Essays on empirical likelihood in economics

Gao, Z.

Publication date
2012

4

GEOMETRIC INTERPRETATIONS FOR CONSTRAINED GMC AND GEL

Constrained optimization, as a general mathematical tool, can narrow the potential parameter space and refine irregular problems in statistical models. In econometrics, this tool often consists of moment constraints and various criterion functions[1], such as Mahalanobis distance (GMM Hansen, 1982), log-likelihood ratio (Empirical Likelihood, hereafter EL, Owen, 2001), and Kullback-Leibler divergence (Kitamura and Stutzer, 1997). Smith (1997) and Newey and Smith (2004) show that using a general representation, many criterion functions with moment constraints can be covered in the framework of Generalized Empirical Likelihood (GEL). The statistics of members in GEL share similar asymptotic properties. A class competing with GEL is the Generalized Minimum Contrast (GMC) class which has been discussed in Kitamura (2006). This chapter will consider a subclass in GMC whose members have similar properties as those in GEL. By geometry techniques, we connect the representations of GEL and GMC through polyhedral approximations and then we show that a sub-Sobolev class, called $\mathcal{W}$-class, in GMC will have similar properties as GEL and will obtain standard weak convergence result. The result simply induces the Fisher type efficiency.

A discussion about the connection between GEL and GMC has been given by Kitamura (2006). Kitamura (2006) shows that the standard inferential problems in GEL, e.g. EL, Exponential Tiling or continuous updating GMM, can be also presented in the GMC setup. GMC has a primal-dual representation, while GEL is based on the dual parameters of the moment constraints. It implies that GMC has a standard mathematical programming representation as its primal form. This representation is impor-

1 The word "criterion function" has the same meaning as the objective function in constrained optimization problems. Thereafter, these two terminologies are used simultaneously.

tant for both theoretical analyses and practical implementations. However, the Fisher type efficiency of GEL is not derivable if one concentrates on the general form of GMC. The reason is that GMC consists of some criterion functions which are incomparable to those in GEL in the Fisher type information metric. Then what could be the primal representation of GEL? This chapter is intended to give an explicit clue.

4.1 THE MODEL

We consider a criterion function $\rho(\cdot,\cdot)$ that is a contrast function measuring the discrepancy between its inputs. The inputs are the densities or the likelihoods of a general probability family $\mathcal{P}$ such that the Radon–Nikodym derivative exists with respect to (w.r.t.) the counting measure μ:

$$\mathcal{P}_\mu := \left\{ pd\mu |\, p = \frac{dP}{d\mu},\, P \in \mathcal{P} \right\}$$

So $\rho(\cdot,\cdot) : \mathcal{P}_\mu \times \mathcal{P}_\mu \mapsto \mathbb{R}$. It is not necessary for the contrast function ρ to be a metric. If $\mathcal{P}$ consists of parametric families, we will use the notation $\mathcal{P}_\theta := \{p_\theta d\mu \in \mathcal{P}_\mu, \theta \in \mathbb{R}^d\}$ and let p_0 denote the true density. The minimum dissimilarity between p_θ and p_0 is attainable at $\theta = \theta_0$ w.r.t. $\rho(\cdot,\cdot)$.

The criterion functions in this chapter belong to a general class, $\mathcal{W}$ which is a subspace of the *Sobolev space* $\mathcal{W}^{k,2}$ with $k \geq 2$. To be more specific, any $\rho \in \mathcal{W}^{k,2}$, satisfies:

$$\left(\sum_{j=0}^{k} \left\| \rho^{(j)} \right\|^2 \right)^{1/2} < \infty,$$

where $\rho^{(j)}$ means the j-th (functional) derivative[2] of ρ and $\|\cdot\|$ is L^2-norm. The subspace $\mathcal{W}$-class that we will use only involves those ρs that have the same order as squared Hellinger distance in the sense that for $p_\theta,\, p_0 \in \mathcal{P}_\theta$:

$$\lim_{\theta \to \theta_0} \frac{\rho(p_\theta,\, p_0)}{\rho^{(H)}(p_\theta,\, p_0)} = \frac{\rho(p_{\theta_0},\, p_0)}{\rho^{(H)}(p_{\theta_0},\, p_0)} = c,$$

where c is a constant and $\rho^{(H)}(p,\, g)$ is squared Hellinger distance:

$$\rho^{(H)}(p,\, g) := \frac{1}{2} \int (\sqrt{p} - \sqrt{g})^2 d\mu.$$

2 The notation $\rho^{(j)}(x,y)$ means that $\frac{\partial^j \rho(x,y)}{\partial x^p y^q}$ with any positive p, q and $p+q=j$.

We assume a sample space $\mathcal{X}$. The model specifies d moment restrictions $\mathbb{E}[m(X,\theta)] = 0$ for a unique $\theta_0 \in \mathbb{R}^d$ where $\mathbb{E}[\cdot]$ is taken w.r.t. $p_0 d\mu$ so the model is exactly identified[3]. The estimator is exactly the solution[4] to the moment condition $\sum_{i=1}^n m(x_i,\theta) = 0$. With a sample of n observations we can define the empirical support $\mathcal{X}_n$ as all those points in $\mathcal{X}$ that have been observed in the sample. On $\mathcal{X}_n$ we can define an empirical measure using the Dirac function δ_n such that $\delta_n(x) = 1$ for $x \in \mathcal{X}_n$ and $\delta_n(x) = 0$ otherwise. We consider a mathematical programming problem:

$$(M) := \begin{cases} \min_{\mathbf{p}_n,\theta} & \rho(p_\theta(x), \delta_n(x)n^{-1}) \text{ for any } x \in \mathcal{X}_n \\ s.t. & \mathbf{p}_n = (p_1,\ldots,p_n)^T \in \mathbb{S}_{n-1}, \sum_{i=1}^n p_i m(x_i,\theta) = \mathbf{0}. \end{cases}$$

where $p_\theta(x_i) = p_i$, $\mathbb{S}_{n-1}$ is a regular $(n-1)$-simplex $\{\mathbf{p}_n : \sum_i^n p_i = 1, p_i \geq 0 \text{ for all } i\}$. For every θ, $p_\theta(x_i) = p_i$ is the (weighted empirical) density for x_i under θ. For simplicity, throughout the chapter, we assume that the model (M) satisfies *common regularity assumptions*.

Assumption 4.1.
(i) $N^{-1}\sum_{i=1}^N [m(x_i,\theta)^T m(x_i,\theta)]$ *has finite variance.*
(ii) $\lambda^T m(x_i,\theta)$ *has finite variance where* λ *is introduced in* (4.1).
(iii) $m(\cdot,\theta)$ *belongs to the P-Donsker class.*
(iv) $\Theta \subset \mathbb{R}^d$ *is compact.*
(v) $m(X,\theta)$ *is continuous w.r.t.* θ.

Remark 4.2. The $\mathcal{W}$-class is a sub-class in the GMC. The smoothness requirement of ρ is a crucial feature because the smoothness is a necessary condition for defining a classical information metric in the sense of Fisher. There are many other options for measuring the divergence of distributions. For example, f-divergence is a rather broad class in GMC. However,

3 Note that even if $\theta = \theta_0$, p_{θ_0} is not necessarily equivalent to p_0, since p_{θ_0} is simply a pseudo true density of $m(X,\theta_0)$ and θ_0 itself does not fully capture the parameter information of the distribution of X or $m(X,\theta_0)$. For example, if p_0 is Poisson or log-Gaussian, the implied probability of EL or ET

$$\tilde{p}_{\theta_0}(x_i) = \frac{1}{n}\frac{1}{1+\lambda^T m(x_i,\theta_0)}, \text{ or } \tilde{p}_{\theta_0}(x_i) = \frac{e^{\lambda^T m(x_i,\theta_0)}}{\sum_i e^{\lambda^T m(x_i,\theta_0)}}$$

will not be equivalent to p_0 no matter how many $m(X,\theta_0)$ are used.

4 This chapter is intended to discuss the connection between GEL and GMC. While the exact identification case does diminish the advantages of using over-identified moment constraints in GEL or GMC, the implied density and the analysis based on this implied density do not rely on whether the constraints are over-identified or not.

f-divergence cannot guarantee the smoothness of ρ, e.g., the total variation distance, a special case in the f-divergence. We consider $\mathcal{W}$-class instead of f-divergence because of the important role of the smoothness condition.

Remark 4.3. Many essential contrast functionals belong to $\mathcal{W}$-class. Besides Hellinger distance, Likelihood ratio and χ^2, there are Kullback-Leibler divergence,

$$\rho^{(KL)}(p, g) := \int \left(\log \frac{p}{g}\right) p d\mu$$

Mahalanobis distance,

$$\rho^{(M)}(p, g) := \int (p - g)^2 w^{-1} d\mu$$

Jeffreys divergence, Chernoff information divergence, etc. Eguchi (1985) introduces this class from a differential geometry point of view in statistics.

4.2 DUAL REPRESENTATION ON CONVEX BODIES: CONSTRAINTS

Neither closed form solutions nor explicit expressions of the optimal $p_\theta(x)$ in (M) are available. However, one can get the first-order-condition (FOC) by the Lagrangian method:

$$\mathcal{L}_n = \rho(p_\theta(x), \delta_n(x)n^{-1}) - \lambda^T \left(\sum_{i=1}^n p_i m(x_i, \theta)\right) - \zeta(1 - \sum_i^n p_i). \tag{4.1}$$

Taking derivative of (4.1) w.r.t. $\mathbf{p}_n$, one has the FOC:

$$\nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) = \lambda^T m(x_i, \theta) - \zeta, \tag{4.2}$$

where ∇_i is gradient w.r.t. p_i. If we multiply $\{p_1, \ldots, p_n\}$ on both sides and add up these equations, we have

$$\sum_{i=1}^n p_i \nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) = -\zeta,$$

where $\lambda^T m(x_i, \theta)$ drops out because $\sum_i p_i m(x_i, \theta) = 0$. Just like EL, if we substitute the expression for ζ into (4.2), we have:

$$\nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) - \sum_{i=1}^n p_i \nabla_i \rho(p_\theta(x), \delta_n(x)n^{-1}) = \lambda^T m(x_i, \theta). \tag{4.3}$$

Equation (4.3) shows that the optimal p_θ, the solution of this nonlinear differential equation, is a function of λ. In general, there is no closed form expression for $p_\theta(\lambda)$. A straightforward *conjecture* about the solution of (4.3) is that $p_\theta(\lambda)$ is a function[5] of $\lambda^T m(x_i,\theta)$ only. The following Proposition is to formalize this conjecture.

Proposition 4.4. *The density $p_\theta(\lambda)$ in the problem (M) only depends on $\lambda^T m(x_i,\theta)$ for any $1 \leq i \leq n$.*

One can easily relate this result to GEL (Smith, 1997; Newey and Smith, 2004) whose definition is based on a criterion function of $\lambda^T m(x_i,\theta)$. The setup of GEL is:

$$\hat{\theta} = \arg\min_\theta \max_\lambda \left[\frac{1}{n} \sum_{i=1}^{n} \tilde{\rho}(\lambda^T m(x_i,\theta)) \right]$$

where $\tilde{\rho}(\cdot)$ is strictly concave and $\tilde{\rho}^{(1)}(0) = \tilde{\rho}^{(2)}(0) = -1$. The notation $\tilde{\rho}^{(j)}(0)$ means the j-th derivative of $\tilde{\rho}$ at zero.

As the optimal solution $p_\theta(\lambda)$ in the problem (M) only depends on $\lambda^T m(x_i,\theta)$, we can see that problem (M) is exactly the same as GEL except a different definition of the objective function ρ. However, since ρ is in the $\mathcal{W}$-class, it is smooth, differentiable, continuous and approximately quadratic. Hence, the fact that the objective function is in the $\mathcal{W}$-class plays no additional role other than in GEL in proving consistency. The consistency result for GEL (Newey and Smith, 2004) or for minimum Hellinger distance (Kitamura et al., 2009) can be applied directly to the estimator in the $\mathcal{W}$-class. As a consequence, we have following Proposition which we state without proof.

Proposition 4.5. *For any estimator $\hat{\theta}$ such that*

$$\rho(p_{\hat{\theta}}(x), \delta_n(x) n^{-1}) \leq \rho(p_{\theta_0}(x), \delta_n(x) n^{-1}) + o_p(1),$$

we have for any $\epsilon > 0$, $\Pr(\|\hat{\theta} - \theta_0\| \leq \epsilon) \to 1$.

Remark 4.6. We briefly describe the geometric meaning of Proposition 4.4 but for details, please refer to the proof in the appendix. The geometric role of the moment constraints in (M) is to construct a convex hull of $m(X,\theta)$, $\mathrm{conv}(m(X,\theta))$. The convex hull is approximated by some geometric objects. Mathematically speaking, the border of $\mathrm{conv}(m(X,\theta))$ is empirically approximated by the polyhedron, $\mathcal{P}_m = \{(x_1,\ldots,x_n,\mathbf{b}) | \lambda^T m(x_i,\theta) \leq$

5 Because fundamental solution of the ordinary differential equation $dX/dt = g(X)$ is a function of $g(X)$.

$\mathbf{b}\}$ for the given sample, where $\mathbf{b}$ is chosen optimally. The dual problem of (M) is a linear programming problem of the polyhedron set $\mathcal{P}_m$. Then the optimal solution of this linear programming problem will give a unique representation of $\text{conv}(m(X,\theta))$ and the solution will only depend on $\lambda^T m(x_i,\theta)$. The density $p_\theta(x)$ is the *generator* of $\text{conv}(m(X,\theta))$, so the optimal $p_\theta(x_i)$ will also only depend on $\lambda^T m(x_i,\theta)$.

4.3 WEAK CONVERGENCE ON THE UNIT SPHERE: CRITERION FUNCTIONS

The criterion function ρ plays an important role in weak convergences. The distance between two points in $\mathcal{P}_\theta$ measures the amount of their dissimilarity. The functional form of ρ defines the criterion of such a measurement. We will show that the geometric analysis of weak convergence for the $\mathcal{W}$-class is simple and intuitive.

Given a hypothesis testing problem $H_0 : \theta = \tilde{\theta}$, the Hellinger distance on $\mathcal{P}_\theta$ between p_0 and the empirical distribution $\delta_n(x)n^{-1}$ is:

$$\rho^{(H)}(p_{\tilde{\theta}}(x),\, \delta_n(x)n^{-1}) = -\frac{1}{2}\sum_{i=1}^{n}\left[\sqrt{p_i^*} - \sqrt{\frac{1}{n}}\right]^2,$$

where the notation p_i^* is the solution of problem (M) such that $p_{\tilde{\theta}}(x_i) = p_i^*$ given $\theta = \tilde{\theta}$. Our geometric interpretation of $\rho^{(H)}$ in (M) is the metric distance over a unit sphere $\mathcal{S}^{n-1}$, a n-dimensional Hilbert space $\mathcal{H}$. Note that any square-root likelihood $\xi_i := (p_i^*)^{\frac{1}{2}}$ satisfies the normalizing condition $\sum_i^n \xi_i^2 = 1$ and is nonnegative and is located within $[0,1]$. Thus, ξ can be regarded as a unit vector in the Hilbert space $\mathcal{S}^{n-1}$. The empirical square-root measure $\delta_n(x)n^{-\frac{1}{2}}$, considered as a special case of $(p_i^*)^{\frac{1}{2}}$, also belongs to this space. For any two sequences $\xi^{(1)}$ and $\xi^{(2)}$, the Hellinger distance induces an inner product of this space

$$\cos\beta := \langle \xi_a, \xi_b \rangle = \sum_i^n \xi_i^{(1)}\xi_i^{(2)} = 1 - \frac{1}{2}\sum_{i=1}^{n}\left[\sqrt{\xi_i^{(1)}} - \sqrt{\xi_i^{(2)}}\right]^2$$

which is the so called *Hellinger affinity*. In addition, the double-angle formula $\cos\beta = 1 - 2\sin^2\frac{\beta}{2}$ implies that:

$$4\sin^2\frac{\beta}{2} = \sum_{i=1}^{n}\left[\sqrt{\xi_i^{(1)}} - \sqrt{\xi_i^{(2)}}\right]^2 = -2\rho^{(H)}(p_{\tilde{\theta}}(x),\, \delta_n(x)n^{-1}).$$

We should reiterate that p_θ is a pseudo true density so even $\tilde{\theta} = \theta_0$, $p_{\tilde{\theta}}$ is not necessarily equivalent to p_0.

It is obvious that the angle β can be interpreted as a distance between two probability distributions on $\mathcal{S}^{n-1}$ equipped with the Hellinger metric. The maximal possible distance, corresponding to orthogonal sequences, is given by $\beta = \pi/2$. The double-angle formula clearly shows that the distribution of $\cos\beta$ is equivalent to that of $4\sin^2(\beta/2)$.

Theorem 4.7. *If the hypothesis $H_0 : \theta = \tilde{\theta}$ is true, the statistics $\rho(p_{\tilde{\theta}}(x), \delta_n(x)n^{-1})$ in the $\mathcal{W}$-class converges in distribution to a χ^2_d-distributed random variable when $n \to \infty$.*

Remark 4.8. From basic trigonometrics, we know that when the angle β is very small, $\sin^2\beta \approx \beta^2$. Thus Hellinger distance implies that when $\rho(\cdot,\cdot)$ is small, the distance in the $\mathcal{W}$-class can be described in terms of β^2. Hellinger distance is also an important concept when one defines Fisher's information metric in the spherical geometry. The inner product $\cos\beta$ could be written as $\langle \xi_a, \xi_b \rangle_{g_{ab}} := \int g_{ab}\xi_a\xi_b d\mathcal{S}^{n-1}$ where $d\mathcal{S}^{n-1}$ is volume measure on the sphere associated with the Riemannian metric g_{ab}. If the derivatives of ξ_a and ξ_b are available[6], then $4g_{ab}\partial\xi_a\partial\xi_b$ is the Fisher information metric which can be used to see whether the efficiency bound can be attained on this sphere $\mathcal{S}^{n-1}$.

4.4 CONCLUSION

In this chapter, we show that the GEL problem has a dual representation, a constrained GMC problem for the $\mathcal{W}$ contrast functional class. This connection is established from a geometric perspective. The dual representation transfers the task of estimating implied densities for GEL to a task of solving a standard mathematical programming problem. We also show that the statistics in the $\mathcal{W}$-class share the same first order statistical properties with GEL.

6 The derivative $\partial\xi$ is $d\sqrt{p} = \sqrt{p}d\ell$ where $d\ell$ is the functional derivative of the empirical log-likelihood in semi-parametric models. $(\partial_1\ell, \ldots, \partial_r\ell)^T$ is a co-ordinate basis on the tangent space of $\mathcal{P}_\theta$, $T\mathcal{P}_\theta$, hence one can use it to construct an inner product on $T\mathcal{P}_\theta$.

APPENDIX TO CHAPTER 4

Let $m(\mathbf{X}_n,\theta) = (m(x_1,\theta),\ldots,m(x_n,\theta))$ where $m(x_i,\theta)$ is a $d \times 1$ vector.

Proof of Proposition 4.4

Proof. (*Sufficient condition*) The constraint constructs a set of $m(\mathbf{X}_n,\theta)$, the convex hull of a finite set $\{p_1,\ldots,p_n\}$,:

$$\text{conv}(m(\mathbf{X}_n,\theta)) := \left\{ \sum_{i=1}^{n} p_i m(x_i,\theta) \,|, \sum_{i=1}^{n} p_i = 1, p \succeq 0 \right\},$$

which is called a finitely generated convex set. The symbol $\succeq$ defines a vector inequality in $\mathbb{R}^n$ such that $p_i \geq 0$ for $i = 1,\ldots n$. The set $\text{conv}(m(\mathbf{X}_n,\theta))$ is a *polyhedron* $\mathcal{M}$ (Theorem 19.1 Rockafellar, 1996), namely an intersection of finitely many closed half spaces:

$$\text{conv}(m(\mathbf{X}_n,\theta)) = \mathcal{M} := \left\{ m(x,\theta) \,|\, \lambda_p^T m(x,\theta) \succeq \mathbf{b} \right\},$$

where λ_p^T and $\mathbf{b} = (b_1,\ldots,b_n)^T$ define the half-space for each $m(x_i,\theta)$ such that $\{\lambda_p^T m(x_i,\theta) \geq b_i\}$. Define a function

$$\delta_{\mathcal{M}}(\theta) := \mathbb{I}\{m(\mathbf{X}_n,\theta) \in \mathcal{M}\}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function. The notation

$$\{m(\mathbf{X}_n,\theta) \in \mathcal{M}\} := \{m(x_1,\theta),\ldots,m(x_n,\theta) \in \mathcal{M}\}$$

means that $m(x_i,\theta)$ belongs to $\mathcal{M}$ for $0 < i \leq n$. Laplace's inequality states that $\mathbb{I}\{y\} \leq \exp(y)$ for any $y \geq 0$. The polyhedron set introduces an exponential bound such that:

$$\delta_{\mathcal{M}}(\theta) \preceq \exp\left(\lambda_p^T m(\mathbf{X}_n,\theta) - \mathbf{b}\right).$$

Taking expectation w.r.t. $\mathbf{X}_n$, we have the Chernoff bound

$$\Pr\left(m(\mathbf{X}_n,\theta) \in \mathcal{M}\right) \preceq \mathbb{E}\exp\left(\lambda_p^T m(\mathbf{X}_n,\theta) - \mathbf{b}\right), \tag{4.4}$$

which implies the probability p_θ is controlled by $\lambda_p^T m(\mathbf{X}_n,\theta)$ and $\mathbf{b}$.

Now we need to relate the tangent parameter λ_p in the polyhedron with the multiplier λ in optimization of $\rho(p_\theta(x), \delta_x n^{-1})$.

(*Necessary condition*) By the differentiability of the $\mathcal{W}$-class, the (functional) linearization of $\rho(p_\theta(x), \delta_x n^{-1})$ at the point $p^\#$ is

$$\begin{aligned}\rho(p_\theta(x), \delta_x n^{-1}) =& \rho(p^\#, \delta_x n^{-1}) + (p_\theta - p^\#)^T \nabla \rho(p_\theta, \delta_x n^{-1}) \\ &+ O_p(\|p_\theta - p^\#\|^2).\end{aligned} \tag{4.5}$$

By FOC (4.2), we can replace the gradient vector $\nabla \rho(p_\theta, \delta_x n^{-1})$ in (4.5) by $(\lambda^T m(X,\theta) - \zeta)$. Then within a small region of $p^\#$ where $O_p(\|p_\theta - p^\#\|^2)$ is negligible, minimizing $\rho(p_\theta(x), \delta_x n^{-1})$ in this region is equivalent to minimizing the linear approximation w.r.t. λ, θ and ζ:

$$\mathcal{Q}(p^\#) := \inf_{\lambda,\theta,\zeta} \left\{ \rho(p^\#, \delta_x n^{-1}) + (p_\theta - p^\#) \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] \right\}, \tag{4.6}$$

$$= \inf_{\lambda,\theta,\zeta} \left\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \mathcal{Q}^*(\lambda,\theta,\zeta) \right\}, \tag{4.7}$$

where $\mathcal{Q}^*(\lambda,\theta,\zeta) = p_\theta \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \rho(p^\#, \delta_x n^{-1})$ is the conjugate[7] of $\mathcal{Q}(p^\#)$ in the region around $p^\#$. Then by *Legendre–Fenchel transformation* (conjugate functional) [8], the dual programming of problem (4.6) is:

$$\begin{aligned}[\mathcal{Q}(p^\#)]^* =& \mathcal{Q}^*(\lambda,\theta,\zeta) = \inf_{p^\#} \left\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \mathcal{Q}(p^\#) \right\}, \\ =& \inf_{p^\#} \Bigg\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \\ & \underbrace{\inf_{\lambda,\theta,\zeta} \left\{ -p^\# \left[\lambda^T m(\mathbf{X}_n, \theta) - \zeta \right] + \mathcal{Q}^*(\lambda,\theta,\zeta) \right\}}_{(i)} \Bigg\}.\end{aligned}$$

7 Mathematically, a nonlinear programming problem can be pinned down as many linear programming problems using the linearization technique. In these linear programming problems, finding an optimal linear approximation ($\mathcal{Q}^*(\lambda,\theta,\zeta)$) is a dual or a conjugate problem of finding the optimal parameters for such a linear approximation ($\mathcal{Q}(p^\#)$). The construction of equations (4.6) and (4.7) is a functional *Legendre transformation*.

8 Kitamura (2006) gives a general introduction of Legendre-Fenchel transformation in context of EL and GMC.

We write the expression in a simpler way using a slack variable ω for (i) in the above equation:

$$\inf_{p^{\#},\omega} \quad \left\{ p^{\#} \left[\lambda^T m(\mathbf{X}_n,\theta) - \zeta \right] + \omega \right\} \tag{4.8}$$

$$\text{subject to} \quad p^{\#} \left[\lambda^T m(\mathbf{X}_n,\theta) - \zeta \right] - \mathcal{Q}^*(\lambda,\theta,\zeta) \succeq \omega. \tag{4.9}$$

Constraint (4.9) can be written as $\lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*$ where

$$\mathbf{b}^* = (p^{\#})^{-1}(\omega + \mathcal{Q}^*(\lambda,\theta,\zeta)) + \zeta.$$

The polyhedral $\mathcal{M}$ of $m(\mathbf{X}_n,\theta)$ is $\{m(\mathbf{X}_n,\theta) \,|\lambda_p^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}\}$. Constraint (4.9) sets a polyhedral shape using $\lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*$. The problem becomes

$$(M') := \begin{cases} \inf_{p^{\#},\omega} & \left\{ p^{\#} \left[\lambda^T m(\mathbf{X}_n,\theta) - \zeta \right] + \omega \right\} \\ \text{subject to} & \lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*. \end{cases}$$

The optimal $p^{\#}$ and ω in (M') imply optimal edges of the polyhedron $\lambda^T m(\mathbf{X}_n,\theta) \succeq \mathbf{b}^*$. Since the value of $\mathbf{b}$ in (4.4) is arbitrary, we set $\mathbf{b} \equiv \mathbf{b}^*$, then the multiplier λ is equivalent to the tangent parameter λ_p. Except tuning parameter $p^{\#}$ and ω, the variables in (M') are $\lambda^T m(\mathbf{X}_n,\theta)$, $\mathcal{Q}^*(\lambda,\theta,\zeta)$ and ζ. By definition, $\mathcal{Q}^*(\lambda,\theta,\zeta)$ depends on $\lambda^T m(\mathbf{X}_n,\theta)$ and ζ. From the FOC, ζ depends on $\lambda^T m(\mathbf{X}_n,\theta)$. Hence for problem (M'), $\lambda^T m(\mathbf{X}_n,\theta)$ is the only necessary element and $p^{\#}$ will depend on $\lambda^T m(\mathbf{X}_n,\theta)$ only. Note that $p^{\#}$ is the conjugate of p_θ.

By Chernoff bound (4.4) and (M'), the density p_θ will only depend on $\lambda^T m(\mathbf{X}_n,\theta)$. □

Proof of Proposition 4.5

Proof. The consistent result of GEL (Newey and Smith, 2004) can be directly applied to the estimators in $\mathcal{W}$-class. □

Proof of Theorem 4.7

Proof. Completeness of $\mathcal{S}^{n-1}$ (or $\mathcal{H}$) states that every subsequence in $\mathcal{P}_\mu$ equipped with $\rho^{(H)}$ has a limiting point. The consistency result shows that any estimated $\{p_i^*\}_{i\leq n}$ from $\mathcal{W}$-class will converge to $1/n$. All metrics or divergences $\rho(\cdot,\cdot)$ in $\mathcal{W}$-class have the same leading-term as $\rho^{(H)}$. If we consider all the subsequences $\{p_i^*\}_{i\leq n}$ from $\mathcal{W}$-class with the limit point $1/n$

in $\mathcal{S}^{n-1}$ equipped with $\rho^{(H)}$, the subsequences cover all members of p^*s in $\mathcal{W}$-class. In other words, because of completeness and $\mathcal{W}$-class, two sequences of $(p_1^*, \dots p_n^*)$ can be studied in a unified framework even if they are generated by different criterion functions in the problem (M).

We only need to study the convergence result of Hellinger affinity:

$$\begin{aligned}\cos\beta &= \sum_{i=1}^{n} \sqrt{p_i^*} \cdot \sqrt{\frac{1}{n}} = \sum_{i=1}^{n} \frac{1}{n} \left[1 + n\left(p_i^* - \frac{1}{n}\right)\right]^{\frac{1}{2}}, \\ &= \sum_{i=1}^{n} \frac{1}{n} \left[1 + \frac{1}{2} n \left(p_i^* - \frac{1}{n}\right) - \frac{1}{8} n^2 \left(p_i^* - \frac{1}{n}\right)^2 \right. \\ &\quad \left. + O_p\left(|np_i^* - 1|^3\right)\right],\end{aligned}$$

where remainder term will be dropped out asymptotically. The second line is the binomial series expansion because $|n(p_i^* - \frac{1}{n})| \leq 1$ for all i. By consistency, $\lim_{n\to\infty}(p_i^* - \frac{1}{n}) = 0$ for all i. The expression becomes

$$\cos\beta = 1 - \frac{1}{8} \sum_{i=1}^{n} n \left(p_i^* - \frac{1}{n}\right)^2 + o_p(1)$$

the second term is nothing but Neyman's χ^2, also known as Euclidean log-likelihood[9]. Thus, by definition

$$2\sin^2 \frac{\beta}{2} = \frac{1}{8} \sum_{i=1}^{n} n \left(p_i^* - \frac{1}{n}\right)^2 \sim \frac{1}{8} \chi_d^2.$$

This weak convergency result will hold for $p_{\tilde{\theta}}$ of any member in the $\mathcal{W}$-class. □

9 Euclidean log-likelihood belongs to the so called Cressie-Read family, $\mathrm{CR}(k) = [2/(k^2 + k)] \sum_i^n [n(n/p_i)^k - 1]$, for $k = -2$. Please refer to Owen (Chapter 3.15 and 3.16 2001) for more discussion.