notation in this lemma, we have to show that

$$\left.\frac{dS_{\phi_\tau}(p)}{d\tau+}\right|_{\tau=\tau_0} = \left.\frac{dS_{\phi_\tau}(p)}{d\tau-}\right|_{\tau=\tau_0}. \tag{A.2}$$

We refer to the expressions given in (G.14) and (G.15). For $p \in \mathbb{S}_{ns}$ then $\Xi(p, \phi_{\tau_0})$ is not a singleton. However, since $\nabla_\phi \Psi(\theta, \phi_\tau)$ does not depend on σ^2 and since for $p = (\pm 1, 0)$, $\Xi(p, \phi)$ is not a singleton only in the dimension of σ^2 then, we still get the equality (A.2). For $p = (0, 1)$, $\lambda(p, \phi_{\tau_0}) = 0$ so, by using (G.14) and (G.15), this implies that the equality (A.2) still hods.

Proof of Lemma G.13: the proof does not change except for the analysis of term $\mathcal{A}_2$ in CASE II. Let us start by considering $p = (1, 0)$ which corresponds to case 5 above. If $\Xi(p, \phi_0) = (\bar{\mu}, \sigma^2)$, $\forall \sigma^2 \in (\phi_{01}\bar{\mu}^2 - 2\phi_{02}\bar{\mu} + \phi_{03}, \bar{\sigma}^2]$, then the constraint is not binding so that $\lambda(p, \phi_0) = 0$ and $\mathcal{A}_2 = 0$. If we are in the other case, then $\Xi(p, \phi)$ is a singleton in μ. Due to this and to the continuity of $\Psi(\theta, \phi)$ in ϕ, the term $[\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)] = 0$ and $\mathcal{A}_2 = 0$. Proving that $\mathcal{A}_2 = 0$ for $p = (-1, 0)$ and $p = (0, 1)$ proceeds exactly in the same way and then it is omitted.

B Posterior Concentration for ϕ

Much of the literature on posterior concentration rate for Bayesian nonparametrics relies on the notion of *entropy cover number*, which we now define as follows. Recall that for i.i.d. data, the likelihood function can be written as $l_n(\phi, \eta) = \prod_{i=1}^n l(X_i; \phi, \eta)$, where $l(x; \phi, \eta)$ denotes the density of the sampling distribution. Let

$$G = \{l(\cdot; \phi, \eta) : \phi \in \Phi, \eta \in \mathcal{P}\}$$

be the family of likelihood functions. We assume $\mathcal{P}$ is a metric space with a norm $\|.\|_\eta$, which then induces a norm $\|.\|_G$ on G such that $\forall l(\cdot; \phi, \eta) \in G$,

$$\|l(\cdot; \phi, \eta)\|_G = \|\phi\| + \|\eta\|_\eta.$$

In the examples of intervel censoring data and interval regression, $l(x; \phi, \eta) = \eta(x - \phi)$ and $\|\eta\|_\eta = \|\eta\|_1 = \int |\eta(x)| dx$. Then in this case $\|l(., \phi, \eta)\|_G = \|\phi\| + \|\eta\|_1$. Let $B(l, \rho)$ denote a closed ball in G centered at $l \in G$ with radius ρ.

Define the entropy cover number $\mathcal{N}(\rho, G, \|.\|_G)$ to be the minimum number of balls with radius ρ needed to cover G. The importance of the entropy cover number on nonparametric Bayesian asymptotics has been realized for a long history. We refer the audience to Tihkomirov (1961) and van der Vaart and Wellner (1996) for good early references.

We first present the assumptions that are sufficient to derive the posterior concentration rate for the point identified ϕ. The first one is placed on the entropy cover number.

Assumption B.1. *Suppose for all n large enough,*

$$\mathcal{N}(n^{-1/2}(\log n)^{1/2}, G, \|.\|_G) \leq n.$$

This condition requires that the "model" G be not too big. Once condition holds, then for all $r_n \geq n^{-1/2}(\log n)^{1/2}$, $\mathcal{N}(r_n, G, \|.\|_G) \leq \exp(nr_n^2)$. Morever, it ensures the existence of certain tests as given in Lemma below, and hence it can be replaced by the test condition that are commonly used in the literature of posterior concentration rate, i.e., Jiang (2007), Ghosh and Ramamoorthi (2003). Same condition has been imposed by Ghosal et al. (2000) when considering Hellinger rates, and Bickel and Kleijn (2012) when considering semi-parametric posterior asymptotic normality, among others. When η_0 belongs to the family of location mixtures, this condition was verified by Ghosal et al. (1999, Theorem 3.1).

The next assumption places conditions on the prior for (ϕ, η). For each (ϕ, η), define

$$K_{\phi,\eta} = E\left[\log \frac{l(X;\phi_0,\eta_0)}{l(X;\phi,\eta)}\middle|\phi_0,\eta_0\right] = \int \log\left(\frac{l(x;\phi_0,\eta_0)}{l(x;\phi,\eta)}\right) l(x;\phi_0,\eta_0)dx$$

$$V_{\phi,\eta} = \mathrm{var}\left[\log \frac{l(X;\phi_0,\eta_0)}{l(X;\phi,\eta)}\middle|\phi_0,\eta_0\right] = \int \log^2\left(\frac{l(x;\phi_0,\eta_0)}{l(x;\phi,\eta)}\right) l(x;\phi_0,\eta_0)dx - K^2_{\phi,\eta}.$$

Assumption B.2. *The prior $\pi(\phi, \eta)$ satisfies:*

$$\pi\left(K_{\phi,\eta} \leq \frac{\log n}{n}, \quad V_{\phi,\eta} \leq \frac{\log n}{n}\right) n^M \to \infty$$

for some $M > 2$.

Intuitively, when (ϕ, η) is close to $(\phi_0.\eta_0)$, both $K_{\phi,\eta}$ and $V_{\phi,\eta}$ are close to zero. Hence this assumption requires the prior have sufficient amount of support around the true point identified parameters in terms of the Kullback-Leibler distance. Such a prior condition through the Kullback-Leibler neighborhood as $\max\{K_{\phi,\eta}, V_{\phi,\eta}\} \leq \frac{\log n}{n}$ has also been commonly imposed in the literature of semi-parametric posterior concentration, e.g., Ghosal et al. (1999 (2.10), 2000 Condition 2.4), Shen and Wasserman (2001, Theorem 2) and Bickel and Kleijn (2012, (3.13)). Moreover, it has been verified in the literature that the sieve prior (Shen and Wasserman 2001), Dirichlet mixture prior (Ghosal et al. 1999) and Normal mixture prior (Ghosal and van der Vaart 2007).

We are now ready to present the posterior concentration rate for ϕ.

Theorem B.1. *Suppose the data $X_1, ..., X_n$ are i.i.d. Under Assumptions B.1 and B.2, for some $C > 0$,*

$$P(\|\phi - \phi_0\| \leq Cn^{-1/2}(\log n)^{1/2}|D_n) \to^p 1.$$

The proof of this theorem requires two technical lemmas. The first is taken from Shen and Wasserman (2001).

Lemma B.1. *Under Assumption B.2,*

$$P_{D_n}\left(\iint \frac{l_n(\phi,\eta)}{l_n(\phi_0,\eta_0)}\pi(\phi,\eta)d\eta d\phi \geq \frac{1}{2n^2}\pi(K_{\phi,\eta} \leq \log n/n, V_{\phi,\eta} \leq \log n/n)\right) \to 1.$$

Proof. The proof follows the same argument of that of Lemma 1 in Shen and Wasserman (2001), and hence is omitted. □

The following lemma is regarding the existence of an exponential test, which is essential in establishing the posterior consistency and concentration rate in the nonparametric Bayesian literature. The idea of using the exponential test for posterior consistency dates by at least to Schwartz (1965).

For a function of the data $T(D_n)$, define

$$E_{\phi,\eta}T(D_n) = E[T(D_n)|\phi,\eta] = \int T(x)l_n(x;\phi,\eta)dx.$$

Lemma B.2. *Under Assumption B.1, there exists a test T and a constant $L > 4$ and $L \geq M + 2$ (for M defined in Assumption B.2) such that*
(i)

$$E_{\phi_0,\eta_0}T = o(1)$$

(ii) for $r_n = \sqrt{(\log n)/n}$,

$$\sup_{\eta\in\mathcal{P},\|\phi-\phi_0\|>Lr_n} E_{\phi,\eta}(1-T) \leq \exp\left(-\frac{9}{16}L^2nr_n^2\right).$$

Proof. For any natural number j, and some $L > 0$, define

$$H_j = \{l(.,\phi,\eta) \in G : \eta \in \mathcal{P}, jLr_n \leq \|\phi_0 - \phi\| \leq (j+1)Lr_n\}.$$

We cover H_j using N_j balls like: $B(\bar{g}, r) = \{l \in G : \|l-\bar{g}\|_G \leq r\}$ for some small $r = 4^{-1}jLr_n$ and center $\bar{g} \in G$. Then the H_j can be covered by N_j (to be characterized later) balls like $B(g_i, 4^{-1}jLr_n)$, with centers $g_{j1}, ..., g_{j,N_j} \in H_j$:

$$H_j \subset \cup_{i=1}^{N_j} B(g_{ji}, 4^{-1}jLr_n).$$

Let those centers be chosen such that N_j is the minimum number to make such a cover. Let $l_0 = l(.,\phi_0,\eta_0)$. Then for any ball $B(g_{ji}, 4^{-1}jLr_n)$, the center satisfies $\|g_{ji} - l_0\|_G^2 \geq \|\phi_{ji} - \phi_0\|^2 \geq j^2L^2r_n^2$. The last inequality follows since $g_{ji} \in H_j$. Then for any $l(.,\phi,\eta) \in B(g_{ji}, 4^{-1}jLr_n)$, $\|l - l_0\|_G \geq \|l_0 - g_{ji}\|_G - \|l - g_{ji}\|_G \geq jLr_n - 4^{-1}jLr_n = \frac{3}{4}jLr_n$. So we have shown that each element in the small ball $B(g_{ji}, 4^{-1}jLr_n)$ is $3jLr_n/4$ away from l_0, and due to the convexity of such a ball, by the standard minimax result (see Le Cam, 1986, Birgé, 1983), there exists a test T_{ji} such that

$$\begin{aligned}\max\left\{E_{\phi_0,\eta_0}T_{ji}, \sup_{l\in B(g_{ji},4^{-1}jLr_n)} E_{\phi,f}(1-T)\right\} &\leq \exp(-nd(l_0, B(g_{ji}, 4^{-1}jLr_n))^2)\\ &\leq \exp\left(-n\frac{9}{16}j^2L^2r_n^2\right),\end{aligned}$$

where $d(l_0, B) = \inf_{l\in B}\|l - l_0\|_G$, and we have shown that $d(l_0, B(g_{ji}, 4^{-1}jLr_n)) \geq \frac{3}{4}jLr_n$.

Now define

$$T = \sup_{j\geq 1} \max_{1\leq i\leq N_j} T_{ij}.$$

Then for any $l = l(., \phi, \eta)$ such that $\|\phi - \phi_0\| > Lr_n$, exists j^*, so that $l \in H_{j^*}$. By the cover, there exists $i^* \leq N_{j^*}$, and a ball so that $l \in B(g_{j^*i^*}, 4^{-1}j^*Lr_n)$. Due to $-T \leq -T_{ij}$ for any i, j, $E_{\phi,\eta}(1-T) \leq \sup_{g\in B(g_{j^*i^*},4^{-1}j^*Lr_n)} E_g(1-T) \leq \exp(-n\frac{9}{16}j^{*2}L^2r_n^2) \leq \exp(-n\frac{9}{16}L^2r_n^2)$. Hence

$$\sup_{\|\phi-\phi\|>Lr_n,\eta\in\mathcal{P}} E_{\phi,f}(1-T) \leq \exp\left(-n\frac{9}{16}L^2r_n^2\right).$$

This proves the second assertion (type II error) of the lemma.

For the first assertion (type I error), $E_{\phi_0,\eta_0}T \leq \sum_{j\geq 1}\sum_{i\leq N_j} ET_{ij} \leq \sum_j N_j \exp(-n\frac{9}{16}j^2L^2r_n^2)$. Note that $N_j = \mathcal{N}(4^{-1}jLr_n, H_j, \|.\|_G) \leq \mathcal{N}(4^{-1}jLr_n, G, \|.\|_G) \leq \mathcal{N}(r_n, G, \|.\|_G) \leq \exp(nr_n^2)$, where we used $L > 4$ so $4^{-1}jL \geq 1$, and the number of covers should be bigger if the radius is smaller. Hence

$$E_{\phi_0,\eta_0}T \leq \sum_j N_j \exp\left(-n\frac{9}{16}j^2L^2r_n^2\right) \leq \exp(nr_n^2)\sum_j \exp\left(-n\frac{9}{16}j^2L^2r_n^2\right) = o(1)$$

This is $o(1)$ since $L > 4$ and $nr_n^2 \to \infty$. □

Proof of Theorem B.1

Proof. Let $E = E_{f_0,\phi_0}$ be expectation operator with respect to the distribution of data, given the true parameters. For some $M > 0$, let U denote the ball centered at ϕ_0 with radius $Mn^{-1/2}(\log n)^{1/2}$. Denote U^c as the complement of U. Then It suffices to show that for some $M > 0$, $EP(\phi \in U^c|D_n) = o(1)$. In fact, for the test T in Lemma B.2,

$$EP(\phi \in U^c|D_n) = E[P(\phi \in U^c|D_n)T] + E[P(\phi \in U^c|D_n)(1-T)]$$

$$\leq ET + EP(\phi \in U^c|D_n)(1-T) = o(1) + EP(\phi \in U^c|D_n)(1-T).$$

The last equality follows from Lemma B.2(i). Let

$$\beta_n = \pi\left(K_{\phi,\eta} \leq \frac{\log n}{n}, \quad V_{\phi,\eta} \leq \frac{\log n}{n}\right)\frac{1}{2n^2},$$

and define an event

$$A = \iint \frac{l_n(\phi,\eta)}{l_n(\phi_0,\eta_0)}\pi(\phi,\eta)d\phi d\eta \geq \beta_n.$$

Then Lemma B.1 shows that $P_{D_n}(A) \to 1$. In addition, $\exp(-Lnr_n^2)\beta_n^{-1} = o(1)$ for $L \geq M+2$ in Lemma B.2, by Assumption B.2. Then

$$E[P(\phi \in U^c|D_n)(1-T)] = E[P(\phi \in U^c|D_n)(1-T)I_A] + E[P(\phi \in U^c|D_n)(1-T)I_{A^c}]$$

$$\leq EP(\phi \in U^c|D_n)(1-T)I_A + 2EI_{A^c} \leq EP(\phi \in U^c|D_n)(1-T)I_A + o(1).$$

The last equality follows from $P_{D_n}(A) \to 1$.

It remains to upper bound $EP(\phi \in U^c|D_n)(1-T)I_A$. We need to lower bound the denominator of the posterior probability, and upper bound the numerator as well. Because of I_A, the lower bound of denominator can be realized on A. Then

$$E[P(\phi \in U^c|D_n)(1-T)I_A] \leq \frac{1}{\beta_n} E\left\{ \iint_{U^c\times\mathcal{P}} \prod_{i=1}^{n} \frac{l(X_i;\phi,\eta)}{l(X_i;\phi_0,\eta_0)} \pi(d\eta,d\phi)(1-T) \right\}$$

$$= \beta_n^{-1} \iiint_{\mathcal{X}\times U^c\times\mathcal{P}} \prod_{i=1}^{n} \frac{l(X_i;\phi,\eta)}{l(X_i;\phi_0,\eta_0)} (1-T) \prod_{i=1}^{n} l(X_i;\phi_0,\eta_0)\pi(d\eta,d\phi)dX_1...dX_n$$

$$= \beta_n^{-1} \iiint_{\mathcal{X}\times U^c\times\mathcal{P}} \prod_{i=1}^{n} l(X_i;\phi,\eta)(1-T)\pi(d\eta,d\phi)dX_1...dX_n$$

Here I used the fact that $EV = E_{\phi_0,\eta_0}V = \int V \prod_i l(x_i;\phi_0,\eta_0)dx_1...dx_n$. Also note that $\int V \prod_i l(X_i;\phi,\eta)dx_1...dx_n = E_{\phi,\eta}V$. Using the Fubini's theorem by changing the integration order, we have the expression above also equals:

$$\beta_n^{-1} \iiint_{\mathcal{X}\times U^c\times\mathcal{P}} \prod_{i=1}^{n} l(X_i;\phi,\eta)(1-T)\pi(df,d\phi)dX_1...dX_n = \beta_n^{-1} \iint_{U^c\times\mathcal{P}} E_{\phi,\eta}(1-T)\pi(d\eta,d\phi)$$

$$\leq \beta_n^{-1}\pi(\phi \in U^c) \sup_{\phi\in U^c,\eta\in\mathcal{P}} E_{\phi,\eta}(1-T) \leq \exp(-Lnr_n^2)\beta_n^{-1} = o(1).$$

C Proofs for Section 3

C.1 Proof of Theorem 3.1

In this proof we use the notation $\iota\epsilon$ to denote: ϵ if $\iota = 1$ and $-\epsilon$ if $\iota = -1$. We start by stating and proving the following lemma.

Lemma C.1. *Let $\pi(\theta|\phi)$ be a regular conditional distribution. Under assumption 3.2 there exists a sequence of bounded and continuous functions $\{h_{m,\iota}(\phi)\}_m$ defined on Φ for $\iota \in \{-1,1\}$ such that $|h_{m,\iota}(\phi)| \leq 1$ and*

$$\pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi) = \lim_{m\to\infty} h_{m,\iota}(\phi), \qquad \pi(\phi) - a.s.$$

for $\iota \in \{-1,1\}$.

Proof. Denote by $\mathcal{C}(\Phi)$ the set of continuous function on Φ. Since $\pi(\theta|\phi)$ is a regular conditional distribution then there exists a transition probability from $(\Phi,\mathfrak{B}_\phi)$ to $(\Theta,\mathfrak{B}_\theta)$ that characterizes it, where $\mathfrak{B}_\phi$ and $\mathfrak{B}_\theta$ denote the σ-fields associated with Φ and Θ, respectively. This means that $\pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi)$ is a measurable function of ϕ for $\iota \in \{-1,1\}$.

Next, remark that if $\phi \notin A_{\epsilon,\iota}$ then $\Theta(\phi_0)^{\iota\epsilon}$ is not supported by the conditional prior distribution $\pi(\theta|\phi)$. Therefore, $\pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi) = 0$, $\forall\phi \notin A_{\epsilon,\iota}$. It follows by the Lusin's theorem (see *e.g.* Rudin (1986) page 55) that, on a compact set $K_\iota \subset (\Phi,\mathfrak{B}_\phi)$ of almost full $\pi(\phi)$-probability, $\pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi)$ is equal to a continuous function of ϕ. Finally, since for any $\phi \in \Phi$, $|\pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi)| \leq 1$, by the corollary page 56 in Rudin (1986), there exists a sequence $h_{m,\iota} \in \mathcal{C}(\Phi)$, $|h_{m,\iota}| \leq 1$ such that $\pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi) = \lim_{m\to\infty} h_{m,\iota}(\phi)$, $\pi(\phi)$-a.s.

□

Under assumption 3.2 and by lemma C.1 we can apply the *Dominated Convergence Theorem* so that $\lim_{m\to\infty} h_{m,\iota}(\phi) = \pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi)$, $\pi(\phi)$-a.s. implies

$$\lim_{m\to\infty} \int_\Phi h_{m,\iota}(\phi)\pi(\phi)d\phi = \int_\Phi \lim_{m\to\infty} h_{m,\iota}(\phi)\pi(\phi)d\phi.$$

We consider $P(\theta \in \Theta(\phi_0)^{\iota\epsilon}|D_n)$ and show that it converges to 1 for $\iota = 1$ and to something smaller than 1 for $\iota = -1$. Under assumption 3.2 and by lemma C.1, this probability can be rewritten as

$$\begin{aligned} P(\theta \in \Theta(\phi_0)^{\iota\epsilon}|D_n) &= \int_\Phi \pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi)\pi(\phi|D_n)d\phi = \int_\Phi \lim_{m\to\infty} h_{m,\iota}(\phi)\pi(\phi|D_n)d\phi \\ &= \lim_{m\to\infty} \int_\Phi h_{m,\iota}(\phi)\pi(\phi|D_n)d\phi \qquad \text{(C.1)} \end{aligned}$$

We analyse separately the case with a nonparametric prior and the case with a semi-parametric prior.

Nonparametric prior. In this case, assumption 3.1 *(i)* holds. The expression in (C.1) must be developed further:

$$P(\theta \in \Theta(\phi_0)^{\iota\epsilon}|D_n) = \lim_{m\to\infty} \int_\mathcal{F} h_{m,\iota}(\phi(F))\pi(F|D_n)dF.$$

Therefore, since ϕ is a continuous function of F (by the first part of assumption 3.1 *(i)*) we have that the composed function $h_{m,\iota} \circ \phi$ is a continuous and bounded function of F and under the second part of assumption 3.1 *(i)* we obtain

$$\begin{aligned} &\lim_{n\to\infty} P(\theta \in \Theta(\phi_0)^{\iota\epsilon}|D_n) = \lim_{n\to\infty}\lim_{m\to\infty} \int_\mathcal{F} h_{m,\iota}(\phi(F))\pi(F|D_n)dF \\ =\;& \lim_{m\to\infty} \int_\mathcal{F} h_{m,\iota}(\phi(F)) \lim_{n\to\infty} \pi(F|D_n)dF = \lim_{m\to\infty} \int_\mathcal{F} h_{m,\iota}(\phi(F))\delta_{F_0}(dF) \\ =\;& \lim_{m\to\infty} h_{m,\iota}(\phi(F_0)) = \pi(\theta \in \Theta(\phi_0)^{\iota\epsilon}|\phi(F_0)) \end{aligned}$$

where δ_{F_0} denotes the Dirac mass in F_0. Since $\pi(\theta|\phi(F_0))$ has support equal to $\Theta(\phi_0)$ and $\Theta(\phi_0)^\epsilon \subset \Theta(\phi_0) \subset \Theta(\phi_0)^\epsilon$ then by using assumption 3.3

$$\lim_{n\to\infty} P(\theta \in \Theta(\phi_0)^\epsilon|D_n) = 1 \quad \text{and} \quad \lim_{n\to\infty} P(\theta \in \Theta(\phi_0)^{-\epsilon}|D_n) < 1, \quad F_0 - a.s.$$

Semi-parametric prior. In this case, assumption 3.1 *(ii)* holds. Since $h_{m,\iota}(\cdot)$ is a continuous function of ϕ, then equation (C.1) and assumption 3.1 *(ii)* imply

$$\begin{aligned} & \lim_{n\to\infty} P(\theta\in\Theta(\phi_0)^{\iota\epsilon}|D_n) = \lim_{n\to\infty}\lim_{m\to\infty}\int_\Phi h_{m,\iota}(\phi)\pi(\phi|D_n)d\phi \\ = & \lim_{m\to\infty}\int_\Phi h_{m,\iota}(\phi)\lim_{n\to\infty}\pi(\phi|D_n)d\phi = \lim_{m\to\infty}\int_\Phi h_{m,\iota}(\phi)\delta_{\phi_0}(d\phi) \\ = & \lim_{m\to\infty} h_{m,\iota}(\phi_0) = \pi(\theta\in\Theta(\phi_0)^{\iota\epsilon}|\phi_0) \end{aligned}$$

where δ_{ϕ_0} denotes the Dirac mass in ϕ_0. Since $\pi(\theta|\phi_0)$ has support equal to $\Theta(\phi_0)$ then by using assumption 3.3

$$\lim_{n\to\infty} P(\theta\in\Theta(\phi_0)^{\epsilon}|D_n) = 1 \quad \text{and} \quad \lim_{n\to\infty} P(\theta\in\Theta(\phi_0)^{-\epsilon}|D_n) < 1, \quad F_0 - a.s.$$

This concludes the proof.

D Proofs for Section 4

D.1 Proof of Theorem 4.1

Define $Q(\theta,\phi) = \|\max(\Psi(\theta,\phi),0)\| = \left[\sum_{i=1}^k (\max(\Psi_i(\theta,\phi),0))^2\right]^{1/2}$.

Lemma D.1. *There exists $C>0$, for any $\phi_1,\phi_2\in\Phi$,*

$$\sup_{\theta\in\Theta} |Q(\theta,\phi_1) - Q(\theta,\phi_2)| \le C\|\phi_1-\phi_2\|.$$

Proof. Define $f(x) = xI(x\ge 0)$, where $x\in\mathbb{R}$. It is straightforward to show that $\forall x_1,x_2$, $|f(x_1)-f(x_2)|\le |x_1-x_2|$. One the other hand, for any $\phi_1,\phi_2\in\Phi$,

$$\begin{aligned} & |Q(\theta,\phi_1) - Q(\theta,\phi_2)| = |\|\max(\Psi(\theta,\phi_1),0)\| - \|\max(\Psi(\theta,\phi_2),0)\|| \\ \le & \|\max(\Psi(\theta,\phi_1),0) - \max(\Psi(\theta,\phi_2),0)\| \\ = & \left(\sum_{i=1}^d [\max(\Psi_i(\theta,\phi_1),0) - \max(\Psi_i(\theta,\phi_2),0)]^2\right)^{1/2} \\ = & \left(\sum_{i=1}^d [f(\Psi_i(\theta,\phi_1)) - f(\Psi_i(\theta,\phi_2))]^2\right)^{1/2} \le \left(\sum_{i=1}^d [\Psi_i(\theta,\phi_1) - \Psi_i(\theta,\phi_2)]^2\right)^{1/2} \\ = & \|\Psi\theta,\phi_1) - \Psi(\theta,\phi_2)\| \le C\|\phi_1-\phi_2\| \end{aligned} \tag{D.1}$$

where C does not depend on θ, by Assumption 4.3.

□

Lemma D.2. *There exists a closed neighborhood $U(\phi_0)$, for any $a_n = O(1)$, there exists $K>0$ that does not depend on ϕ, so that*

$$\inf_{\phi\in U(\phi_0)} \inf_{d(\theta,\Theta(\phi))\ge Ka_n} \max_{i\le k} \Psi_i(\theta,\phi) > a_n.$$

Proof. For any $C > 0$, define $A_C = \{\phi \in U(\phi_0) : \inf_{\theta: d(\theta,\Theta(\phi)) \geq Ca_n} \max_{i\leq k} \Psi_i(\theta,\phi) > a_n\}$. Then by Assumption 4.4, $\forall \phi \in U(\phi_0)$, there exists $C_\phi > 0$ so that $\phi \in A_{C_\phi}$. Thus, we have $U(\phi_0) \subset \cup_{\phi\in U(\phi_0)} A_{C_\phi}$. Since $U(\phi_0)$ is a closed neighborhood inside $\mathbb{R}^{\dim(\phi)}$, which is compact, hence there exist constants $C_1, ..., C_N$ for some finite $N > 0$ to form a finite cover so that $U(\phi_0) \subset \bigcup_{i=1}^N A_{C_i}$. Then $\forall \phi \in U(\phi_0)$, there exists $j \leq N$ so that $\phi \in A_{C_j}$, which is $\inf_{\theta: d(\theta,\Theta(\phi)) \geq C_j a_n} \max_{i\leq d} \Psi_i(\theta,\phi) > a_n$. On the other hand, let $K = \max\{C_i : i \leq N\}$, then

$$\inf_{\theta: d(\theta,\Theta(\phi)) \geq Ka_n} \max_{i\leq k} \Psi_i(\theta,\phi) \geq \inf_{\theta: d(\theta,\Theta(\phi)) \geq C_j a_n} \max_{i\leq k} \Psi_i(\theta,\phi) > a_n.$$

This is true for any $\phi \in U(\phi_0)$. Hence the result follows. □

Lemma D.3. *For any $M > 0$, there exists $\delta > 0$, and a neighborhood $U(\phi_0)$ of ϕ_0, so that*

$$\inf_{\phi\in U(\phi_0)} \inf_{d(\theta,\Theta(\phi)) \geq \delta\sqrt{(\log n)/n}} Q(\theta,\phi) > M\sqrt{\frac{\log n}{n}}.$$

Proof. For any $M > 0$, by Lemma D.2, there exist $U(\phi_0)$ and $\delta > 0$ so that

$$\inf_{\phi\in U(\phi_0)} \inf_{d(\theta,\Theta(\phi)) \geq \delta\sqrt{\log n/n}} \max_{i\leq d} \Psi_i(\theta,\phi) > M\sqrt{\frac{\log n}{n}}. \tag{D.2}$$

Now for any $(\theta,\phi) \in \{(\theta,\phi) \in \Theta \times U(\phi_0) : d(\theta,\Theta(\phi)) \geq \delta\sqrt{\log n/n}\}$, since $\theta \notin \Theta(\phi)$, $\max_{i\leq k} \Psi_i(\theta,\phi) > 0$, which implies that $\max_{i\leq k} \Psi_i(\theta,\phi) = \max_{i\leq k} \Psi_i(\theta,\phi) I(\Psi_i(\theta,\phi) > 0)$.

For notational simplicity, let $\Psi_i = \Psi_i(\theta,\phi)$, and $\Psi = (\Psi_1, ..., \Psi_k)^T$. Then using the fact that $\max_i A_i^2 = (\max_i A_i)^2$ if $A_i \geq 0$, we have,

$$\begin{aligned} Q(\theta,\phi) = \|\max(\Psi,0)\| &= \left(\sum_{i=1}^k [\max(\Psi_i,0)]^2\right)^{1/2} \geq \left(\max_{i\leq k}[\max(\Psi_i,0)]^2\right)^{1/2} \\ = \quad &\left([\max_{i\leq k}\max(\Psi_i,0)]^2\right)^{1/2} = \max_{i\leq k}\max(\Psi_i,0) = \max_{i\leq k}\Psi_i I(\Psi_i \geq 0) = \max_{i\leq k}\Psi_i(\theta,\phi). \end{aligned}$$

The result follows immediately from (D.2). □

To simplify our notation, let us define

$$r_n = \sqrt{\frac{\log n}{n}}.$$

Lemma D.4. *There exists a constant $C > 0$ so that*

$$P(\Theta(\phi) \subset \Theta(\phi_0)^{Cr_n} | D_n) \to^p 1,$$

where $\Theta(\phi_0)^{Cr_n} = \{\theta \in \Theta : d(\theta,\Theta(\phi_0)) \leq Cr_n\}$, and $P(.|D_n)$ denotes the marginal posterior probability of ϕ.

Proof. For any $C > 0$, let $\Theta\backslash\Theta(\phi_0)^{Cr_n} = \{\theta \in \Theta : d(\theta, \Theta(\phi_0)) > Cr_n\}$. Suppose it is true that there exists $C > 0$ so that

$$P\left(\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi) > \sup_{\theta\in\Theta(\phi)} Q(\theta,\phi)\middle| D_n\right) \to^p 1, \tag{D.3}$$

then the lemma holds, this is because on the event $\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi) > \sup_{\theta\in\Theta(\phi)} Q(\theta,\phi)$, we have $\Theta(\phi) \subset \Theta(\phi_0)^{Cr_n}$. Thus it suffices to show (D.3). Note that $\sup_{\theta\in\Theta(\phi)} Q(\theta,\phi) = 0$, since $\forall\theta \in \Theta(\phi)$, $g(\theta,\phi) \leq 0$, which is equivalent to $Q(\theta,\phi) = 0$. On the other hand, we have $P(\|\phi - \phi_0\| < r_n|D_n) \to^p 1$. Therefore, it remains to show

$$P\left(\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi) > 0\middle| D_n\right) \to^p 1. \tag{D.4}$$

In fact, for any ϕ so that $\|\phi - \phi_0\| \leq r_n$, by Lemma D.1, there exists $K > 0$, for any $C > 0$,

$$\begin{aligned}\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi) &\geq \inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi_0) - \sup_{\theta\in\Theta}|Q(\theta,\phi) - Q(\theta,\phi_0)| \\ \geq \quad &\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi_0) - Kr_n.\end{aligned} \tag{D.5}$$

Now by Lemma D.3, there exists $C > 0$ such that

$$\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi_0) = \inf_{d(\theta,\Theta(\phi_0))\geq Cr_n} Q(\theta,\phi_0) \geq 3Kr_n.$$

Hence we have shown that whenever $\|\phi - \phi_0\| \leq r_n$, $\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi) \geq 2Kr_n > 0$. Therefore, by the posterior concentration for ϕ, (D.4) holds from

$$P\left(\inf_{\theta\in\Theta\backslash\Theta(\phi_0)^{Cr_n}} Q(\theta,\phi) > 0\middle| D_n\right) \geq P\left(\|\phi - \phi_0\| \leq r_n | D_n\right) \to^p 1.$$

□

Lemma D.5. *There exists $L > 0$ so that $P(\Theta(\phi_0) \subset \Theta(\phi)^{Lr_n}|D_n) \to^p 1$.*

Proof. By Lemma D.1, there exists $K > 0$ so that whenever $\|\phi - \phi_0\| \leq r_n$,

$$\sup_{\theta\in\Theta}|Q(\theta,\phi) - Q(\theta,\phi_0)| \leq Kr_n.$$

We now fix such a ϕ, then for all large enough n, $\phi \in U(\phi_0)$ where $U(\phi_0)$ is the neighborhood defined in Lemma D.3. For such a K, by Lemma D.3, there exists $L > 0$ that does not depend on ϕ (since the following inequality holds uniformly for $\phi \in U(\phi_0)$ by Lemma D.3),

$$\inf_{d(\theta,\Theta(\phi))\geq Lr_n} Q(\theta,\phi) > Kr_n,$$

which then implies that $\{\theta : Q(\theta,\phi) \leq Kr_n\} \subset \{\theta : d(\theta,\Theta(\phi)) \leq Lr_n\}$. On the other hand, for any $\theta \in \Theta(\phi_0)$, $Q(\theta,\phi_0) = 0$, which implies $Q(\theta,\phi) \leq 0 + |Q(\theta,\phi) - Q(\theta,\phi_0)| \leq Kr_n$.

Therefore, $\Theta(\phi_0) \subset \{\theta : Q(\theta, \phi) \leq Kr_n\} \subset \{\theta : d(\theta, \Theta(\phi)) \leq Lr_n\}$. Hence we have in fact shown that, the event $\|\phi - \phi_0\| \leq r_n$ implies the event $\Theta(\phi_0) \subset \Theta(\phi)^{Lr_n}$. Moreover, the event $\|\phi - \phi_0\| \leq r_n$ occurs with probability approaching one under the posterior distribution of ϕ, which then implies the result. □

Lemma D.6. *For two sets A, B, if $A \subset B^{r_1}$ and $B \subset A^{r_2}$ for some r_1, r_2, then*

$$d_H(A, B) \leq \max\{r_1, r_2\}.$$

Proof. $d_H(A, B) = \max\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\}$. Then $\forall a \in A$, since $A \subset B^{r_1}$, $a \in B^{r_1}$, which is $d(a, B) \leq r_1$. This simplies $\sup_{a\in A} d(a, B) \leq r_1$. Similarly we can show $\sup_{b\in B} d(b, A) \leq r_2$. □

Proof of Theorem 4.1
Theorem 4.1 follows from combining Lemmas D.4-D.6. Q.E.D.

E Proofs for Section 5

Lemma E.1. *If $\Psi(\theta, \phi_0)$ contains a subvector of strictly convex functions $\Psi_S(., \phi_0)$ of θ, then the function $\theta \mapsto \Psi_S(\theta, \phi)$ is strictly convex for all $\phi \in B(\phi_0, \delta)$.*

Proof. We fix a constant $\delta > 0$ to be determined later. Then for any $\phi \in B(\phi_0, \delta)$, $\lambda \in [0, 1]$ and $\theta_1, \theta_2 \in \Theta$, we want to show $\Psi_S(\theta_1\lambda + (1-\lambda)\theta_2, \phi) < \lambda\Psi_S(\theta_1, \phi) + (1-\lambda)\Psi_S(\theta_2, \phi)$. In fact, since $\Psi_S(\theta, \phi_0)$ is strictly convex in θ, there is $\epsilon_0 > 0$ such that $\Psi_S(\theta_1\lambda + (1-\lambda)\theta_2, \phi_0) < \lambda\Psi_S(\theta_1, \phi_0) + (1-\lambda)\Psi_S(\theta_2, \phi_0) - \epsilon_0$. Then by the continuity of $\phi \to \Psi_S(\theta, \phi)$ at ϕ_0, there is $\delta > 0$ such that whenever $\|\phi - \phi_0\| < \delta$, we have

$$\Psi_S(\theta_1, \phi_0) < \Psi_S(\theta_1, \phi) + \epsilon_0/3, \quad \Psi_S(\theta_2, \phi_0) < \Psi_S(\theta_2, \phi) + \epsilon_0/3$$

and $\Psi_S(\theta_1\lambda + (1-\lambda)\theta_2, \phi) < \Psi_S(\theta_1\lambda + (1-\lambda)\theta_2, \phi_0) + \epsilon_0/3$. Therefore,

$$\begin{aligned} &\Psi_S(\theta_1\lambda + (1-\lambda)\theta_2, \phi) < \Psi_S(\theta_1\lambda + (1-\lambda)\theta_2, \phi_0) + \epsilon_0/3 \\ < \quad &\lambda\Psi_S(\theta_1, \phi_0) + (1-\lambda)\Psi_S(\theta_2, \phi_0) - \epsilon_0 + \epsilon_0/3 < \lambda\Psi_S(\theta_1, \phi) + (1-\lambda)\Psi_S(\theta_2, \phi). \end{aligned}$$

□

E.1 Proof of Theorem 5.1

For any $\tau \in [0, 1]$ and any $\phi_1, \phi_2 \in B(\phi_0, r_n)$, define $\phi_\tau = \tau\phi_1 + (1-\tau)\phi_2$ with $\phi_2 = \phi_\tau|_{\tau=0}$ and $\phi_1 = \phi_\tau|_{\tau=1}$. For every $p \in \mathbb{S}^d$ the support function may be rewritten as a function of τ: $S_{\phi_{(\cdot)}}(p) : [0, 1] \to \mathbb{R}$. By lemma G.14 the support function $S_{\phi_\tau}(p)$ is differentiable at $\tau = \tau_0 \in (0, 1)$ and then we can apply the mean value theorem to $S_{\phi_\tau}(p)$:

$$S_{\phi_1}(p) - S_{\phi_2}(p) = \frac{\partial}{\partial\tau} S_{\phi_\tau}(p)\bigg|_{\tau=\tau_0\in(0,1)}. \tag{E.1}$$

By defining $\tau_0 : \mathbb{S}^d \to (0,1)$ a measurable and differentiable function of p and by using the result of Lemma G.14 we obtain

$$\frac{\partial}{\partial \tau} S_{\phi_\tau}(p)\bigg|_{\tau=\tau_0(p)} = \lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)})[\phi_1 - \phi_2] \tag{E.2}$$

for some $\tilde{\theta}(p) \in \Xi(p, \phi_{\tau_0(p)})$, $p \in \mathbb{S}^d$. By plugging (E.2) in (E.1) and developing further we obtain

$$\begin{aligned} S_{\phi_1}(p) - S_{\phi_2}(p) &= \lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)})[\phi_1 - \phi_2] \\ &= \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)[\phi_1 - \phi_2] \qquad \text{(E.3)} \\ &\quad + \Big(\lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)}) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)\Big)[\phi_1 - \phi_2] \end{aligned}$$

where $\theta_* : \mathbb{S}^d \to \Theta$ is a Borel measurable mapping satisfying $\theta_*(p) \in \Xi(p, \phi_0)$. Denote

$$f(\phi_1, \phi_2) = \sup_{p \in \mathbb{S}^d} \Big(\lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)}) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)\Big)[\phi_1 - \phi_2].$$

By lemma G.15, $\frac{f(\phi_1,\phi_2)}{\|\phi_1-\phi_2\|}$ converges to 0 uniformly in $(\phi_1, \phi_2) \in B(\phi_0, r_n)$ as $r_n \to 0$.

Finally, we analyze the first term in E.3. By lemma G.9 and lemma G.10, the function $p \mapsto \lambda(p, \phi_0)$ is continuous in $p \in \mathbb{S}^d$ and therefore it attains its supremum. Moreover, $\sup_{p \in \mathbb{S}^d} \nabla_\phi \Psi(\theta_*(p), \phi_0) \leq \sup_{\theta \in \Theta} \nabla_\phi \Psi(\theta, \phi_0)$ and the supremum is attained since, under assumptions 5.2 and 5.3 (i), $\theta \mapsto \nabla_\phi \Psi(\theta, \phi_0)$ is uniformly continuous on Θ. We have

$$\begin{aligned} &\sup_{p \in \mathbb{S}^d} \left|(S_{\phi_1}(p) - S_{\phi_2}(p)) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)[\phi_1 - \phi_2]\right| \qquad \text{(E.4)} \\ &\quad = \sup_{p \in \mathbb{S}^d} \left|\Big(\lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)}) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)\Big)[\phi_1 - \phi_2]\right| \\ &\quad =: f(\phi_1, \phi_2). \qquad \text{(E.5)} \end{aligned}$$

E.2 Proof of Theorem 5.2

Proof. Denote $r_n = (\log n)^{1/2} n^{-1/2}$ and $\Omega = \{\phi \in B(\phi_0, r_n)\}$. Under assumption 4.1: $P(\Omega^c | D_n) = o_p(1)$. Then, by using the expansion of the support function given in lemma 5.1

we have

$$
\begin{aligned}
P\Big(\sup_{p\in\mathbb{S}^d}|S_\phi(p)-S_{\phi_0}(p)|\geq Cr_n\Big|D_n\Big) &= P\Big(\{\sup_{p\in\mathbb{S}^d}|S_\phi(p)-S_{\phi_0}(p)|\geq Cr_n\}\cap\Omega\Big|D_n\Big)\\
&\quad + P\Big(\sup_{p\in\mathbb{S}^d}|S_\phi(p)-S_{\phi_0}(p)|\geq Cr_n\cap\Omega^c\Big|D_n\Big)\\
&\leq P\Big(\sup_{p\in\mathbb{S}^d}|S_\phi(p)-S_{\phi_0}(p)|\geq Cr_n\cap\Omega\Big|D_n\Big)+P(\Omega^c|D_n)\\
&\leq P\Big(f(\phi_1,\phi_2)+\sup_{p\in\mathbb{S}^d}|\lambda(p,\phi_0)'\nabla_\phi\Psi(\theta_*(p),\phi_0)[\phi-\phi_0]|\geq Cr_n\cap\Omega\Big|D_n\Big)+o_p(1)\\
&\leq P\Big(o(\|\phi-\phi_0\|)+\sup_{p\in\mathbb{S}^d}|\lambda(p,\phi_0)'\nabla_\phi\Psi(\theta_*(p),\phi_0)|\|\|\phi-\phi_0\|\geq Cr_n\Big|D_n\Big)P\left(\Omega|D_n\right)+o_p(1)
\end{aligned}
$$

which converges to 0 in probability under assumption 4.1. □

E.3 Proof of Theorem 5.3

Proof. Denote $r_n=(\log n)^{1/2}n^{-1/2}$, $\Omega:=\{\phi\in B(\phi_0,r_n)\}$ and $h_n:=\sqrt{n}\sup_{p\in\mathbb{S}^d}(S_\phi(p)-S_{\phi_0}(p))$. Since the Total Variation distance is bounded by 2 we have:

$$
\begin{aligned}
\mathbf{E}\|P_{h_n|D_n}-\mathcal{N}(\bar{\Delta}_{n,\phi_0},\bar{I}_{\phi_0}^{-1})\|_{TV} &= \mathbf{E}\|P_{h_n|D_n}-\mathcal{N}(\bar{\Delta}_{n,\phi_0},\bar{I}_{\phi_0}^{-1})\|_{TV}I_\Omega\\
&\quad+\ \mathbf{E}\|P_{h_n|D_n}-\mathcal{N}(\bar{\Delta}_{n,\phi_0},\bar{I}_{\phi_0}^{-1})\|_{TV}I_{\Omega^c}\\
&\leq\ \mathbf{E}\|P_{h_n|D_n}-\mathcal{N}(\bar{\Delta}_{n,\phi_0},\bar{I}_{\phi_0}^{-1})\|_{TV}I_\Omega+2P(\Omega^c).
\end{aligned}
$$

By the expansion of the support function given in lemma G.16 the element h_n is asymptotically equal to $\sqrt{n}|\lambda(p,\phi_0)'\nabla_\phi\Psi(\theta_*(p),\phi_0)[\phi-\phi_0]$. Moreover, under assumption 4.1, $P(\Omega^c|D_n)=o_p(1)$. Therefore, $\mathbf{E}\|P_{h_n|D_n}-\mathcal{N}(\bar{\Delta}_{n,\phi_0},\bar{I}_{\phi_0}^{-1})\|_{TV}$ equals

$$
\mathbf{E}\|P_{\sqrt{n}\sup_{p\in\mathbb{S}^d}|\lambda(p,\phi_0)'\nabla_\phi\Psi(\theta_*(p),\phi_0)[\phi-\phi_0]\|D_n}-\mathcal{N}(\bar{\Delta}_{n,\phi_0},\bar{I}_{\phi_0}^{-1})\|_{TV}I_\Omega+o(1)+o_p(1)
$$

which converges to 0 under assumption 5.7.

□

F Proofs for Section 6

F.1 Proof of Theorem 6.1

Lemma F.1. *For any consistent estimator $\|\hat{\phi}-\phi_0\|=o_p(1)$, $P(\theta\notin\Theta(\hat{\phi})|D_n)=o_p(1)$.*

Proof. Straightforward calculation shows

$$
\begin{aligned}
&P(\theta\notin\Theta(\hat{\phi})|D_n)=\int\pi(\theta\notin\Theta(\hat{\phi})|\phi)p(\phi|D_n)d\phi\leq\int\pi(\theta\notin\Theta(\phi_0)|\phi)p(\phi|D_n)d\phi\\
&+\int|\pi(\theta\notin\Theta(\hat{\phi})|\phi)-\pi(\theta\notin\Theta(\phi_0)|\phi)|p(\phi|D_n)d\phi\equiv A+B.
\end{aligned}
$$

We investigate A and B respectively. First of all, by the posterior concentration of ϕ, there is $r_n = o(1)$ such that (note that $\pi(\theta|\phi) = \pi(\theta|\phi)I_{\theta\in\Theta(\phi)}$)

$$\begin{aligned} A &= \int_{\|\phi-\phi_0\|\leq r_n} \pi(\theta \notin \Theta(\phi_0)|\phi)p(\phi|D_n)d\phi + o_p(1) \\ &= \int_{\|\phi-\phi_0\|\leq r_n} \int_{\theta\notin\Theta(\phi_0),\theta\in\Theta(\phi)} \pi(\theta|\phi)d\theta p(\phi|D_n)d\phi + o_p(1). \end{aligned}$$

By the expansion of the support function, $\|\phi - \phi_0\| \leq r_n$ implies $d_H(\Theta(\phi), \Theta(\phi_0)) \leq Cr_n$ for some $C > 0$. Hence $\Theta(\phi) \subset \Theta(\phi_0)^{Cr_n}$, which yields

$$A \leq \int_{\|\phi-\phi_0\|\leq r_n} \int_{\theta\notin\Theta(\phi_0),\theta\in\Theta(\phi_0)^{Cr_n}} \pi(\theta|\phi)d\theta p(\phi|D_n)d\phi + o_p(1) = o_p(1),$$

where the last equality follows since $r_n = o(1)$ and $\sup_{\theta,\phi} \pi(\theta|\phi) < \infty$. On the other hand, let $H = \{\theta \in \Theta(\phi_0), \theta \notin \Theta(\hat{\phi})\} \cup \{\theta \in \Theta(\hat{\phi}), \theta \notin \Theta(\phi_0)\}$. Then

$$\sup_{\phi\in\Phi} |\pi(\theta \notin \Theta(\phi_0)|\phi) - \pi(\theta \notin \Theta(\hat{\phi}_0)|\phi)| \leq \sup_{\phi\in\Phi} \pi(\theta \in H|\phi).$$

Due to the consistency of $\hat{\phi}$, and the expansion of the support function, $d_H(\Theta(\phi_0), \Theta(\hat{\phi})) = o_p(1)$, and therefore $\mu(H) = o_p(1)$ which implies $\sup_{\phi\in\Phi} \pi(H|\phi) = o_p(1)$ since $\sup_{\theta,\phi} \pi(\theta|\phi) < \infty$. We conclude that $B = o_p(1)$. □

Now to finish proving part (i) of the theorem, noting that $\Theta(\hat{\phi}) \subset \text{FCS}(\tau)$, we have

$$P(\theta \notin \text{FCS}(\tau)|D_n) \leq P(\theta \notin \Theta(\hat{\phi})|D_n) = o_p(1),$$

which by definition, leads to the conclusion of part (i).

For part (ii), by the definition of BCS that $P(\theta \in \text{BCS}(\tau)|D_n) = 1 - \tau$, we have

$$P(\theta \in \text{FCS}(\tau), \theta \notin \text{BCS}(\tau)|D_n) \leq P(\theta \notin \text{BCS}(\tau)|D_n) = \tau.$$

On the other hand, Lemma F.1 implies $P(\theta \notin \Theta(\hat{\phi}) \cup \theta \in \text{BCS}(\tau)|D_n) \leq o_p(1) + 1 - \tau$. Hence

$$\begin{aligned} &P(\theta \in \text{FCS}(\tau), \theta \notin \text{BCS}(\tau)|D_n) \geq P(\theta \in \Theta(\hat{\phi}), \theta \notin \text{BCS}(\tau)|D_n) \\ &\geq 1 - (o_p(1) + 1 - \tau) = \tau + o_p(1). \end{aligned}$$

F.2 Proof of Theorems 6.2 and 6.3

Theorem 6.2 has been proved in the main text. We now prove Theorem 6.3.

Lemma F.2. *Suppose that assumptions 5.1-5.6 hold with $\delta = r_n$. Then, for any $x \geq 0$,*

$$P(\sqrt{n} \sup_{\|p\|=1} |S_\phi(p) - S_{\hat{\phi}_M}(p)| \leq x|D_n) - P_{D_n}(\sqrt{n} \sup_{\|p\|=1} |S_{\phi_0}(p) - S_{\hat{\phi}_M}(p)| \leq x) = o_p(1).$$

Proof. For $\theta_*(p)$ and $\lambda(p,\phi_0)$ defined in Lemma G.16, define

$$f_p^n(\phi_1,\phi_2) = \sqrt{n}\lambda(p,\phi_0)^T\nabla\Psi(\theta_*(p),\phi_0)(\phi_1-\phi_2)$$

where $\theta_*(p)$ and $\lambda(p,\phi_0)$ do not depend on specific choice of ϕ_1 and ϕ_2. Then Lemma G.16 implies that

$$\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)}\sup_{\|p\|=1}|\sqrt{n}(S_{\phi_1}(p)-S_{\phi_2}(p)) - f_p^n(\phi_1,\phi_2)| = o(1). \tag{F.1}$$

For notational simplicity, we further write $g_n(\phi_1,\phi_2) = \sqrt{n}\sup_{\|p\|=1}|S_{\phi_1}(p)-S_{\phi_2}(p)|$. Then

$$\begin{aligned} & |P(g_n(\phi,\hat{\phi}_M)\le x|D_n) - P_{D_n}(g_n(\phi_0,\hat{\phi}_M)\le x)| \\ \le\ & |P(\sup_{\|p\|=1}|f_p^n(\phi,\hat{\phi}_M)|\le x|D_n) - P_{D_n}(\sup_{\|p\|=1}|f_p^n(\phi_0,\hat{\phi}_M)|\le x)| \\ & +|P(g_n(\phi,\hat{\phi}_M)\le x|D_n) - P(\sup_{\|p\|=1}|f_p^n(\phi,\hat{\phi}_M)|\le x|D_n)| \\ & +|P_{D_n}(\sup_{\|p\|=1}|f_p^n(\phi_0,\hat{\phi}_M)|\le x) - P_{D_n}(g_n(\phi_0,\hat{\phi}_M)\le x)| \\ =\ & a_1+a_2+a_3 \end{aligned}$$

It remains to show that $a_i = o_p(1)$ for $i=1,2,3$. By the Bernstein von Mises theorem and asymptotic normality of $\hat{\phi}_M$ (Assumption 6.2), the posterior $\sqrt{n}(\phi-\hat{\phi}_M)|D_n$ and the sampling distribution $\sqrt{n}(\hat{\phi}_M-\phi_0)$ are asymptotically identically distributed. This implies that $a_1 = o_p(1)$. On the other hand, (F.1) implies $\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)}|g_n(\phi_1,\phi_2) - \sup_{\|p\|=1} f_p^n(\phi_1,\phi_2)| = o(1)$, where we used the inequality $|\sup_x g_1(x) - \sup_x g_2(x)| \le 3\sup_x|g_1(x)-g_2(x)|$ for any wo functions $g_1(x)$ and $g_2(x)$. Therefore, if we write $\Delta = |g_n(\phi,\hat{\phi}_M) - \sup_{\|p\|=1} f_p^n(\phi,\hat{\phi}_M)|$,

$$\begin{aligned} a_2 \ \le\ & |P(g_n(\phi,\hat{\phi}_M)\le x|D_n) - P(g_n(\phi,\hat{\phi}_M)\le x+\Delta|D_n)| \\ & +|P(g_n(\phi,\hat{\phi}_M)\le x|D_n) - P(g_n(\phi,\hat{\phi}_M)\le x-\Delta|D_n)| = o_p(1) \end{aligned}$$

since $\Delta = o_p(1)$. Similarly, $a_3 = o_p(1)$. □

Proof of Theorem 6.3

$$\begin{aligned} & P_{D_n}(\Theta(\hat{\phi}_M)^{-q_\tau/\sqrt{n}}\subset\Theta(\phi_0)\subset\Theta(\hat{\phi}_M)^{q_\tau/\sqrt{n}}) = P_{D_n}(\sqrt{n}\sup_{\|p\|=1}|S_{\phi_0}(p)-S_{\hat{\phi}_M}(p)|\le q_\tau) \\ \ge\ & P(J(\phi)\le q_\tau|D_n) + o_p(1) = 1-\tau+o_p(1), \end{aligned}$$

where the inequality follows from Lemma F.2.

Proof of Corollary 6.1

Proof. For any fixed $\theta\in\Theta(\phi_0)$,

$$P_{D_n}(\theta\in\Theta(\hat{\phi}_M)^{\tilde{q}_\tau/\sqrt{n}}) \ge P_{D_n}(\Theta(\phi_0)\subset\Theta(\hat{\phi}_M)^{\tilde{q}_\tau/\sqrt{n}}) \ge 1-\tau+o_p(1)$$

where $o_p(1)$ is uniformly in $\theta\in\Theta(\phi_0)$. This gives the result. □

G Technical Lemmas

We remind some technical notation that will be used throughout this section.

- $d = \dim(\theta)$, $d_\phi = \dim(\phi)$;
- $B(\phi_0, \delta) = \{\phi \in \Phi;\ \|\phi - \phi_0\| \le \delta\}$;
- $\Psi_S(\theta, \phi)$ is the k_S-subvector of $\Psi(\theta, \phi)$ containing the constraints that are strictly convex functions of θ and $\lambda_S(p, \phi)$ are the corresponding Lagrange multipliers for $p \in \mathbb{S}^d$;
- $\Psi_L(\theta, \phi)$ is the k_L-subvector of $\Psi(\theta, \phi)$ containing the constraints that are linear in θ and $\lambda_L(p, \phi)$ are the corresponding Lagrange multipliers for $p \in \mathbb{S}^d$;
- $\Xi(p, \phi) = \arg\max_{\theta \in \Theta}\{p^T\theta;\ \Psi(\theta, \phi) \le 0\}$ is the *support set* of $\Theta(\phi)$;
- $\nabla_\phi \Psi(\theta, \phi)$ the $k \times d_\phi$ matrix of partial derivatives of Ψ with respect to ϕ;
- $\forall \theta \in \Theta(\phi)$ and $\phi \in B(\phi_0, r_n)$, we denote by $Act(\theta, \phi) := \{i;\ \Psi_i(\theta, \phi) = 0\}$ the set of the inequality active constraint indices and by $d_A(\theta, \phi)$ the number of its elements;
- $\forall i \in Act(\theta, \phi)$, $\nabla_\theta \Psi_i(\theta, \phi)$ denotes the d-vector of partial derivatives of Ψ_i with respect to θ.

Lemma G.1. *Under assumptions 5.1 and 5.3 (iii) with $\delta = r_n$ if $(\tilde\theta, \tilde\phi) \in \Theta \times B(\phi_0, r_n)$ are such that $\Psi(\tilde\theta, \tilde\phi) < 0$, then there exists a N such that whenever $n \ge N$ we have that $\Psi(\tilde\theta, \phi) < 0$ for every $\phi \in B(\phi_0, r_n)$.*

Proof. Under assumption 5.3 (iii) with $\delta = r_n$ for every $\phi \in B(\phi_0, r_n)$ there exists a $\theta \in \Theta$ such that $\Psi(\theta, \phi) < 0$. Denote by $(\tilde\theta, \tilde\phi)$ this value (*i.e.* $\Psi(\tilde\theta, \tilde\phi) < 0$). By assumption 5.1 the function $\Psi(\theta, \phi)$ is continuous in (θ, ϕ), then there is a N such that whenever $n \ge N$ we have that $\Psi(\tilde\theta, \phi) < 0$ for every $\phi \in B(\phi_0, r_n)$.
Q.E.D.

Lemma G.2. *Let assumptions 5.1 and 5.2 be satisfied with $\delta = r_n$. For every $\epsilon_n > 0$ there exists a N such that for every $n \ge N$ and $\phi \in B(\phi_0, r_n)$*

$$\sup_{\theta \in \Theta} \|\Psi(\theta, \phi) - \Psi(\theta, \phi_0)\| < \epsilon_n. \tag{G.1}$$

Proof. Under assumption 5.1, the function $\phi \to \Psi(\theta, \phi)$ is continuous on Φ, for every $\theta \in \Theta$, and uniformly continuous on $B(\phi_0, r_n)$, due to the compactness of $B(\phi_0, r_n)$. Therefore, $\forall \epsilon_n > 0$ there exists a $\delta_\theta > 0$ such that $\forall \phi \in B(\phi_0, r_n)$: $\|\Psi(\theta, \phi) - \Psi(\theta, \phi_0)\| < \epsilon_n$ for every $\theta \in \Theta$. Now, for every $\phi \in B(\phi_0, r_n)$ denote $f_\phi(\theta) = \|\Psi(\theta, \phi) - \Psi(\theta, \phi_0)\|$ and

$$A_{\delta_\theta} := \left\{\tilde\theta \in \Theta;\ f_\phi(\tilde\theta) < \epsilon_n,\ \forall \phi \in B(\phi_0, r_n);\ r_n < \delta_\theta\right\}$$

for every $\theta \in \Theta$, $\epsilon_n > 0$ and $\delta_\theta > 0$. This means that $\forall \theta \in \Theta$ there is a δ_θ such that $\theta \in A_{\delta_\theta}$. Under assumption 5.1 $f_\phi(\theta)$ is continuous in θ, hence A_{δ_θ} is an open set and $\bigcup_{\theta \in \Theta} A_{\delta_\theta}$ is an open cover of Θ: $\Theta \subset \bigcup_{\theta \in \Theta} A_{\delta_\theta}$. Due to compactness of Θ (see assumption 5.2) there exists a finite set $\{\delta_1, \dots, \delta_K\}$, $K < \infty$ such that $\{A_{\delta_i}\}_{i=1}^K$ is a subcover of Θ, that is, $\Theta \subset \bigcup_{i=1}^K A_{\delta_i}$.

Let $\delta^* = \min\{\delta_1, \ldots, \delta_K\}$ so that $A_{\delta_i} \subseteq A_{\delta^*}$ for every $i = 1, \ldots, K$ and for any $\theta \in \Theta$ we have $\theta \in A_{\delta^*}$. Remark that this δ^* does not depend on θ. This then implies that for any $\phi \in B(\phi_0, r_n)$ and $r_n < \delta^*$

$$\sup_{\theta \in \Theta} \|\Psi(\theta, \phi) - \Psi(\theta, \phi_0)\| < \epsilon_n.$$

Q.E.D.

Lemma G.3. *Under assumptions 5.1, 5.2 and 5.3 (iii) with $\delta = r_n$, there exists a N such that for every $n \geq N$ the correspondence $\phi \mapsto \Theta(\phi)$ is well defined and continuous at all $\phi \in B(\phi_0, r_n)$, that is, it is upper and lower hemicontinuous.*

Proof. This proof follows the lines of the proof of Lemma B.3 in Kaido and Santos (2011) with minor modifications. First, under assumptions 5.1 and 5.3 (iii) the set $\Theta(\phi)$ is a convex set with nonempty interior for every $\phi \in B(\phi_0, r_n)$.

Next, we have to show that the correspondence $\phi \mapsto \Theta(\phi)$ is continuous (for a definition of continuity of a correspondence see for instance Definition 17.2 in Aliprantis and Border (2006)). First, we show that $\phi \mapsto \Theta(\phi)$ is lower hemicontinuous at any $\phi \in B(\phi_0, r_n)$. We show this by showing Theorem 17.19 (ii) in Aliprantis and Border (2006), that is, for any $\theta^* \in \Theta(\phi)$ (*i.e.* $\Psi(\theta^*, \phi) \leq 0$) and net $\{\phi_j\}$ with $\phi_j \to \phi$, there exists a subnet $\{\phi_{j_\beta}\}_{\beta \in \Upsilon}$ and a net $\{\theta_\beta\}_{\beta \in \Upsilon}$ such that $\theta_\beta \in \Theta(\phi_{j_\beta})$, for every $\beta \in \Upsilon$, and $\theta_\beta \to \theta^*$. In order to show this, consider a net $\phi_j \to \phi$. Then we distinguish between two cases. *Case I: $\theta^* \in int(\Theta)(\phi)$, i.e.* $\Psi(\theta^*, \phi) < 0$. By Lemma G.1 there exists a N such that for every $n \geq N$ there exists a j_n such that $\phi_{j_n} \in B(\phi_0, r_n)$ and $\Psi(\theta^*, \phi_j) < 0$ for every $j \geq j_n$. Define $\Upsilon \equiv \{j \geq j_n\}$ and $\phi_j = \phi_\beta$ with $\beta \in \Upsilon$. Fix $\theta_\beta = \theta^*$ (*i.e.* θ_β is a constant net equal to θ^*) so that $\theta_\beta \in \Theta(\phi_\beta)$ for every $\beta \in \Upsilon$ and $\theta_\beta \to \theta^*$. *Case II: $\theta^* \in \partial\Theta(\phi)$.* Since $\Theta(\phi)$ is convex with non-empty interior then there exists $\{\tilde{\theta}_\lambda\}$ that belongs to $int(\Theta)(\phi)$, for every λ, with $\tilde{\theta}_\lambda \to \theta^*$. By Lemma G.1 there exists a N such that for every $n \geq N$ there exists a $j_{n,\lambda}$ such that $\phi_j \in B(\phi_0, r_n)$ and $\Psi(\tilde{\theta}_\lambda, \phi_j) < 0$ for every $j \geq j_{n,\lambda}$ (*i.e.* $\tilde{\theta}_\lambda \in \Theta(\phi_j)$ for every $j \geq j_{n,\lambda}$). Since $B(\phi_0, r_n)$ is compact then every convergent net admits a convergent subnet, that is, there exists l_n such that $\{\phi_{j_\beta}\}_{\beta \in \Upsilon}$ converges to ϕ where $\Upsilon \equiv \{j \geq \max\{l_0, j_{0,\lambda}\}\}$. The corresponding $\theta_\beta \equiv \tilde{\theta}_\lambda$ satisfies $\theta_\beta \to \theta^*$ by construction, $\theta_\beta \in \Theta(\phi_{j_\beta})$ and $\theta^* \in \partial\Theta(\phi)$.

Now, let us show that the correspondence $\phi \mapsto \Theta(\phi)$ is upper hemicontinuous at any $\phi \in B(\phi_0, r_n)$. By theorem 17.16 in Aliprantis and Border (2006) it is sufficient to show that for every net $\{\phi_j, \theta_j\}$ such that $\theta_j \in \Theta(\phi_j)$ for each j (*i.e.* (ϕ_j, θ_j) is in the graph of $\Theta(\cdot)$), if $\phi_j \to \phi$ then $\theta_j \to \theta^* \in \Theta(\phi)$.

To show this, first observe that, since Θ is compact then a convergent net $\theta_j \in \Theta(\phi_j)$ has a subnet $\theta_{j_\beta} \in \Theta(\phi_{j_\beta})$ which is convergent, that is, $\theta_{j_\beta} \to \theta^*$ for some $\theta^* \in \Theta$. Therefore, we have to show that $\theta^* \in \Theta(\phi)$. To show this first remark that there exists $j_n > 0$ such that for every $j \geq j_n$, $\phi_j \in B(\phi_0, r_n)$. Since $B(\phi_0, r_n)$ is compact then every convergent net $\phi_j \to \phi$ admits a convergent subnet $\{\phi_{j_\beta}\}_{\beta \in \Upsilon}$ such that $\phi_{j_\beta} \to \phi$. By the result of Lemma G.2 we have that

$$\Psi(\theta_{j_\beta}, \phi_{j_\beta}) - \Psi(\theta_{j_\beta}, \phi)$$

converges to 0 uniformly in θ. Moreover, under assumption 5.1 the function $\Psi(\cdot, \phi)$ is continuous in θ. This allows to conclude that

$$\Psi(\theta_{j_\beta}, \phi_{j_\beta}) \to \Psi(\theta^*, \phi) \tag{G.2}$$

(because $\Psi(\theta_{j_\beta}, \phi_{j_\beta}) - \Psi(\theta_{j_\beta}, \phi) = \big(\Psi(\theta_{j_\beta}, \phi_{j_\beta}) - \Psi(\theta_{j_\beta}, \phi)\big) + \big(\Psi(\theta_{j_\beta}, \phi) - \Psi(\theta^*, \phi)\big)$). Since $\Psi(\theta_{j_\beta}, \phi_{j_\beta}) \leq 0$ because $\theta_{j_\beta} \in \Theta(\phi_{j_\beta})$ then $\Psi(\theta^*, \phi) \leq 0$. We conclude that $\theta^* \in \Theta(\phi)$ and upper hemicontinuity is established. Q.E.D.

Lemma G.4. *Let Assumptions 5.1, 5.2, and 5.3 (ii)-(iii) hold with $\delta = r_n$. Then, there exists a N such that for every $n \geq N$ the correspondence*

$$(p, \phi) \mapsto \Xi(p, \phi) = \arg\max_{\theta \in \Theta} \{p^T \theta;\ \Psi(\theta, \phi) \leq 0\}$$

has non-empty compact values and it is upper hemicontinuous on $\mathbb{S}^d \times B(\phi_0, r_n)$.

Proof. Let $\tau_0 : \mathbb{S}^d \to (0,1)$ be a measurable and differentiable function of p. For every $\phi_1, \phi_2 \in B(\phi_0, r_n)$ define $\phi_{\tau_0(p)} = \tau_0(p)\phi_1 + (1-\tau_0(p))\phi_2$. Lemma G.3 implies that there exists a N such that for every $n > N$ the correspondence $\phi \mapsto \Theta(\phi)$ is well defined and continuous at all $\phi \in B(\phi_0, r_n)$. Therefore, for any $\phi_1, \phi_2 \in B(\phi_0, r_n)$, the correspondence $p \mapsto \Theta(\phi_{\tau_0(\cdot)}) : \mathbb{S}^d \to \mathbb{R}^d$ is continuous because it is the composition of continuous functions.

Under assumption 5.1 the function $\theta \mapsto \Psi(\theta, \phi)$ is continuous in θ for every $\phi \in \Phi$ then it is also lower semi-continuous. Therefore, for every $i = 1, \dots, k$ and $\phi \in \Phi$, the lower level sets $\{\theta \in \Theta;\ \Psi_i(\theta, \phi) \leq 0\}$ are closed and the set $\Theta(\phi)$ is closed because it is a finite intersection of closed sets. Because $\forall \phi \in \Phi$, $\Theta(\phi) \subseteq \Theta$ and Θ is compact (under assumption 5.2) then the set $\Theta(\phi)$ is also compact.

Now, under assumption 5.3 (ii) we can apply the "Berge Maximum Theorem", see *e.g.* theorem 17.31 in Aliprantis and Border (2006), which guarantees that the correspondence

$$p \mapsto \Xi(p, \phi_{\tau_0(p)}) = \arg\max_{\theta \in \Theta(\phi_{\tau_0(p)})} p^T \theta$$

has nonempty compact values and it is upper hemicontinuous for every $\phi_1, \phi_2 \in B(\phi_0, r_n)$.
By the definition of upper hemicontinuity (see *e.g.* theorem 17.16 in Aliprantis and Border (2006)), for every net $\{p_j, \theta_j\}$ such that $\theta_j \in \Xi(p_j, \phi_{\tau_0(p_j)})$, if $p_j \to p$ then $\theta_j \to \theta \in \Xi(p, \phi_{\tau_0(p)})$. The correspondence $p \mapsto \Xi(p, \phi_{\tau_0(p)})$ may be rewritten as a correspondence of two arguments: $(p, \phi_{\tau_0(p)}) \mapsto \Xi(p, \phi_{\tau_0(p)})$ – where $(p, \phi_{\tau_0(p)})$ in turn is the value taken by the map $(p, \phi_1, \phi_2) \mapsto (p, \phi_{\tau_0(p)})$ – and, since for every net $(p_j, \phi_{\tau_0(p_j)}, \theta_j)$ such that $\theta_j \in \Xi(p_j, \phi_{\tau_0(p_j)})$, if $(p_j, \phi_{\tau_0(p_j)}) \to (p, \phi_{\tau(p)})$ then $\theta_j \to \theta \in \Xi(p, \phi_{\tau_0(p)})$, it follows that the correspondence $(p, \phi_{\tau_0(p)}) \mapsto \Xi(p, \phi_{\tau_0(p)})$ is also upper hemicontinuous. Since every $\phi \in B(\phi_0, r_n)$ can be represented as $\phi_{\tau_0(p)}$ (if we choose ϕ_1 and ϕ_2 on the boundary of $B(\phi_0, r_n)$) then we can rewrite the correspondence as $(p, \phi) \mapsto \Xi(p, \phi)$ which is upper hemicontinuous. Q.E.D.

Lemma G.5. *Let assumptions 5.1, 5.2 and 5.3 (ii)-(iii) be satisfied with $\delta = r_n$. Let $W \subseteq \mathbb{S}^d$ be compact and $\Xi(p, \phi_0)$ be a singleton $\forall p \in W$. Then there exists a N such that for every $n \geq N$ and $\phi \in B(\phi_0, r_n)$ there exists a $\varepsilon_n > 0$ which goes to 0 as $r_n \to 0$ such that for $\theta_*(p) = \Xi(p, \phi_0)$,*

$$\sup_{p \in W} \sup_{\theta \in \Xi(p, \phi)} \|\theta - \theta_*(p)\| < \varepsilon_n$$

Proof. For every $(p, \phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$ define $\Xi^\delta(p, \phi) := \{\theta \in \mathbb{R}^d;\ \inf_{\tilde{\theta} \in \Xi(p, \phi)} \|\theta - \tilde{\theta}\| < \delta\}$. Since $\|\phi - \phi_0\| \leq r_n$ and $\Xi(p, \phi) : \mathbb{S}^d \times B(\phi_0, r_n) \to \mathbb{R}^d$ is upper hemicontinuous by lemma G.4 (for n sufficiently large), then whenever $\phi \in B(\phi_0, r_n)$ for any $p \in \mathbb{S}^d$ there exists a $\varepsilon_n > 0$ such that $\Xi(p, \phi) \subseteq \Xi^{\varepsilon_n}(p, \phi_0)$ where $\varepsilon_n \to 0$ as $r_n \to 0$. This implies that

$$\sup_{\theta \in \Xi(p, \phi)} \|\theta - \theta_*(p)\| \leq \sup_{\theta \in \Xi^{\varepsilon_n}(p, \phi_0)} \|\theta - \theta_*(p)\| < \varepsilon_n.$$

Now, fix the sequence ε_n, denote $f_\phi(p) := \sup_{\theta\in\Xi(p,\phi)} \|\theta - \theta_*(p)\|$ for $\phi \in B(\phi_0, r_n)$ and

$$A_{\delta_p} := \{\tilde{p} \in W;\ f_\phi(\tilde{p}) < \varepsilon_n,\ \forall\phi \in B(\phi_0, r_n) \text{ with } r_n < \delta_p\}$$

for every $p \in W$. This means that for any $p \in W$ there is a δ_p so that $p \in A_{N_p}$. Since $f_\phi(p)$ is continuous in p (by the result of lemma G.4), hence A_{δ_p} is an open set and $\bigcup_{p\in W} A_{\delta_p}$ is an open cover of W: $W \subset \bigcup_{p\in W} A_{\delta_p}$.
Due to the compactness of W there exists a finite set $\{\delta_1, \ldots, \delta_K\}$, $K < \infty$ such that $\{A_{\delta_i}\}_{i=1}^K$ is a subcover of W: $W \subset \bigcup_{i=1}^K A_{\delta_i}$. Let $\delta^* = \min\{\delta_1, \ldots, \delta_K\}$ so that $A_{\delta_i} \subseteq A_{\delta^*}$ for every $i = 1, \ldots, K$ and for any $p \in W$ we have $p \in A_{\delta^*}$. This then implies that for any $r_n < \delta^*$

$$\sup_{p\in W} \sup_{\theta\in\Xi(p,\phi)} \|\theta - \theta_*(p)\| < \varepsilon_n.$$

□

Lemma G.6. *Let assumptions 5.1, 5.2 and 5.3 (ii)-(iii) be satisfied with $\delta = r_n$. Let $W \subseteq \mathbb{S}^d$ be compact and $\Xi(p, \phi_0)$ be a singleton $\forall p \in W$. Moreover, let $\varepsilon_n > 0$ satisfy assumption 5.6 (iv). Then for $\theta_*(p) = \Xi(p, \phi_0)$,*

$$\sup_{\phi\in B(\phi_0,r_n)} \sup_{p\in W} \sup_{\theta\in\Xi(p,\phi)} \|\theta - \theta_*(p)\| = O(r_n).$$

Proof. By lemma G.5 we know that $\sup_{p\in W} \sup_{\theta\in\Xi(p,\phi)} \|\theta - \theta_*(p)\| < \varepsilon_n$ where now ε_n is chosen such that $\varepsilon_n = O(r_n)$. Such an ε_n exists by assumption 5.6 (iv). Next, for $\phi \in B(\phi_0, r_n)$, denote $f(\phi) := \sup_{p\in W} \sup_{\theta\in\Xi(p,\phi)} \|\theta - \theta_*(p)\|$ and

$$A_{\delta_\phi} := \left\{\tilde{\phi} \in B(\phi_0, r_n);\ f(\tilde{\phi}) < \varepsilon_n,\ \forall r_n < \delta_\phi\right\}.$$

Remark that δ_ϕ must be such that $\phi \in A_{\delta_\phi}$ for every $\phi \in B(\phi_0, r_n)$. Since $f(\phi)$ is continuous in ϕ, hence A_{δ_ϕ} is an open set and $\bigcup_{\phi\in B(\phi_0,r_n)} A_{\delta_\phi}$ is an open cover of $B(\phi_0, r_n)$, that is, $B(\phi_0, r_n) \subset \bigcup_{\phi\in B(\phi_0,r_n)} A_{\delta_\phi}$. Due to compactness of $B(\phi_0, r_n)$ there exists a finite set $\{\delta_1, \ldots, \delta_K\}$, $K < \infty$ such that $\{A_{\delta_i}\}_{i=1}^K$ is a subcover of $B(\phi_0, r_n)$, that is,

$$B(\phi_0, r_n) \subset \bigcup_{i=1}^K A_{\delta_i}.$$

Let $\delta^* = \min\{\delta_1, \ldots, \delta_K\}$ so that $A_{\delta_i} \subseteq A_{\delta^*}$ for every $i = 1, \ldots, K$ and for every $\phi \in B(\phi_0, r_n)$ we have $\phi \in A_{\delta^*}$. This then implies that for any $r_n < \delta^*$

$$\sup_{\phi\in B(\phi_0,r_n)} \sup_{p\in W} \sup_{\theta\in\Xi(p,\phi)} \|\theta - \theta_*(p)\| = O(r_n).$$

Q.E.D.

Lemma G.7. *Let assumptions 5.1, 5.2, 5.3 (iii) and 5.4 (i) be satisfied with $\delta = r_n$. For every $\theta \in \Theta(\phi_0)$ denote by $Act(\theta, \phi_0) := \{i;\ \Psi_i(\theta, \phi_0) = 0\}$ the set of constraints in $\Psi(\theta, \phi_0) \leq 0$ which are active and by $d_A(\theta, \phi)$ the number of elements in $Act(\theta, \phi)$. Then, there exists an N such that $d_A(\theta, \phi) \leq d$ for every $\theta \in \Theta(\phi)$, $\phi \in B(\phi_0, r_n)$ and $n \geq N$.*

Proof. Consider the correspondence $\Theta(\cdot) : B(\phi_0, r_n) \to \Theta$ and a net $(\phi_\alpha, \theta_\alpha)$ in the graph of $\Theta(\phi)$, that is, $\phi_\alpha \in B(\phi_0, r_n)$ and $\theta_\alpha \in \Theta(\phi_\alpha)$. Under Assumption 5.2, $\Theta(\phi)$ is compact $\forall \phi \in B(\phi_0, r_n)$ since it is a closed subset of a compact space. Because, by lemma G.3, the correspondence $\Theta(\cdot) : B(\phi_0, r_n) \to \Theta$ is upper-hemicontinuous it follows by theorem 17.16 in Aliprantis and Border (2006) and equation (G.2) that for every $i \in Act^c(\theta, \phi_0)$, for any $\epsilon > 0$, there exists an N such that for every $\phi_\alpha \in B(\phi_0, r_n)$ with $n > N$ we have

$$|\Psi_i(\theta_\alpha, \phi_\alpha) - \Psi_i(\theta^*, \phi_0)| < \epsilon, \quad \Psi_i(\theta_\alpha, \phi_\alpha) < 0$$

where $\theta^* \in \Theta(\phi_0)$. This is because, since $\Psi_i(\theta^*, \phi_0) < 0$ for every $i \in Act^c(\theta, \phi_0)$ then there exists a $\tilde{\epsilon} > 0$ such that $\Psi_i(\theta^*, \phi_0) < -\tilde{\epsilon}_i$ and, for any $\epsilon < \tilde{\epsilon}_i$, we can always find a N such that $\Psi_i(\theta_\alpha, \phi_\alpha) < 0$ for every $\phi_\alpha \in B(\phi_0, r_n)$ and $n > N$.
This means that for every $\theta \in \Theta(\phi_0)$, $Act^c(\theta, \phi_0) \subseteq Act^c(\theta_\alpha, \phi_\alpha)$ for every $\phi_\alpha \in B(\phi_0, r_n)$ and n sufficiently large. Therefore, the reverse inclusion holds for the complements of these sets:

$$Act(\theta_\alpha, \phi_\alpha) \subseteq Act(\theta, \phi_0), \qquad \theta \in \Theta(\phi_0),\ \theta_\alpha \in \Theta(\phi_\alpha) \tag{G.3}$$

for every $\phi_\alpha \in B(\phi_0, r_n)$ and n sufficiently large. By assumption 5.4 (i) we have $d_A(\theta, \phi_0) \leq d$ which, together with (G.3), implies that for any $\theta \in \Theta(\phi)$, $d_A(\theta, \phi) \leq d$ for every $\phi \in B(\phi_0, r_n)$ and n sufficiently large. Q.E.D.

Lemma G.8. *Let Assumptions 5.1, 5.2, 5.3 (iii), 5.3 (v) and 5.4 (ii) be satisfied with $\delta = r_n$. Then, there exists an N such that for every $n \geq N$, $\phi \in B(\phi_0, r_n)$ and $\theta \in \Theta(\phi)$ the vectors*

$$\{\nabla_\theta \Psi_i(\theta, \phi)\}_{i \in Act(\theta, \phi)}$$

are linearly independent.

Proof. By (G.3) we have the inclusion $Act(\theta, \phi) \subseteq Act(\theta, \phi_0)$, for every $\theta \in \Theta(\phi_0)$, $\phi_\alpha \in B(\phi_0, r_n)$ and n sufficiently large. Therefore, we can prove the results by considering the indices in the biggest set $Act(\theta, \phi_0)$.
Since, by lemma G.3, the correspondence $\phi \mapsto \Theta(\phi)$ is upper hemicontinuous then for every net $\{\phi_\alpha, \theta_\alpha\}$ in the graph of $\Theta(\cdot)$ (*i.e.* such that $\theta_\alpha \in \Theta(\phi_\alpha)$, $\forall \alpha$ such that $\phi_\alpha \in B(\phi_0, r_n)$) we have that if $\phi_\alpha \to \phi_0$ then $\theta_\alpha \to \theta^*$ where θ^* is some element of $\Theta(\phi_0)$ (see *e.g.* theorem 17.16 in Aliprantis and Border (2006)). Because by assumption 5.3 (v) the vectors $\nabla_\theta \Psi_i(\theta, \phi)$, $i \in Act(\theta, \phi_0)$ with $\theta \in \Theta(\phi_0)$, are continuous in (θ, ϕ) it follows that

$$\nabla_\theta \Psi_i(\theta_\alpha, \phi_\alpha) \to \nabla_\theta \Psi_i(\theta^*, \phi_0), \quad \forall i \in Act(\theta, \phi_0),\ \theta \in \Theta(\phi_0). \tag{G.4}$$

Now, denote by $\nabla_\theta \Psi^A(\theta, \phi)$ the $(d \times d_A)$ matrix obtained by stacking columnwise the vectors $\{\nabla_\theta \Psi_i(\theta, \phi)\}_{i \in Act(\theta, \phi_0)}$ and by $\{\rho_i(\theta, \phi)\}_{i \in Act(\theta, \phi_0)}$ its singular values. By assumption 5.4 (ii) the matrix $\nabla_\theta \Psi^A(\theta, \phi_0)$ is full-column rank and then there exists a $\epsilon > 0$ such that $\inf_{i \in Act(\theta, \phi_0)} \rho_i(\theta, \phi_0) > \epsilon$. Continuity of the singular values (which follows from the continuity of $\nabla_\theta \Psi^A(\theta, \phi)$, see *e.g.* theorem II.5.1 in Kato (1995)) and (G.4) imply

$$\rho_i(\theta_\alpha, \phi_\alpha) \to \rho_i(\theta^*, \phi_0) \quad \forall i \in Act(\theta, \phi_0),\ \theta \in \Theta(\phi_0).$$

We conclude that there exists a N such that for every $n \geq N$, $\phi \in B(\phi_0, r_n)$ and $\theta \in \Theta(\phi)$ the eigenvalues $\{\rho_i(\theta, \phi)\}_{i \in Act(\theta, \phi_0)}$ are strictly positive which implies that $\nabla_\theta \Psi^A(\theta, \phi)$ is non-singular. Henceforth, $\{\nabla_\theta \Psi_i(\theta, \phi)\}_{i \in Act(\theta, \phi)}$ are linearly independent. Q.E.D.

Lemma G.9. *Let assumptions 5.1, 5.2, 5.3 (iii), 5.3 (v) and 5.4 (ii) hold with $\delta = r_n$. Then, there exists a N such that for every $n \geq N$, $\phi \in B(\phi_0, r_n)$ and $p \in \mathbb{S}^d$ there exists a unique $\lambda(p,\phi) \in \mathbb{R}_+^k$ satisfying*

$$\sup_{\theta\in\Theta(\phi)} p^T\theta = \sup_{\theta\in\Theta}\left\{p^T\theta - \lambda(p,\phi)^T\Psi(\theta,\phi)\right\}. \tag{G.5}$$

Proof. For every $\phi \in \Phi$, $\Theta(\phi)$ is a compact set since it is a closed subset of Θ, which is compact under assumption 5.2. This implies that $\sup_{\theta\in\Theta(\phi)} < p, \theta >$ is finite. Hence, if assumption 5.3 (iii) holds then, the conditions of Corollary 28.2.1 in Rockafellar (1970) are satisfied and by applying this corollary we obtain that, for every $\phi \in B(\phi_0, r_n)$ and $p \in \mathbb{S}^d$, there exists $\lambda(p,\phi)$ satisfying equation (G.5).

In order to show uniqueness of $\lambda(p,\phi)$, suppose that for $(\phi, p) \in B(\phi_0, r_n) \times \mathbb{S}^d$ there exist two different vectors $\lambda_1(p,\phi)$ and $\lambda_2(p,\phi)$ that satisfy equation (G.5). Since assumption 5.3 (iv) implies $\Xi(p,\phi) \subset int\Theta$ for all $(p,\phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$, where $int\Theta$ denotes the interior of Θ, it follows that for any $\tilde{\theta} \in \Xi(p,\phi)$, $\lambda_1(p,\phi)$ and $\lambda_2(p,\phi)$ satisfy the first order condition:

$$p - \nabla_\theta\Psi(\tilde{\theta},\phi)\lambda_1(p,\phi) = p - \nabla_\theta\Psi(\tilde{\theta},\phi)\lambda_2(p,\phi) = 0. \tag{G.6}$$

By the complementary slackness condition the Lagrange multipliers of the non-binding constraints are equal to 0. Therefore, equation (G.6) simplifies to

$$p - \sum_{i=1}^{d_A} \lambda_1^i(p,\phi)\nabla_\theta\Psi_i(\tilde{\theta},\phi) = p - \sum_{i=1}^{d_A} \lambda_2^i(p,\phi)\nabla_\theta\Psi_i(\tilde{\theta},\phi) = 0 \tag{G.7}$$

which, after simplifications, gives

$$\sum_{i=1}^{d_A} \left(\lambda_1^i(p,\phi) - \lambda_2^i(p,\phi)\right)\nabla_\theta\Psi_i(\tilde{\theta},\phi) = 0. \tag{G.8}$$

By lemma G.8 the vectors $\{\nabla_\theta\Psi_i(\tilde{\theta},\phi)\}_{i\in Act(\tilde{\theta},\phi)}$ are linearly independent for $\phi \in B(\phi_0, r_n)$, $\tilde{\theta} \in \Theta(\phi)$ and n sufficiently large. Therefore, the same holds for $\tilde{\theta} \in \Xi(p,\phi)$ with $p \in \mathbb{S}^d$ since $\Xi(p,\phi) \subset \Theta(\phi)$. This and (G.8) contradict $\lambda_1(p,\phi) \neq \lambda_2(p,\phi)$. Q.E.D.

Lemma G.10. *Let Assumptions 5.1, 5.2, 5.3 (ii)-(v) and 5.4 (ii) hold with $\delta = r_n$. Then, there exists an N such that for every $n \geq N$ the vector $\lambda(p,\phi)$ is continuous in $(p,\phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$.*

Proof. For every $\phi \in B(\phi_0, r_n)$ and $\theta \in \Theta(\phi)$, denote by $\lambda^A(p,\phi)$ the d_A-vector with components $\{\lambda^i(p,\phi)\}_{i\in Act(\theta,\phi)}$ and by $\nabla_\theta\Psi^A(\theta,\phi)$ the $(d \times d_A)$ matrix obtained by stacking columnwise the vectors $\{\nabla_\theta\Psi_i(\theta,\phi)\}_{i\in Act(\theta,\phi)}$. By lemma G.8, there exists a N such that $\forall n \geq N$, $\phi \in B(\phi_0, r_n)$ and $\theta \in \Theta(\phi)$, the matrix $[\nabla_\theta\Psi^A(\theta,\phi)]^T\nabla_\theta\Psi^A(\theta,\phi)$ is invertible. It follows from the first order condition in (G.7) – which is valid under assumption 5.3 (iv) – that for n sufficiently large we can write

$$\lambda^A(p,\phi) = \left([\nabla_\theta\Psi^A(\theta,\phi)]^T\nabla_\theta\Psi^A(\theta,\phi)\right)^{-1}[\nabla_\theta\Psi^A(\theta,\phi)]^T p,$$

$\forall (p,\phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$ and $\theta \in \Xi(p,\phi)$. Since $\lambda^i(p,\phi) = 0$ for every $i \notin Act(\theta,\phi)$ then $\|\lambda(p,\phi)\| =$

$\|\lambda^A(p,\phi)\|$ and

$$\begin{aligned}\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\|\lambda(p,\phi)\| &= \sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\|\lambda^A(p,\phi)\|\\ &\leq \sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\left\|[\nabla_\theta\Psi^A(\tilde{\theta}(p,\phi),\phi)]^T\nabla_\theta\Psi^A(\tilde{\theta}(p,\phi),\phi)\right\|^{-1}\|\nabla_\theta\Psi^A(\tilde{\theta}(p,\phi),\phi)\|\|p\|\\ &\leq \sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\left(\underline{\rho}(\tilde{\theta}(p,\phi),\phi)\right)^{-2}\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\|\nabla_\theta\Psi^A(\tilde{\theta}(p,\phi),\phi)\|\end{aligned}$$

where $\tilde{\theta}(p,\phi)\in\Xi(p,\phi)$ and $\underline{\rho}(\theta,\phi)$ denotes the smallest singular values of $\nabla_\theta\Psi^A(\theta,\phi)$. Under assumption 5.3 (v), and because $Act(\theta,\phi)\subseteq Act(\theta,\phi_0)$, $\forall\theta\in\Theta(\phi)$, $\phi\in B(\psi_0,r_n)$ and n sufficiently large under assumption 5.4 (ii), the matrix $\nabla_\theta\Psi^A(\theta,\phi)$ is continuous in (θ,ϕ). Further, because $\Xi(p,\phi)\subset\Theta(\phi)$ and $\mathbb{S}^d\times B(\phi_0,r_n)\times\Xi(p,\phi)$ is compact (compactness of $\Xi(p,\phi)$ follows from lemma G.4), it follows from the *extreme value theorem* and lemma G.4 that $\nabla_\theta\Psi^A(\theta,\phi)$ attains its maximum value on $\mathbb{S}^d\times B(\phi_0,r_n)\times\Xi(p,\phi)$ so that there exists a constant $0<C_1<\infty$ such that

$$\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\|\nabla_\theta\Psi^A(\tilde{\theta}(p,\phi),\phi)\|\leq\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\sup_{\theta\in\Xi(p,\phi)}\|\nabla_\theta\Psi^A(\theta,\phi)\|<C_1.$$

Continuity of the singular values (which follows from the continuity of $\nabla_\theta\Psi^A(\theta,\phi)$, see *e.g.* theorem II.5.1 in Kato (1995)) and compactness of $\mathbb{S}^d\times B(\phi_0,r_n)\times\Xi(p,\phi)$ implies that there exists a constant $0<C_2<\infty$ such that

$$\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\left(\underline{\rho}(\tilde{\theta}(p,\phi),\phi)\right)^{-2}\leq\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\sup_{\theta\in\Xi(p,\phi)}\left(\underline{\rho}(\theta,\phi)\right)^{-2}<C_2.$$

This shows that $\lambda(p,\phi)$ is uniformly bounded in $(p,\phi)\in\mathbb{S}^d\times B(\phi_0,r_n)$ which implies that it is continuous. Q.E.D.

Lemma G.11. *Let assumptions 5.1, 5.2, 5.3 (ii)-(v) and 5.4 (ii) hold with $\delta=r_n$. Then, there exists a N such that for every $n\geq N$ and for any $\epsilon_n>0$ which converges to 0 as $r_n\to 0$ we have*

$$\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\|\lambda(p,\phi)-\lambda(p,\phi_0)\|=O(\epsilon_n).$$

Proof. By lemmas G.9 and G.10 there exists a N such that for every $n\geq N$ the function $\lambda:\mathbb{S}^d\times B(\phi_0,r_n)\to\mathbb{R}^k_+$ is singleton valued and continuous in $(p,\phi)\in\mathbb{S}^d\times B(\phi_0,r_n)$. Therefore, by compactness of $B(\phi_0,r_n)$ the function $\phi\mapsto\lambda(p,\phi)$ is uniformly continuous on $B(\phi_0,r_n)$ for every $p\in\mathbb{S}^d$. This means that for every p and any $\epsilon_n>0$ which converges to 0 as $r_n\to 0$ there exists a natural number N_p that depends on p such that for all $n\geq N_p$,

$$\sup_{\phi\in B(\phi_0,r_n)}\|\lambda(p,\phi)-\lambda(p,\phi_0)\|<\epsilon_n. \tag{G.9}$$

For a fixed ϵ_n define $f_n(p):=\sup_{\phi\in B(\phi_0,r_n)}\|\lambda(p,\phi)-\lambda(p,\phi_0)\|$ and

$$A_{N_p}:=\left\{\tilde{p}\in\mathbb{S}^d;\ f_n(\tilde{p})<\epsilon_n,\ \forall n>N_p\right\}$$

for every $p\in\mathbb{S}^d$. This means that for any $p\in\mathbb{S}^d$ there is a N_p so that $p\in A_{N_p}$. Since $f_n(p)$ is continuous in p, hence A_{N_p} is an open set and $\bigcup_{p\in\mathbb{S}^d}A_{N_p}$ is an open cover of $\mathbb{S}^d$: $\mathbb{S}^d\subset\bigcup_{p\in\mathbb{S}^d}A_{N_p}$.

Due to compactness of $\mathbb{S}^d$ there exists a finite set $\{N_1,\dots,N_K\}$, $K<\infty$ such that $\{A_{N_i}\}_{i=1}^K$ is a subcover of $\mathbb{S}^d$: $\mathbb{S}^d\subset\bigcup_{i=1}^K A_{N_i}$.
Let $N^*=\max\{N_1,\dots,N_K\}$ so that $A_{N_i}\subseteq A_{N^*}$ for every $i=1,\dots,K$ and for any $p\in\mathbb{S}^d$ we have $p\in A_{N^*}$. Remark that this N^* does not depend on p. This then implies that for any $n>N^*$

$$\sup_{p\in\mathbb{S}^d}\sup_{\phi\in B(\phi_0,r_n)}\|\lambda(p,\phi)-\lambda(p,\phi_0)\|<\epsilon_n.$$

□

Lemma G.12. *Let Assumptions 5.1, 5.2, 5.3 (ii)-(iii) hold with $\delta=r_n$. If $\Psi(\theta,\phi_0)$ contains a subvector of strictly convex functions $\Psi_S(\cdot,\phi_0)$ of θ, then there exists an N such that for every $n\geq N$ and any $\phi\in B(\phi_0,r_n)$, $p\in\mathbb{S}^d$ for which $\Xi(p,\phi)$ is not a singleton we have*

$$\Psi_S(\tilde\theta,\phi)<0\quad\textit{for some }\tilde\theta\in\Xi(p,\phi).$$

Proof. By lemma E.1 there exists an N such that for every $n\geq N$ and $\phi\in B(\phi_0,r_n)$ the function $\theta\to\Psi_S(\theta,\phi)$ is strictly convex. Moreover, since by lemma G.4 $\Xi(p,\phi)$ is compact for every $(p,\phi)\in\mathbb{S}^d\times B(\phi_0,r_n)$ for n sufficiently large, then $\Xi(p,\phi)$ is closed and bounded (by the Heine-Borel theorem) and convex (since it is a closed subset of the convex sets $\Theta(\phi)$). For some $\tilde\theta_1,\tilde\theta_2\in\Xi(p,\phi)$, $\tilde\theta_1\neq\tilde\theta_2$, and $\nu\in[0,1]$ define $\tilde\theta:=\nu\tilde\theta_1+(1-\nu)\tilde\theta_2$. It follows that $\tilde\theta$ belongs to $\Xi(p,\phi)$. Since $\Psi_S(\cdot,\phi)$ is strictly convex in θ for every $\phi\in B(\phi_0,r_n)$ we conclude that

$$\Psi_S(\tilde\theta,\phi)<\nu\Psi_S(\tilde\theta_1,\phi)+(1-\nu)\Psi_S(\tilde\theta_2,\phi)\leq 0$$

where the last inequality is due to the fact that $\tilde\theta_1,\tilde\theta_2\in\Xi(p,\phi)\subset\Theta(\phi)$. Q.E.D.

Lemma G.13. *Let $\tau_0:\mathbb{S}^d\to(0,1)$ be a measurable and differentiable function of p. For every $\phi_1,\phi_2\in B(\phi_0,r_n)$ define $\phi_{\tau_0(p)}=\tau_0(p)\phi_1+(1-\tau_0(p))\phi_2$. Let assumptions 5.1, 5.2, 5.3 (i)-(v), 5.4 (ii) and 5.5 hold. Then, there exists a N such that $\forall n\geq N$ there exists a constant $\bar\lambda(\phi_0)>0$ and an $\varepsilon_n>0$ such that $\varepsilon_n\to0$ as $r_n\to0$ and :*

$$\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)}\sup_{p\in\mathbb{S}^d}\sup_{\theta\in\Xi(p,\phi_{\tau_0(p)})}\|\left[\nabla_\phi\Psi(\theta,\phi_{\tau_0(p)})-\nabla_\phi\Psi(\theta_*(p),\phi_0)\right]^T\lambda(p,\phi_0)\|<\bar\lambda(\phi_0)\varepsilon_n.$$

Proof. First, remark that

$$\begin{aligned}\|\left[\nabla_\phi\Psi(\theta,\phi_{\tau_0(p)})-\nabla_\phi\Psi(\theta_*(p),\phi_0)\right]^T\lambda(p,\phi_0)\|&\leq\|\left[\nabla_\phi\Psi(\theta,\phi_{\tau_0(p)})-\nabla_\phi\Psi(\theta,\phi_0)\right]^T\lambda(p,\phi_0)\|\\&+\|\left[\nabla_\phi\Psi(\theta,\phi_0)-\nabla_\phi\Psi(\theta_*(p),\phi_0)\right]^T\lambda(p,\phi_0)\|\equiv\mathcal{A}_1+\mathcal{A}_2\end{aligned}\tag{G.10}$$

and the function $p\mapsto\lambda(p,\phi_0)$ is continuous in $p\in\mathbb{S}^d$ by lemmas G.9 and G.10. Therefore, it attains its supremum. We start by analyzing the first term $\mathcal{A}_1$:

$$\begin{aligned}&\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)}\sup_{p\in\mathbb{S}^d}\sup_{\theta\in\Xi(p,\phi_{\tau_0(p)})}\|\left[\nabla_\phi\Psi(\theta,\phi_{\tau_0(p)})-\nabla_\phi\Psi(\theta,\phi_0)\right]^T\lambda(p,\phi_0)\|\\&\qquad\leq\sup_{\phi\in B(\phi_0,r_n)}\sup_{\theta\in\Theta}\|\nabla_\phi\Psi(\theta,\phi)-\nabla_\phi\Psi(\theta,\phi_0)\|\sup_{p\in\mathbb{S}^d}\|\lambda(p,\phi_0)\|\end{aligned}$$

since by convexity of $B(\phi_0,r_n)$ we have $\phi_1,\phi_2\in B(\phi_0,r_n)$ implies $\phi_{\tau_0(p)}\in B(\phi_0,r_n)$, $\forall p\in\mathbb{S}^d$. In order to show convergence to zero of this term we follow the proof of lemma G.11, therefore we shorten explanations. By compactness of $B(\phi_0,r_n)$ and under assumption 5.3 (i), the function

$\phi \mapsto \nabla_\phi \Psi(\theta, \phi)$ is uniformly continuous on $B(\phi_0, r_n)$ for every $\theta \in \Theta$. Hence, for every θ and any $\varepsilon_n > 0$ which converges to 0 as $r_n \to 0$ there exists a natural number N_θ that depends on θ such that for all $n \geq N_\theta$ we have

$$f_n(\theta) := \sup_{\phi \in B(\phi_0, r_n)} \|\nabla_\phi \Psi(\theta, \phi) - \nabla_\phi \Psi(\theta, \phi_0)\| < \varepsilon_n. \tag{G.11}$$

Define $A_{N_\theta} := \left\{\tilde{\theta} \in \Theta;\, f_n(\tilde{\theta}) < \varepsilon_n,\, \forall n > N_\theta\right\}$ for every $\theta \in \Theta$. Since $f_n(\theta)$ is continuous in θ, hence A_{N_θ} is an open set and $\bigcup_{\theta \in \Theta} A_{N_\theta}$ is an open cover of Θ. Due to compactness of Θ there exists a finite set $\{N_1, \ldots, N_K\}$, $K < \infty$ such that $\{A_{N_i}\}_{i=1}^K$ is a subcover of Θ: $\Theta \subset \bigcup_{i=1}^K A_{N_i}$.

Let $N^* = \max\{N_1, \ldots, N_K\}$ so that $A_{N_i} \subseteq A_{N^*}$ for every $i = 1, \ldots, K$ and for any $\theta \in \Theta$ we have $\theta \in A_{N^*}$. Remark that this N^* does not depend on θ so for any $n > N^*$,

$$\sup_{\theta \in \Theta} \sup_{\phi \in B(\phi_0, r_n)} \|\nabla_\phi \Psi(\theta, \phi) - \nabla_\phi \Psi(\theta, \phi_0)\| < \varepsilon_n.$$

We conclude that the first term is upper bounded by ε_n.

Now we analyze the second term $\mathcal{A}_2$ of (G.10). We have to consider two cases: *I. for every $(p, \phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$ the set $\Xi(p, \phi)$ is a singleton*; *II. for some $(p, \phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$ the set $\Xi(p, \phi)$ is not a singleton.*

Case I. This case corresponds to the situation where assumption 5.5 *(i)* holds. Hence, the correspondences $\theta_*(p)$ and $\Xi(p, \phi_{\tau_0(p)})$ are single-valued and

$$\begin{aligned}
&\sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| \left[\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\right]^T \lambda(p, \phi_0)\| \\
&\qquad = \sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \| \left[\nabla_\phi \Psi(\Xi(p, \phi_{\tau_0(p)}), \phi_0) - \nabla_\phi \Psi(\Xi(p, \phi_0), \phi_0)\right]^T \lambda(p, \phi_0)\|.
\end{aligned}$$

By Lemma G.4, $\theta(p) = \Xi(p, \phi_{\tau_0(p)})$ is a continuous function of p and by assumption 5.3 (i) the matrix $\nabla_\phi \Psi(\theta, \phi)$ is continuous in θ. Because the composition of two continuous functions is continuous, this implies that $\nabla_\phi \Psi(\Xi(\cdot, \phi_{\tau_0(\cdot)}), \phi_0)$ is a continuous function on $\mathbb{S}^d \times B(\phi_0, r_n)$. Moreover, the compactness of $\mathbb{S}^d \times B(\phi_0, r_n)$ implies that $\nabla_\phi \Psi(\Xi(\cdot, \phi_{\tau_0(\cdot)}), \phi_0)$ is uniformly continuous in (p, ϕ_1, ϕ_2). From this and that $\sup_{B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \|\phi_{\tau_0(p)} - \phi_0\| \leq r_n$, it follows that there exists a $\varepsilon_n > 0$ such that

$$\sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \|\nabla_\phi \Psi(\Xi(p, \phi_{\tau_0(p)}), \phi_0) - \nabla_\phi \Psi(\Xi(p, \phi_0), \phi_0)\| < \varepsilon_n.$$

Remark that this ε_n converge to 0 as $r_n \to 0$. Finally, because $\lambda(p, \phi_0)$ is uniformly bounded in p by some constant, say $\bar{\lambda}_1(\phi_0) > 0$, (by Lemma G.10 and by compactness of $\mathbb{S}^d$) we conclude that

$$\begin{aligned}
&\sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| \left[\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\right]^T \lambda(p, \phi_0)\| \\
&\leq \sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \|\nabla_\phi \Psi(\Xi(p, \phi_{\tau_0(p)}), \phi_0) - \nabla_\phi \Psi(\Xi(p, \phi_0), \phi_0)\| \sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\| = \bar{\lambda}_1(\phi_0)\varepsilon_n.
\end{aligned}$$

Case II. This case corresponds to the situation where assumption 5.5 *(ii)* holds. For some

$\delta_n := \delta(r_n) > 0$ that converges to 0 with r_n define:

$$\mathbb{S}_{ns} := \left\{ p \in \mathbb{S}^d;\ \Xi(p, \phi_0) \text{ is not a singleton } \right\}$$
$$\mathbb{S}_{ns}^{\delta} = \left\{ p \in \mathbb{S}^d;\ \inf_{\tilde{p} \in \mathbb{S}_{ns}} \|p - \tilde{p}\| < \delta_n \right\}$$

and $\mathbb{S}_{ns} \subseteq \mathbb{S}_{ns}^{\delta} \subseteq \mathbb{S}^d$. Therefore,

$$\sup_{p \in \mathbb{S}^d} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| [\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)]^T \lambda(p, \phi_0)\| =$$
$$\sup_{p \in \mathbb{S}_{ns}^{\delta}} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| [\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)]^T \lambda(p, \phi_0)\|$$
$$+ \sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| [\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)]^T \lambda(p, \phi_0)\| =: \mathcal{B}_1 + \mathcal{B}_2 \quad \text{(G.12)}$$

where $(\mathbb{S}_{ns}^{\delta})^c$ denotes the complement of $\mathbb{S}_{ns}^{\delta}$ and is a closed and compact set.

We start by analyzing term $\mathcal{B}_1$. By lemma 5.1 if there exists a subvector of strictly convex constraints $\Psi_S(\cdot, \phi_0)$ of θ then the function $\theta \to \Psi_S(\cdot, \phi)$ is strictly convex for every $\phi \in B(\phi_0, r_n)$. Then, by lemma G.12, for all $p \in \mathbb{S}^d$ for which $\Xi(p, \phi_0)$ is not a singleton we have $\Psi_S(\tilde{\theta}, \phi_0) < 0$ (where the inequality holds componentwise) for some $\tilde{\theta} \in \Xi(p, \phi_0)$. This means that these constraints are not binding and the corresponding Lagrange multipliers, say $\lambda_S(p, \phi_0)$, are equal to 0 (by the complementary slackness condition). Therefore, by the uniqueness of the Lagrange multiplier (see lemma G.9), $\lambda_S(p, \phi_0) = 0$ is the optimum value of the Lagrange multipliers and $\mathcal{B}_1$ simplifies as

$$\sup_{p \in \mathbb{S}_{ns}^{\delta}} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| [\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)]^T \lambda(p, \phi_0)\|$$
$$= \sup_{p \in \mathbb{S}_{ns}^{\delta}} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| [\nabla_\phi \Psi_L(\theta, \phi_0) - \nabla_\phi \Psi_L(\theta_*(p), \phi_0)]^T \lambda_L(p, \phi_0)\|$$
$$= \sup_{p \in \mathbb{S}_{ns}^{\delta}} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| [\nabla_\phi A_2(\phi_0) - \nabla_\phi A_2(\phi_0)]^T \lambda_L(p, \phi_0)\| = 0.$$

Let us consider term $\mathcal{B}_2$. By the result in lemma G.5 with $W = (\mathbb{S}_{ns}^{\delta})^c$, there exists an $\tilde{\varepsilon}_n > 0$ such that $\sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \sup_{\theta \in \Xi(p, \phi)} \|\theta - \theta_*(p)\| < \tilde{\varepsilon}_n$. Moreover, since $\theta \mapsto \nabla_\phi \Psi(\theta, \phi_0)$ is uniformly continuous on Θ (under assumptions 5.1 and 5.2) it follows that for every $\phi \in B(\phi_0, r_n)$ there exists an $\varepsilon_n > 0$ such that

$$\sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \sup_{\theta \in \Xi(p, \phi)} \|\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\| < \varepsilon_n.$$

Since $B(\phi_0, r_n)$ is compact we can easily show, by using a proof similar to that one used in lemma G.11, that $\sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \sup_{\theta \in \Xi(p, \phi)} \|\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\| < \varepsilon_n$.
Therefore, by Lemma G.10 and compactness of $\mathbb{S}^d$ there exists a constant, say $\bar{\lambda}_2(\phi_0) > 0$, such that $\sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \|\lambda(p, \phi_0)\| < \bar{\lambda}_2(\phi_0)$ and $\sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \mathcal{B}_2$ is upper bounded by

$$\sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \sup_{\theta \in \Xi(p, \phi)} \|\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\| \sup_{p \in (\mathbb{S}_{ns}^{\delta})^c} \|\lambda(p, \phi_0)\| < \bar{\lambda}_2(\phi_0)\varepsilon_n.$$

We conclude that

$$\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \sup_{\theta\in\Xi(p,\phi_{\tau_0(p)})} \| \left[\nabla_\phi\Psi(\theta,\phi_0) - \nabla_\phi\Psi(\theta_*(p),\phi_0)\right]^T \lambda(p,\phi_0)\|$$
$$= \sup_{\phi\in B(\phi_0,r_n)} (\mathcal{B}_1+\mathcal{B}_2) < \bar{\lambda}_2(\phi_0)\varepsilon_n.$$

Q.E.D.

Lemma G.14. *For any $\phi_1,\phi_2 \in B(\phi_0,r_n)$ and $\tau\in(0,1)$ define $\phi_\tau = \tau\phi_1 + (1-\tau)\phi_2$ with $\phi_2 = \phi_\tau|_{\tau=0}$ and $\phi_1 = \phi_\tau|_{\tau=1}$. Let assumptions 5.1, 5.2, 5.3 (i)-(iii), 5.3 (v), 5.4 and 5.5 hold with $\delta = r_n$. Then, there exists a N such that $\forall n \geq N$ and for all $p\in\mathbb{S}^d$ we have*

$$\frac{\partial}{\partial\tau} S_{\phi_\tau}(p)\Big|_{\tau=\tau_0(p)} = \lambda(p,\phi_{\tau_0(p)})'\nabla_\phi\Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})[\phi_1-\phi_2] \tag{G.13}$$

where $\tau_0 : \mathbb{S}^d \to (0,1)$ is a measurable and differentiable function of p, $\tilde{\theta}(p)\in\Xi(p,\phi_{\tau_0(p)})$ and $\nabla_\phi\Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})$ denotes the $(k\times d_\phi)$-matrix $\Psi(\theta,\phi)$ evaluated at $(\theta,\phi) = (\tilde{\theta}(p),\phi_{\tau_0(p)})$.

Proof. Define $L(\theta,\lambda,p,\tau_0(p)) =< p,\theta > -\lambda(p,\phi_{\tau_0(p)})'\Psi(\theta,\phi_{\tau_0(p)})$ for $\lambda(p,\phi) : \mathbb{S}^d\times\mathbb{R}^{d_\phi}\to\mathbb{R}^k_+$. By Corollary 5 in Milgrom and Segal (2002) (which can be applied under assumptions 5.1, 5.2 and 5.3 (i)) the function $S_{\phi_{\tau_0(p)}}(p)$ is directionally differentiable in p and its directional derivatives (in direction p) are given by $\frac{d}{dp+}S_{\phi_{\tau_0(p)}}(p) = \max_{\theta\in\Xi(p,\phi_{\tau_0})}\frac{d}{dp}L(\theta,\lambda,p,\tau_0(p))$ and $\frac{d}{dp-}S_{\phi_{\tau_0(p)}}(p) = \min_{\theta\in\Xi(p,\phi_{\tau_0})}\frac{d}{dp}L(\theta,\lambda,p,\tau_0(p))$. For simplicity we have shorten $\tau_0(p)$ with τ_0.

Now, by denoting $\tau_0' := \frac{d\tau_0}{dp}$ we have

$$\frac{d}{dp}L(\theta,\lambda,p,\tau_0(p)) = \theta - \frac{\partial}{\partial p}\lambda(p,\phi_\tau)\Big|'_{\tau=\tau_0(p)}\Psi(\theta,\phi_\tau)|_{\tau=\tau_0(p)}$$
$$-(\phi_1-\phi_2)'\,\frac{\partial}{\partial\phi_\tau}\lambda(p,\phi_\tau)\Big|'_{\tau=\tau_0(p)}\Psi(\theta,\phi_\tau)|_{\tau=\tau_0(p)}\,\tau_0' - \lambda(p,\phi_{\tau_0(p)})'\,\nabla_\phi\Psi(\theta,\phi_\tau)|_{\tau=\tau_0(p)}\,(\phi_1-\phi_2)\tau_0'.$$

Therefore, the partial derivative of $L(\theta,\lambda,p,\tau)$ with respect to its fourth argument is:

$$\frac{d}{d\tau}L(\theta,\lambda,p,\tau) = -(\phi_1-\phi_2)'\left(\frac{\partial}{\partial\phi_\tau}\lambda(p,\phi_\tau)\right)'\Psi(\theta,\phi_\tau) - \lambda(p,\phi_{\tau(p)})'\nabla_\phi\Psi(\theta,\phi_\tau)(\phi_1-\phi_2)$$

and

$$\frac{dS_{\phi_\tau}(p)}{d\tau+}\Big|_{\tau=\tau_0} = \max_{\theta\in\Xi(p,\phi_{\tau_0})}\Big[-(\phi_1-\phi_2)'\,\frac{\partial}{\partial\phi_\tau}\lambda(p,\phi_\tau)\Big|'_{\tau=\tau_0}\Psi(\theta,\phi_\tau)|_{\tau=\tau_0(p)}$$
$$-\lambda(p,\phi_{\tau_0})'\,\nabla_\phi\Psi(\theta,\phi_\tau)|_{\tau=\tau_0}\,(\phi_1-\phi_2)\Big]$$
$$\frac{dS_{\phi_\tau}(p)}{d\tau-}\Big|_{\tau=\tau_0} = \min_{\theta\in\Xi(p,\phi_{\tau_0})}\Big[-(\phi_1-\phi_2)'\,\frac{\partial}{\partial\phi_\tau}\lambda(p,\phi_\tau)\Big|'_{\tau=\tau_0}\Psi(\theta,\phi_\tau)|_{\tau=\tau_0(p)}$$
$$-\lambda(p,\phi_{\tau_0})'\,\nabla_\phi\Psi(\theta,\phi_\tau)|_{\tau=\tau_0}\,(\phi_1-\phi_2)\Big].$$

The first term on the right hand side of both these equations is equal to zero because $\Psi(\theta,\phi_\tau)|_{\tau=\tau_0} = 0$ for $\theta\in\Xi(p,\phi_{\tau_0})$ since this is the first order condition of the optimization problem in $\Xi(p,\phi_{\tau_0})$

evaluated at the optimum value θ. More precisely, it is the partial derivative of the Lagrangian function with respect to the Lagrange multiplier. Thus,

$$\left.\frac{dS_{\phi_\tau}(p)}{d\tau+}\right|_{\tau=\tau_0} = \max_{\theta\in\Xi(p,\phi_{\tau_0})}\Big[-\lambda(p,\phi_{\tau_0})'\,\nabla_\phi\Psi(\theta,\phi_\tau)|_{\tau=\tau_0}(\phi_1-\phi_2)\Big] \tag{G.14}$$

$$\left.\frac{dS_{\phi_\tau}(p)}{d\tau-}\right|_{\tau=\tau_0} = \min_{\theta\in\Xi(p,\phi_{\tau_0})}\Big[-\lambda(p,\phi_{\tau_0})'\,\nabla_\phi\Psi(\theta,\phi_\tau)|_{\tau=\tau_0}(\phi_1-\phi_2)\Big]. \tag{G.15}$$

If $p\in\mathbb{S}^d$ is such that $\Xi(p,\phi_{\tau_0})$ is a singleton then $\left.\frac{dS_{\phi_\tau}(p)}{d\tau+}\right|_{\tau=\tau_0} = \left.\frac{dS_{\phi_\tau}(p)}{d\tau-}\right|_{\tau=\tau_0}$. If $p\in\mathbb{S}^d$ is such that $\Xi(p,\phi_{\tau_0})$ is not a singleton then, by Lemma G.12 there exists a N such that $\forall n\geq N$ and $\phi_1,\phi_2\in B(\phi_0,r_n)$ there is a $\tilde{\theta}\in\Xi(p,\phi_{\tau_0})$ such that $\Psi_S(\tilde{\theta},\phi_{\tau_0})<0$, where Ψ_S denotes the vector of constraints that are strictly convex in θ. This means that these constraints are not binding and the corresponding Lagrange multipliers, say $\lambda_S(p,\phi_{\tau_0})$, are equal to 0 (by the complementary slackness condition). Therefore, by the uniqueness of the Lagrange multiplier (see Lemma G.9), $\lambda_S(p,\phi_{\tau_0})=0$ is the optimum value of the Lagrange multipliers and the term in $\left.\frac{dS_{\phi_\tau}(p)}{d\tau+}\right|_{\tau=\tau_0}$ and $\left.\frac{dS_{\phi_\tau}(p)}{d\tau-}\right|_{\tau=\tau_0}$ simplifies as

$$\lambda(p,\phi_{\tau_0})'\,\nabla_\phi\Psi(\theta,\phi_\tau)|_{\tau=\tau_0} = \lambda_L(p,\phi_{\tau_0})'\,\nabla_\phi\Psi_L(\theta,\phi_\tau)|_{\tau=\tau_0}$$

over $\Xi(p,\phi_{\tau_0})$. By using the expression given in assumption 5.5 *(ii)* for the linear constraints we get $\nabla_\phi\Psi_L(\theta,\phi_\tau)|_{\tau=\tau_0}=\nabla_\phi A_2(\phi_\tau)|_{\tau=\tau_0}$ which does not depend on θ so that $\left.\frac{dS_{\phi_\tau}(p)}{d\tau+}\right|_{\tau=\tau_0} = \left.\frac{dS_{\phi_\tau}(p)}{d\tau-}\right|_{\tau=\tau_0}$ even when $\Xi(p,\phi_{\tau_0})$ is not a singleton.
Q.E.D.

Lemma G.15. *For every $\phi_1,\phi_2\in B(\phi_0,r_n)$ and $\tau\in[0,1]$ define $\phi_\tau:=\tau\phi_1+(1-\tau)\phi_2$ and:*

$$f(\phi_1,\phi_2):=\sup_{p\in\mathbb{S}^d}\left|\Big(\lambda(p,\phi_{\tau_0(p)})^T\nabla_\phi\Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})-\lambda(p,\phi_0)^T\nabla_\phi\Psi(\theta_*(p),\phi_0)\Big)[\phi_1-\phi_2]\right|.$$

where $\tau_0:\mathbb{S}^d\to(0,1)$ is a measurable and differentiable function of p, $\tilde{\theta}(p)\in\Xi(p,\phi_{\tau_0(p)})$ and $\theta_*(p)\in\Xi(p,\phi_0)$.
Let Assumptions 5.1, 5.2, 5.3 (i)-(v), 5.4 (ii) and 5.5 hold with $\delta=r_n$. Then, there exists a constant $C>0$ and an N (independent of ϕ_1 and ϕ_2) such that for every $n>N$

$$\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)}\frac{f(\phi_1,\phi_2)}{\|\phi_1-\phi_2\|}<C\tilde{\varepsilon}_n$$

where $\tilde{\varepsilon}_n\to0$ as $n\to\infty$.

Proof. Remark that $\tilde{\theta}(p)$ depends also on ϕ_1 and ϕ_2. By the Cauchy-Schwartz inequality we can write:

$$\begin{aligned}&\left|\Big(\lambda(p,\phi_{\tau_0(p)})^T\nabla_\phi\Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})-\lambda(p,\phi_0)^T\nabla_\phi\Psi(\theta_*(p),\phi_0)\Big)[\phi_1-\phi_2]\right|\\&\qquad\leq\|\nabla_\phi\Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})^T\lambda(p,\phi_{\tau_0(p)})-\nabla_\phi\Psi(\theta_*(p),\phi_0)^T\lambda(p,\phi_0)\|\;\|\phi_1-\phi_2\|\end{aligned}$$

so that

$$\frac{f(\phi_1,\phi_2)}{\|\phi_1-\phi_2\|} \leq \|\nabla_\phi \Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})^T \left(\lambda(p,\phi_{\tau_0(p)}) - \lambda(p,\phi_0)\right)\| + \\ \|\left(\nabla_\phi \Psi(\tilde{\theta}(p),\phi_{\tau_0(p)}) - \nabla_\phi \Psi(\theta_*(p),\phi_0)\right)^T \lambda(p,\phi_0)\| =: \mathcal{A}_1 + \mathcal{A}_2.$$

We start by analyzing term $\mathcal{A}_1$. Since $B(\phi_0, r_n)$ is convex then $\phi_1, \phi_2 \in B(\phi_0, r_n)$ implies $\phi_\tau \in B(\phi_0, r_n)$, $\forall \tau \in [0,1]$. Thus,

$$\begin{aligned}
\sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \mathcal{A}_1 &= \sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \|\nabla_\phi \Psi(\tilde{\theta}(p),\phi_{\tau_0(p)})^T \left(\lambda(p,\phi_{\tau_0(p)}) - \lambda(p,\phi_0)\right)\| \\
&\leq \sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \sup_{\tau\in[0,1]} \sup_{\theta\in\Xi(p,\phi_\tau)} \|\nabla_\phi \Psi(\theta,\phi_\tau)^T (\lambda(p,\phi_\tau) - \lambda(p,\phi_0))\| \\
&\leq \sup_{\phi\in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \sup_{\theta\in\Xi(p,\phi_\tau)} \|\nabla_\phi \Psi(\theta,\phi)\| \sup_{\phi\in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \|\lambda(p,\phi) - \lambda(p,\phi_0)\| \\
&\leq \sup_{\phi\in B(\phi_0,r_n)} \sup_{\theta\in\Theta} \|\nabla_\phi \Psi(\theta,\phi)\| \sup_{\phi\in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \|\lambda(p,\phi) - \lambda(p,\phi_0)\|.
\end{aligned}$$

By assumptions 5.2 and 5.3 (i), $\nabla_\phi \Psi$ exists and is continuous in $(\theta, \phi) \in \Theta \times B(\phi_0, r_n)$. Since Θ and $B(\phi_0, r_n)$ are compact it follows that $\|\nabla_\phi \Psi(\theta, \phi)\|$ is uniformly bounded on $\Theta \times B(\phi_0, r_n)$, that is, there exists a constant $\bar{\psi}(\phi_0) > 0$ such that $\sup_{\phi\in B(\phi_0,r_n)} \sup_{\theta\in\Theta} \|\nabla_\phi \Psi(\theta, \phi_\tau)\| < \bar{\psi}(\phi_0)$. By lemma G.11 there exists $\epsilon_n := \epsilon(r_n) > 0$, $\epsilon_n \to 0$ as $r_n \to 0$ such that:

$$\sup_{p\in\mathbb{S}^d} \sup_{\phi\in B(\phi_0,r_n)} \|\lambda(p,\phi) - \lambda(p,\phi_0)\| < \epsilon_n$$

for n sufficiently large and we conclude that $\sup_{\phi_1,\phi_2\in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \mathcal{A}_1 < \bar{\psi}(\phi_0)\epsilon_n$, where $\epsilon_n \to 0$ as $n \to 0$. Next, let us consider term $\mathcal{A}_2$:

$$\begin{aligned}
&\sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \|\left(\nabla_\phi \Psi(\tilde{\theta}(p),\phi_{\tau_0(p)}) - \nabla_\phi \Psi(\theta_*(p),\phi_0)\right)^T \lambda(p,\phi_0)\| \\
&\qquad \leq \sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \sup_{\theta\in\Xi(p,\phi_{\tau_0(p)})} \|\left(\nabla_\phi \Psi(\theta,\phi_{\tau_0(p)}) - \nabla_\phi \Psi(\theta_*(p),\phi_0)\right)^T \lambda(p,\phi_0)\|.
\end{aligned}$$

By using the result of lemma G.13 there exists a constant $\bar{\lambda}(\phi_0) > 0$ and an $\varepsilon_n := \varepsilon(r_n) > 0$, $\varepsilon_n \to 0$ as $r_n \to 0$ such that

$$\sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \sup_{\theta\in\Xi(p,\phi_{\tau_0(p)})} \|\left(\nabla_\phi \Psi(\theta,\phi_{\tau_0(p)}) - \nabla_\phi \Psi(\theta_*(p),\phi_0)\right)^T \lambda(p,\phi_0)\| < \bar{\lambda}(\phi_0)\varepsilon_n.$$

By collecting the two upper bounds and by denoting $\tilde{\varepsilon}_n = \epsilon_n + \varepsilon_n$ we get the result.
Q.E.D.

Lemma G.16. *Let $r_n = \sqrt{(\log n)/n}$. Suppose that assumptions 5.1-5.6 hold with $\delta = r_n$. Then, there exists a N such that for every $n \geq N$, $\phi_1, \phi_2 \in B(\phi_0, r_n)$ we have:*

$$\sup_{\phi_1,\phi_2 \in B(\phi_0,r_n)} \sup_{p\in\mathbb{S}^d} \left|\sqrt{n}\left(S_{\phi_1}(p) - S_{\phi_2}(p)\right) - \sqrt{n}\lambda(p,\phi_0)^T \nabla_\phi \Psi(\theta_*(p),\phi_0)[\phi_1 - \phi_2]\right| = o(1)$$

where $\theta_* : \mathbb{S}^d \to \Theta$ is a Borel measurable mapping satisfying $\theta_*(p) \in \Xi(p, \phi_0)$ for all $p \in \mathbb{S}^d$.

Proof. On $B(\phi_0, r_n)$ we have that $\|\phi_1 - \phi_2\| \leq r_n$ for every $\phi_1, \phi_2 \in B(\phi_0, r_n)$. Then, assumption 5.6 (i) implies that $\|\lambda(p, \phi_1) - \lambda(p, \phi_2)\| \leq K_1 r_n$ and therefore

$$\sup_{\phi \in B(\phi_0, r_n)} \|\lambda(p, \phi_1) - \lambda(p, \phi_2)\| \leq K_1 r_n.$$

This implies that the rate ϵ_n in lemma G.11 is $\epsilon_n = O(r_n)$.
A similar argument may be applied to show that $\varepsilon_n = O(r_n)$ in lemma G.13. To this aim, first consider term $\mathcal{A}_1$ in the proof of lemma G.13:

$$\begin{aligned}
&\sup_{p \in \mathbb{S}^d} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| \left[\nabla_\phi \Psi(\theta, \phi_{\tau_0(p)}) - \nabla_\phi \Psi(\theta, \phi_0)\right]^T \lambda(p, \phi_0)\| \\
&\qquad \leq \sup_{p \in \mathbb{S}^d} \sup_{\phi \in B(\phi_0, r_n)} \sup_{\theta \in \Xi(p, \phi)} \|\nabla_\phi \Psi(\theta, \phi) - \nabla_\phi \Psi(\theta, \phi_0)\| \sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\| \\
&\qquad \leq \sup_{\phi \in B(\phi_0, r_n)} \sup_{\theta \in \Theta} \|\nabla_\phi \Psi(\theta, \phi) - \nabla_\phi \Psi(\theta, \phi_0)\| \sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\| = O(\|\phi - \phi_0\|)
\end{aligned}$$

since $\sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\| = O(1)$ (by lemma G.10) and since assumption 5.6 (ii) implies $\sup_{\phi \in B(\phi_0, r_n)} \sup_{\theta \in \Theta} \|\nabla_\phi \Psi(\theta, \phi) - \nabla_\phi \Psi(\theta, \phi_0)\| \leq K_2 \|\phi - \phi_0\|$. Now, consider term $\mathcal{A}_2$ in the proof of lemma G.13. In *Case I* where $\Xi(p, \phi)$ is a singleton for every $(p, \phi) \in \mathbb{S}^d \times B(\phi_0, r_n)$ we have:

$$\begin{aligned}
&\sup_{p \in \mathbb{S}^d} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| \left[\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\right]^T \lambda(p, \phi_0)\| \\
&\qquad \leq \sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \|\nabla_\phi \Psi(\Xi(p, \phi), \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\| \sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\| \\
&\qquad \leq K_3 \sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \|\Xi(p, \phi) - \theta_*(p)\| \sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\|
\end{aligned}$$

under assumption 5.6 (iii). By lemma G.6, under assumption 5.6 (iv), we have an upper bound: $\sup_{\phi \in B(\phi_0, r_n)} \sup_{\theta \in \Theta} \|\Xi(p, \phi) - \theta_*(p)\| = O(r_n)$ so that we conclude that

$$\sup_{p \in \mathbb{S}^d} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| \left[\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\right]^T \lambda(p, \phi_0)\| = O(r_n).$$

For the *Case II* in the proof of lemma G.13, the analysis of term $\mathcal{B}_1$ does not change while for term $\mathcal{B}_2$ we obtain:

$$\begin{aligned}
\mathcal{B}_2 &= \sup_{p \in (\mathbb{S}^\delta_{ns})^c} \sup_{\theta \in \Xi(p, \phi_{\tau_0(p)})} \| \left[\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\right]^T \lambda(p, \phi_0)\| \\
&\leq \sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in (\mathbb{S}^\delta_{ns})^c} \sup_{\theta \in \Xi(p, \phi)} \|\nabla_\phi \Psi(\theta, \phi_0) - \nabla_\phi \Psi(\theta_*(p), \phi_0)\| \sup_{p \in \mathbb{S}^d} \|\lambda(p, \phi_0)\| \\
&\leq K_3 \sup_{\phi \in B(\phi_0, r_n)} \sup_{p \in (\mathbb{S}^\delta_{ns})^c} \sup_{\theta \in \Xi(p, \phi)} \|\theta - \theta_*(p)\| O(1) = O(r_n)
\end{aligned}$$

under assumptions 5.6 (iii) and (iv) and by lemma G.6.
By replacing these rates in the proof of lemma G.15 we get

$$\sup_{p \in \mathbb{S}^d} \left| \left(\lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)}) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)\right) [\phi_1 - \phi_2] \right| = O(r_n \|\phi_1 - \phi_2\|).$$

This and (E.4) give: $\forall \phi_1, \phi_2 \in B(\phi_0, r_n)$,

$$\begin{aligned}
&\sup_{p \in \mathbb{S}^d} \left|(S_{\phi_1}(p) - S_{\phi_2}(p)) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)[\phi_1 - \phi_2]\right| \\
= \quad & \sup_{p \in \mathbb{S}^d} \left| \left(\lambda(p, \phi_{\tau_0(p)})^T \nabla_\phi \Psi(\tilde{\theta}(p), \phi_{\tau_0(p)}) - \lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0) \right) [\phi_1 - \phi_2] \right| = O(r_n^2),
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\phi_1, \phi_2 \in B(\phi_0, r_n)} \sup_{p \in \mathbb{S}^d} \left|\sqrt{n}\left(S_{\phi_1}(p) - S_{\phi_2}(p)\right) - \sqrt{n}\lambda(p, \phi_0)^T \nabla_\phi \Psi(\theta_*(p), \phi_0)[\phi_1 - \phi_2]\right| \\
&\qquad = \quad O(\sqrt{n} r_n^2) = O\left(\frac{\log(n)}{\sqrt{n}}\right)
\end{aligned}$$

which converges to 0.

References

Aliprantis, C.D. and Border, K.C. (2006). *Infinite dimensional analysis - A Hitchhiker's Guide.* Springer-Verlag, Berlin.

Amewou-Atisso, M., Ghosal, S. and Ghosh, J. (2003) Posterior consistency for semiparametric regression problems. *Bernoulli* **9**, 291-312.

Andrews, D. and Jia, P. (2008). Inference for parameters dened by moment inequalities: a recommended moment selection procedure. *Manuscript.* Yale Univ.

Andrews, D. and Guggenberger, P. (2009). Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. *Econometric Theory.* **25**, 669-709.

Andrews, D. and Soares, G. (2010). Inference for parameters dened by moment inequalities using generalized moment selection. *Econometrica.* **78**, 119-157.

Beresteanu, A., Molchanov, I. and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, **79**, No. 6, 1785-1821.

Beresteanu, A., Molchanov, I. and Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics*, **166**, 17-32.

Beresteanu, A. and Molinari, F. (2008) Asymptotic properties for a class of partially identified models. *Econometrica*, **76**, 763-814.

Bickel, P. and Kleijn, B. (2011). The semiparametric Bernstein-Von Mises theorem. *Annals of Statistics*, **40**, 206-237.

Birgé, L. (1983) Approximation dans les espaces mtriques et thorie de l'estimation. *Z. Wahrsch, Verw. Gebiete* **65**, 181-237.

Bollinger, C. and Hasselt, M. (2009) A Bayesian analysis of binary misclassification: inference in partially identified models. *Manuscript.* University of Kentucky.

Bontemps, C., Magnac, T. and Maurin, E. (2011). Set identified linear models. *Econometrica*, forthcoming.

Bugni, F. (2010) Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set. *Econometrica.* **78**, 735-753.

Canay, I. (2010) EL Inference for partially identified models: large deviations optimality and bootstrap validity. *Journal of Econometrics.* **156**, 408-425.

Castillo, I. (2008) Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics.* **2**, 1281-1299.

Chandrasekhar, A., Chernozhukov, V., Molinari, F. and Schrimpf, P. (2012) Inference for best linear approximations to set identified functions. *Manuscript.* MIT.

Chernozhukov, V., Kocatulum, E. and Menzel, K. (2008) Inference on sets in finance. *Manuscript.* MIT.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econometrics,* **115**, 293-346.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica,* **75**, 1243-1284.

Chiburis, R. (2009). Approximately most powerful tests for moment inequalities. *Manuscript.* UT Austin.

Epstein, L. and Seo, K. (2011) Bayesian inference and non-Bayesian prediction and choice: foundations and an application to entry games with multiple equilibria. *Manuscript.* Boston Univ.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, No.2, 209-230.

Florens, J.P. (2002). Inférence Bayésienne non paramétrique et bootstrap. In *Méthoses Bayésiennes en statistique* (eds. J.J. Droesbeke, J. Fine & G. Saporta), 295-313. Editions TECHNIP.

Florens, J.P. and Simoni, A. (2011). Bayesian identification and partial identification, preprint.

Florens, J.P. and Simoni, A. (2012). Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. *Journal of Econometrics.* **170**, 458-475

Gelfand, A. and Sahu, S. (1999) Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association.* **94**, 247-253.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Annals of Statistics.* **29**, 1264-1280.

Ghosal, S., Ghosh, J. and Ramamoorthi, R. (1999) Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics.* **27**, 143-158.

Ghosal, S., Ghosh, J. and van der Vaart, A. (1999) Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference* **77**, 181-193.

Ghosal, S., Ghosh, J. and van der Vaart, A. (2000) Convergence rates of posterior distributions. *Annals of Statistics.* **28**, 500-531.

Ghosal, S. and van der Vaart, A. (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics.* **29**, 1233-1263.

Ghosal, S. and van der Vaart, A. (2007) Convergence rates of posterior distributions for noniid obsevations. *Annals of Statistics.* **35**, 192-223.

Ghosh, J.K. and Ramamoorthi, R.V.(2003). *Bayesian Nonparametrics.* Springer-Verlag.

Gustafson, P. (2005) On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science.* **20**, 111-140.

Gustafson, P. (2012) On the behaviour of Bayesian credible intervals in partially identified models. *Electronic Journal of Statistics.* **6**, 2107-2124.

Haile, P. and Tamer, E. (2003) Inference with an incomplete model of English auctions. *Journal of Political Economy.* **111**, 1-51.

Hansen, L. and Jagannathan, R. (1991) Implications of security market data for models of dynamic economies. *Journal of Political Economy.* **99**, 225-262.

Imbens, G. and Manski, C. (2004) Confidence Intervals for partially identified parameters. *Econometrica.* **72**, 18451857.

Jiang, W. (2007) Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Annals of Statistics.* **35**, 1487-1511.

Kaido, H. and Santos, A. (2011). Asymptotically efficient estimation of models defined by convex moment inequalities, preprint.

Kaido, H. (2012). A dual approach to inference for partially Identified econometric models, preprint.

Kato, T. (1995). *Perturbation theory for linear operators.* Springer-Verlag Berlin Heidelberg.

Kim, J. (2002) Limited information likelihood and Bayesian analysis. *Journal of Econometrics.* **107**, 175-193.

Kitagawa, T. (2012) Estimation and inference for set-identified parameters using posterior lower probability. *Manuscript.* UCL.

Kline, B. (2011) The Bayesian and frequentist approaches to testing a one-sided hypothesis about a multivariate mean. *Journal of Statistical Planning and Inference.* **141**, 3131-3141.

Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *Ann. Statist.* **20**, No.3, 1222-1235.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

Liao, Y. and Jiang, W. (2010) Bayesian analysis in moment inequality models. *Annals of Statistics.* **38**, 275-316.

Liao, Y. and Jiang, W. (2011) Posterior consistency of nonparametric conditional moment restricted models. *Annals of Statistics.* **39**, 3003-3031.

Lindsay, B. and Basak, P. (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of American Statistical Association* **88**, 468-476.

Manski, C. and Tamer, E. (2002) Inference on regressions with interval data on a regressor or outcome. *Econometrica.* **70**, 519-547.

Manski, C. (2003) *Partial identification of probability distributions.* Spinger, New York.

Menzel, K. (2011) Consistent estimation with many moment inequalities. *Manuscript.* NYU.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, **70**, 583-601.

Molchanov, I. (2005). *Theory of random sets. Springer.*

Moon, H. R. and Schorfheide, F. (2012). Bayesian and frequentist inference in partially-identified models. *Econometrica*, **80**, 755-782.

Neath, A. and Samaniego, F. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *The American Statistician* **51**, 325-332.

Norets, A. and Tang, X. (2012) Semiparametric Inference in dynamic binary choice models. *Manuscript.* Princeton Univ.

Otsu, T. (2006) Large deviation optimal inference for set identified moment inequality models. *Manuscript.* Yale Univ.

Poirier, D. (1998) Revising beliefs in nonidentified models. *Econometric Theory.* **14**, 483-509.

Ray, S. and Lindsay, B. (2005) The topography of multivariate normal mixtures. *Annals of Statistics.* **33**, 2042-2065.

Rivoirard, V. and Rousseau, J. (2012). Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist.*, forthcoming.

Rockafellar, R.T. (1970). *Convex analysis. Princeton University Press.*

Romano, J. and Shaikh, A. (2010) Inference for the identified set in partially identified econometric models. *Econometrica.* **78**, 169-211.

Rosen, A. (2008) Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics.* **146**, 107-117.

Rudin, W. (1986). *Real and complex analysis. McGraw-Hill Science/Engineering/Maths*

SCHENNACH, S. (2005) Bayesian exponentially tilted empirical likelihood. *Biometrika.* **92**, 31-46.

SCHWARTZ, L. (1965) On Bayes procedures. *Z. Wahrsch, Verw. Gebiete* **4**, 10-26..

SETHURAMAN, J. (1994). A constructive definition of the Dirichlet prior. *Statistica Sinica*, **2**, 639-650.

SHEN, X. and WASSERMAN, L. (2001) Rates of convergence of posterior distributions. *Annals of Statistics.* **29**, 687-714.

STOYE, J. (2009) More on confidence intervals for partially identified parameters. *Econometrica* **77**, 1299-1315,

STOYE, J. (2012) New perspectives on statistical decisions under ambiguity. *Annual Review of Economics* **4**, 257-282.

TAMER, E. (2010) Partial identification in econometrics. *Annual Review of Economics* **2**, 167-195.

VAN DER VAART, A.W. (2002). Semiparametric Statistics. In *Ecole d'Eté de St Flour 1999* (eds. P. Bernard), 331-457. Springer Verlag, New York.

VAN DER VAART, A.W. (2000). *Asymptotic Statistics. Cambridge University Press.*

VAN DER VAART, A. and WELLNER, J. (1996) *Weak convergence and empirical processes Springer.*

VAN DER VAART, A.W. and J.H., VAN ZANTEN (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, **36**, No.3, 1435-1463.

WALKER, S., LIJOI, A. and PRUNSTER, I. (2007) On rates of convergence for posterior distributions in infinite-dimensional models. *Annals of Statistics.* **35**, 738-746.

WAN, Y. (2011) An integration-based approach to moment inequality models. *Manuscript.* Univ. of Toronto.

WU, Y. and GHOSAL, S. (2008) Posterior consistency for some semi-parametric problems. *Sankhya: the Indian Journal of Statistics.* **70**, 0-46.