PRIOR INFORMATION IN STOCHASTIC OPTIMIZATION: QUASIGRADIENT METHODS

Francisco Venegas-Martínez*
Centro de Investigación en Finanzas, Tecnológico de Monterrey, Campus Ciudad de México

Gilberto Pérez-Lechuga
Centro de Investigación Avanzada en Ingeniería Industrial, Universidad Autónoma del Estado de Hidalgo

(Received 2 December 2002, accepted 3 March 2003)

Abstract

In this paper, we extend the stochastic quasigradient method when there is prior information on the region where descent directions are likely to be found. Our extension uses maximum entropy and minimum cross-entropy subgradient estimators that incorporate prior information in the form of expectations. We also analyze a number of prior information patterns and provide the convergence conditions for the proposed method. Finally, we obtain a limiting distribution representation for the expected information, which is provided by the sequence of subgradient estimators generated by the proposed method.

Resumen

En este trabajo, se extiende el método de cuasi-gradiente estocástico cuando hay información *a priori* sobre la región en donde es probable encontrar direcciones descendentes. Nuestra extensión utiliza los estimadores de subgradiente de máxima entropía y de mínima entropía cruzada que incorporan la información *a priori* en la forma de valores esperados. Asimismo, analizamos varios patrones información *a priori* y proporcionamos las condiciones de la convergencia para el método propuesto. Por último, obtenemos una representación de la distribución límite para la información esperada, la cual es proporcionada por una sucesión de estimadores de los subgradientes generados por el método propuesto.

Clasificación JEL: C61 C11
Keywords: Stochastic quasigradient methods, Information theory

* Centro de Invesgación en Finanzas, Tecnológico de Monterrey, Campus Ciudad de México, Calle del Puente 222, Aulas 3, Cuarto piso, Col. Ejidos de Huipulco, Del. Tlalpan, 14380 México, D. F., Teléfono: +52(55)54832254, Correo electrónico: fvenegas@itesm.mx

The authors are grateful to the anonymous referees for many comments. Some very useful suggestions were provided by Sam Saunders

1. Introduction

There is a wide range of mathematical programming problems that can be neither solved nor analyzed by using deterministic optimization techniques. For instance, when the objective function and/or the constraints are not differentiable, then stochastic algorithms that use statistical estimators of the subgradients can be more flexible and effective. Most of the deterministic optimization algorithms are myopic in the sense that they are unable to escape from local optima, however, when stochastic subgradients are able to incorporate prior information, it is possible to explore other regions where descent directions can be found with a positive probability. In the present work, we extend the stochastic quasigradient method to more general prior information patterns. We also provide convergence conditions for this extended method.

There are in the literature many stochastic optimization methods that use a limiting distribution approach; we mention, for instance, Dorea (1991) and (1987), de Haan (1981) and Galambos (1978). In this paper, we obtain a limiting distribution representation for the expected information provided by the sequence of subgradient estimators generated by the proposed algorithm.

This paper is organized as follows. In section 2, we define a statistical estimator of a subgradient. Section 3, is an outline of the principles of maximum entropy and minimum cross entropy. In section 4, we construct both the maximum entropy and minimum cross-entropy statistical estimator for subgradients that incorporate prior information in terms of expectations. Through section 5, we extend the stochastic quasigradient method to incorporate prior information. We also provide convergence conditions for this extension. In section 6, by using information theory, we obtain a limiting distribution representation of information provided by the sequence of subgradient estimators generated by the algorithm. Finally, in the last section, we comment on the advantages and delimitations of the obtained results.

2. Subgradient Estimators

For a systematic investigation, we first recall the concept of subgradient of a convex function which is related to the epigraph and supporting hyperplane in optimization theory. We also define a statistical estimator of a subgradient when prior information in terms of expectations is available.

Let $g(\mathbf{X})$, $\mathbf{X} \in \mathbb{R}^m$, be a convex function, not necessarily differentiable, a vector $\widehat{\nabla} g(\mathbf{X})$ is called a subgradient of g at $\mathbf{X}$ if the inequality

$$g(\mathbf{Y}) - g(\mathbf{X}) \geq \langle \widehat{\nabla} g(\mathbf{X}), \mathbf{Y} - \mathbf{X} \rangle \tag{2.1}$$

holds for all $\mathbf{Y} \in \mathbb{R}^m$. Here, $\langle \cdot, \cdot \rangle$ is the usual inner product in the Euclidean space $\mathbb{R}^m$.

One way to construct a statistical estimator, $\xi(\mathbf{X})$, of the subgradient of g at $\mathbf{X}$ is as follows: Consider a random vector $\theta^T = (\theta_1, \theta_2, \ldots, \theta_m)$ with independent, but not necessarily identically distributed components (a superindex T will denote the transposing operation).Let

$$\theta_i^T = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots \theta_m^{(i)}), \quad i = 1, 2, ..., \rho,$$

denote a random sample of size ρ from θ. An estimator, $\xi(\mathbf{X})$, of a subgradient of g at $\mathbf{X}$ may be taken as

$$\xi(\mathbf{X}) = \frac{1}{\rho}\sum_{i=1}^{\rho}\frac{g(\mathbf{X}+\Delta\theta_i)-g(\mathbf{X})}{\Delta}\theta_i, \tag{2.2}$$

where $\Delta > 0$ is a scalar. In this average the factor $[g(\mathbf{X}+\Delta\theta_i)-g(\mathbf{X})]/\Delta$ estimates the rate of change of g with respect to $\mathbf{X}$, and the factor θ_i is a random vector with prior information expressed through

$$\mathcal{I}(\theta): \quad \int_{\Theta} a_k(\theta)\pi(\theta)d\theta = \bar{a}_k, \quad k = 1, 2, ..., s, \quad \Theta \subset \mathbb{R}^m, \tag{2.3}$$

where the functions a_k and tne constants $\bar{a}_k$'s are known. In the sequel, we shall use the more convenient notation for (2.3)

$$\mathcal{I}(\theta) = \{\Theta;\ a_1(\theta), a_2(\theta), ..., a_s(\theta);\ \bar{a}_1, \bar{a}_2, ..., \bar{a}_s\}.$$

3. Maximum and Minimum Cross Entropy

The principle of maximum entropy (Jaynes (1957)) provides a general method of inference about an unknown density, $\pi(\theta)$, when there is new information about $\pi(\theta)$ in terms of expectations. The principle states that of all compatible densities with the new information, we should choose as estimate for $\pi(\theta)$, the one with the greatest entropy. The principle of minimum cross entropy (Kullback (1956)) considers, besides new information in terms of expectations, an initial estimate $p(\theta)$ of $\pi(\theta)$, and in this case, we should choose as final estimate for $\pi(\theta)$, the one with the least cross entropy. Shore and Johnson ((1980) and (1981)) have provided an axiomatic derivation, of these principles through an abstract information operator.

The maximum entropy principle is equivalent to the minimum cross entropy in the special case of discrete spaces and uniform initial estimates. According to Jaynes (1957), to find an posterior estimator of an unknown density function, $\pi(\theta)$, $\theta \in \Theta$, when there is prior information $\mathcal{I}(\theta)$, the maximum entropy principle leads us to solve the following variational problem:

$$\text{Maximize} \quad H(\pi) = -\int_{\Theta}\pi(\theta)\log\pi(\theta)\mathrm{d}\theta,$$

$$\text{subject to} \quad \mathcal{I}(\theta) = \{\Theta;\ 1, a_1(\theta), a_2(\theta), ..., a_s(\theta);\ 1, \bar{a}_1, \bar{a}_2, ..., \bar{a}_s\}.$$

A necessary condition for an estimator $\pi^*(\theta)$ to be a maximum is that

$$\begin{cases} \pi^*(\theta) = \exp\left\{\lambda_0 + \sum_{k=1}^{s}\lambda_k a_k(\theta)\right\}, \\ 1 - \int_{\Theta}\pi^*(\theta)\mathrm{d}\theta = 0, \\ \int_{\Theta}[\bar{a}_k - a_k(\theta)]\pi^*(\theta)\mathrm{d}\theta = 0, \quad k = 1, 2, ..., s, \end{cases} \tag{3.1}$$

where $\lambda_0, \lambda_1, ..., \lambda_s$ are the Lagrange multipliers associated to the given constraints. Substituting $\pi^*(\theta)$ in the other two conditions of (3.1), we readily find that

$$\begin{cases} 0 = \lambda_0 - \log\left\{\int_\Theta \prod_{k=1}^s e^{\lambda_k a_k(\theta)} \mathrm{d}\theta\right\}, \\ 0 = \int_\Theta [a_k(\theta) - \bar{a}_k] \prod_{k=1}^s e^{\lambda_k a_k(\theta)} \mathrm{d}\theta, \ k = 1, 2, ..., s, \end{cases} \tag{3.2}$$

which is a homogeneous nonlinear system in the variables $\lambda_0, \lambda_1, ..., \lambda_s$. Moreover, if the integral determining λ_0 in the above system can be written in a closed form, then the rest of the multipliers can be found from the following relations:

$$\frac{\partial \lambda_0}{\partial \lambda_k} = -a_k, \qquad k = 1, 2, ..., s. \tag{3.3}$$

Following Kullback (1956), to find a posterior estimator of a density function $\pi(\theta)$ $\theta \in \theta$, when we have a prior estimator, $p(\theta)$, and prior information, $\mathcal{I}(\theta)$, we need to solve the following variational problem of minimum cross entropy:

$$\text{Minimize} \quad H(\pi, p) = \int_\Theta \pi(\theta) \log \frac{\pi(\theta)}{p(\theta)} \mathrm{d}\theta,$$

$$\text{subject to} \quad \mathcal{I}(\theta) = \{\Theta;\ 1, a_1(\theta), a_2(\theta), ..., a_s(\theta);\ 1, \bar{a}_1, \bar{a}_2, ..., \bar{a}_s\}.$$

A necessary condition for an estimator $\pi^*(\theta)$ to be minimum is that

$$\begin{cases} \pi^*(\theta) = p(\theta) \exp\left\{ -\lambda_0 - \sum_{k=1}^s \lambda_k a_k(\theta) \right\}, \\ 1 - \int_\Theta \pi^*(\theta) \mathrm{d}\theta = 0, \\ \int_\Theta (\bar{a}_k - a_k(\theta)) \pi^*(\theta) \mathrm{d}\theta = 0, \quad k = 1, 2, ..., s, \end{cases} \tag{3.4}$$

Proceeding as in the maximum entropy case, we find

$$\begin{cases} 0 = \lambda_0 - \log\left\{\int_\Theta p(\theta) \prod_{k=1}^s e^{-\lambda_k a_k(\theta)} \mathrm{d}\theta\right\}, \\ 0 = \int_\Theta [a_k(\theta) - \bar{a}_k] p(\theta) \prod_{k=1}^s e^{-\lambda_k a_k(\theta)} \mathrm{d}\theta, \ k = 1, 2, ..., s. \end{cases} \tag{3.5}$$

Similar relations to those of (3.3) also hold here.

4. Maximum Entropy and Minimum Cross Entropy Subgradient Estimators

In this section we construct maximum entropy and minimum cross-entropy subgradient estimators. We also study a number of prior information patterns,

some of which have been analyzed in Venegas ((1990a) and (1990b)). We start with the following theorem:

Theorem 4.1 *Let g be a convex function, not necessarily differentiable at $\mathbf{X}$, defined on the whole space $\mathbb{R}^m$. If*

$$\xi(\mathbf{X}) = \frac{1}{\rho}\sum_{i=1}^{\rho}\frac{g(\mathbf{X}+\Delta\theta_\ell)-g(\mathbf{X})}{\Delta}\theta_i,$$

as in (2.2), then there exists a symmetric and positive definite (prior information) matrix $\mathbf{G}(\mathcal{I})$ of order $m \times m$, and a vector $\mathbf{g}(\mathbf{X},\mathcal{I}) \in \mathbb{R}^m$, $\mathbf{g}(\mathbf{X},\mathcal{I}) \geq 0$, such that

$$\mathrm{E}\{\xi(\mathbf{X}) \mid \mathbf{X},\ \mathcal{I}\} = \mathbf{G}(\mathcal{I})\widehat{\nabla}g(\mathbf{X}) + \mathbf{g}(\mathbf{X},\mathcal{I}).$$

The matrix $\mathbf{G}(\mathcal{I})$ has entries

$$G_{\ell j}(\mathcal{I}) = \begin{cases} \displaystyle\int_\Theta \theta_\ell^2 \exp\left\{\lambda_0 + \sum_{k=1}^{m}\lambda_k a_k(\theta)\right\}\mathrm{d}\theta, & \ell = j, \\ \displaystyle\left[\int_\Theta \theta_\ell\theta_j \exp\left\{\lambda_0 + \sum_{k=1}^{m}\lambda_k a_k(\theta)\right\}\mathrm{d}\theta\right], & \ell \neq j, \end{cases} \tag{4.1}$$

when there is prior information $\mathcal{I}(\theta)$. Moreover, when there is a prior estimator $p(\theta)$ as well as prior information $\mathcal{I}(\theta)$, $\mathbf{G}(\mathcal{I})$ takes the form

$$G_{\ell j}(\mathcal{I}) = \begin{cases} \displaystyle\int_\Theta \theta_\ell^2 \exp\left\{-\lambda_0 - \sum_{k=1}^{m}\lambda_k a_k(\theta)\right\}\mathrm{d}\theta, & \ell = j, \\ \displaystyle\int_\Theta \theta_\ell\theta_j \exp\left\{-\lambda_0 - \sum_{k=1}^{m}\lambda_k a_k(\theta)\right\}\mathrm{d}\theta, & \ell \neq j. \end{cases} \tag{4.2}$$

Proof: By virtue of (2.1), for each $i = 1, 2..., \rho$, we have

$$\frac{g(\mathbf{X}+\Delta\theta_i)-g(\mathbf{X})}{\Delta}\theta_i \geq \langle\widehat{\nabla}g(\mathbf{X}),\theta_i\rangle\theta_i,$$

hence

$$\mathrm{E}\left\{\frac{g(\mathbf{X}+\Delta\theta_i)-g(\mathbf{X})}{\Delta}\theta_i \mid \mathbf{X},\ \mathcal{I}\right\} \geq \mathrm{E}\left\{\langle\widehat{\nabla}g(\mathbf{X}),\theta_i\rangle\,\theta_i \mid \mathbf{X},\ \mathcal{I}\right\}.$$

The product $\langle\widehat{\nabla}g(\mathbf{X}),\theta_i\rangle\,\theta_i$ may be written as $\mathbf{H}(\theta_i)\widehat{\nabla}g(\mathbf{X})$, where the matrix $\mathbf{H}(\theta_i)$ has entries $H_\ell(\theta_i) = \theta_\ell^{(i)}\theta_j^{(i)}$, $\ell, j = 1, 2, ..., m$. Thus,

$$\mathrm{E}\left\{\frac{g(\mathbf{X}+\Delta\theta_i)-g(\mathbf{X})}{\Delta}\theta_i \mid \mathbf{X},\ \mathcal{I}\right\} \geq \mathrm{E}\{\mathbf{H}(\theta_i) \mid \mathcal{I}\}\,\widehat{\nabla}g(\mathbf{X}).$$

Notice now that the matrix $\mathrm{E}\{\mathbf{H}(\theta_i) \mid \mathcal{I}\}$ is independent of i, so we may write

$$\mathrm{E}\{\mathbf{H}(\theta_i) \mid \mathcal{I}\} = \mathbf{G}(\mathcal{I})$$

for all $i = 1, 2, ..., \rho$. From (3.1) and (3.4) we obtain (4.1) and (4.2), respectively. The conclusion follows then by summing on i and dividing by ρ. □

The previous theorem allows us to write

$$\mathrm{E}\{\mathbf{G}^{-1}(\mathcal{I})\xi(\mathbf{X}) \mid \mathbf{X},\ \mathcal{I}\} = \widehat{\nabla} g(\mathbf{X}) + \mathbf{h}(\mathbf{X}, \mathcal{I}), \tag{4.3}$$

where $\mathbf{h}(\mathbf{X}, \mathcal{I}) = \mathbf{G}^{-1}(\mathcal{I})\mathbf{g}(\mathbf{X}, \mathcal{I})$ is the conditional bias of $\xi(\mathbf{X})$, we shall visit this result in the next section. We now work on a number of applications of theorem 4.1.

Example 4.1 (Applications of Theorem 4.1)

In the following cases there is not a prior estimator:

(*i*).Suppose that prior information about descent directions are likely to be found is around a point $\mu = (\mu_1, \mu_2, ..., \mu_m)$ within a standard deviation σ_ℓ for each component $\ell = 1, 2, ..., m$. That is,

$$\mathcal{I}(\theta) = \left\{\mathbb{R}^m;\ \theta, (\theta - \mu)(\theta - \mu)^T;\ \mu, \Sigma\right\},$$

where $\Sigma = [\sigma_\ell^2]_{\ell=1}^m$ is a diagonal matrix. In such a case, we have $G_{\ell j}(\mathcal{I}) = \sigma_\ell^2 + \mu_\ell^2$, $\ell = j$, and $G_{\ell j}(\mathcal{I}) = \mu_\ell^2$, $\ell \neq j$.

(*ii*). Assuming that prior information is given by

$$\mathcal{I}(\theta) = \Big\{\mathbb{R}_+^m;\ \theta, \log\theta_1, ..., \log\theta_m;\ \alpha_1\beta_1^{-1}, ..., \alpha_m\beta_m^{-1},$$
$$\psi(\alpha_1) - \log\beta_1, ..., \psi(\alpha_m) - \log\beta_m, \Big\},$$

where $\mathbb{R}_+^m = \{\mathbf{X} \in \mathbb{R}^m | \mathbf{X} > 0\}$, $\alpha_\ell, \beta_\ell > 0$, and ψ is the digamma function (see, for instance, Gradshteyn and Ryzhik, 1980). In this case, we find that

$$G_{\ell j}(\mathcal{I}) = \begin{cases} \dfrac{\alpha_\ell(1+\alpha_\ell)}{\beta_\ell^2}, & \ell = j, \\ \dfrac{\alpha_\ell\alpha_j}{\beta_\ell\beta_j}, & \ell \neq j. \end{cases}$$

(*iii*) If prior information has the form

$$\mathcal{I}(\theta) = \Big\{\mathbb{R}_+^m;\ \log\theta_1, ..., \log\theta_m, \theta_1^{\delta_1}, ..., \theta_m^{\delta_m};$$
$$\delta_1^{-1}[\psi(\alpha_1) - \log\beta_1], ..., \delta_m^{-1}[\psi(\alpha_m) - \log\beta_m], \beta_1^{-1}, ..., \beta_m^{-1}\Big\}$$

where $\alpha_\ell, \beta_\ell, \delta_\ell > 0$, then $\mathbf{G}(\mathcal{I})$ satisfies

$$G_{\ell j}(\mathcal{I}) = \begin{cases} \dfrac{\Gamma(\alpha_\ell\beta_\ell + \delta_\ell^{-1})}{\Gamma(\alpha_\ell)} (\alpha_\ell\beta_\ell)^{\alpha_\ell - \alpha_\ell\beta_\ell - 2\delta_\ell^{-1}}, & \ell = j, \\ \dfrac{\Gamma(\alpha_\ell\beta_\ell + \delta_\ell^{-1})\Gamma(\alpha_j\beta_j + \delta_j^{-1})}{\Gamma(\alpha_\ell)\Gamma(\alpha_j)} \times \\ \times\ (\alpha_\ell\beta_\ell)^{\alpha_\ell - \alpha_\ell\beta_\ell - \delta_\ell^{-1}} (\alpha_j\beta_j)^{\alpha_j - \alpha_j\beta_j - \delta_j^{-1}}, & \ell \neq j. \end{cases}$$

Example 4.2 (Further applications of Theorem 4.1)
Suppose that prior information is that θ_ℓ is in some region $\Theta_\ell = (b_{1\ell}, b_{s_\ell+1})$. Suppose also that we assign weights, $\gamma_{1\ell}, \gamma_{2\ell}, ..., \gamma_{s_\ell} \geq 0$, $(\sum_{k=1}^{s_\ell} \gamma_{k\ell} = 1)$, to the events that θ_ℓ belongs to the subregions $A_{k\ell} = (b_{k\ell}, b_{k\ell+1}]$, $k = 1, 2, ..., s_\ell - 1$ and $A_{s_\ell} = (b_{s_\ell,\ell}, b_{s_\ell+1})$, with $b_{1\ell} < b_{2\ell} < \cdots < b_{s_\ell+1}$, $s_\ell \geq 2$, which constitute a partition of $\Theta_\ell = (b_{1\ell}, b_{s_\ell+1})$. Thus, prior information can be written as

$$\int_{\Theta_\ell} I_{A_{k\ell}}(\theta_\ell)\pi(\theta_\ell)\mathrm{d}\theta_\ell = \gamma_{k\ell} > 0, \qquad k = 1, 2, ..., s_\ell, \qquad \sum_{k=1}^{s_\ell} \gamma_{k\ell} = 1. \tag{4.4}$$

If we use the maximum entropy principle, the necessary conditions given in (3.2) are transformed into

$$\begin{cases} 0 = \lambda_{0\ell} - \log\left\{ \displaystyle\int_{\Theta_\ell} \prod_{k=1}^{s_\ell} e^{-\lambda_{k\ell} I_{A_{k\ell}}(\theta_\ell)} \mathrm{d}\theta_\ell \right\}, \\ 0 = \lambda_{0\ell} + \lambda_{k\ell} - \log\left\{ \gamma_{k\ell}^{-1} \displaystyle\int_{\Theta_\ell} I_{A_{k\ell}}(\theta_\ell)\mathrm{d}\theta_\ell \right\}, \quad k = 1, 2, ..., s_\ell. \end{cases} \tag{4.5}$$

If we now define the change of variable

$$\begin{cases} \omega_{0\ell} = e^{\lambda_{0\ell}}, \\ \omega_{k\ell} = e^{-\lambda_{k\ell}}, \quad k = 1, 2, ..., s_\ell, \end{cases}$$

then (4.5) is transformed into the homogeneous linear system:

$$\begin{pmatrix} -1 & u_{1\ell} & u_{2\ell} & \dots & u_{s_\ell} \\ -1 & v_{1\ell} & 0 & \dots & 0 \\ -1 & 0 & v_{2\ell} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \dots & v_{s_\ell} \end{pmatrix} \begin{pmatrix} \omega_{0\ell} \\ \omega_{1\ell} \\ \omega_{2\ell} \\ \vdots \\ \omega_{s_\ell} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{4.6}$$

where $u_{k\ell} = \int_{\Theta_\ell} I_{A_{k\ell}}(\theta_\ell)\mathrm{d}\theta_\ell$, and $v_{k\ell} = \gamma_{k\ell}^{-1} u_{k\ell}$, $k = 1, 2, ..., s_\ell$. We denote the matrix in (4.6) by $\mathbf{M}_\ell$. We shall see that this matrix also plays a role

in cross-entropy minimization (*cf.* Venegas, 1992). The determinant of $\mathbf{M}_\ell$ is given by

$$\left(\frac{\sum_{k=1}^{s_\ell}\gamma_{k\ell}-1}{\prod_{k=1}^{s_\ell}\gamma_{k\ell}}\right)\prod_{k=1}^{s_\ell}u_{k\ell};$$

notice that $\sum_{k=1}^{s_\ell}\gamma_{k\ell}=1$ guarantees the existence of a nontrivial solution. After solving the homogeneous linear system (4.6) we find

$$G_{\ell j}(\mathcal{I})=\begin{cases}\frac{1}{3}\displaystyle\sum_{k=1}^{s_\ell}\gamma_{k\ell}(b_{k\ell}^2+b_{k\ell}b_{k\ell+1}+b_{k\ell+1}^2), & \ell=j,\\ \frac{1}{4}\left[\displaystyle\sum_{k=1}^{s_\ell}\gamma_{k\ell}(b_{k\ell}+b_{k\ell+1})\right]\left[\displaystyle\sum_{k=1}^{s_j}\gamma_{kj}(b_{kj}+b_{kj+1})\right], & \ell\neq j.\end{cases}$$

If we now assume that there is a prior estimator, say,

$$p(\theta_\ell)=\sum_{k=1}^{s_\ell}\beta_{k\ell}u_{k\ell}^{-1}I_{A_{k\ell}}(\theta_\ell),\quad \theta_\ell\in\Theta_\ell,$$

where $\beta_{k\ell}>0$, $k=1,2,...,s$, $\sum_{k=1}^{s_\ell}\beta_{k\ell}=1$, and with prior information as expressed in (4.4), that is, prior information expresses changes in the weights. Then, by using the same change of variable as before in (4.6), we obtain a homogeneous linear system in terms of $\mathbf{M}_\ell$, namely,

$$\mathbf{M}_\ell\mathrm{diag}(1,\beta_{1\ell}u_{1\ell}^{-1},...,\beta_{s_\ell,\ell}u_{s_\ell,\ell}^{-1})\Omega=\mathbf{0}, \tag{4.7}$$

where $\Omega=(\omega_{0\ell},\omega_{1\ell},...,\omega_{s\ell})^T$ and $\mathbf{0}$ is the zero vector. Here, the information provided by the initial estimate is incorporated through the diagonal matrix in (4.7). In this case, the solution is

$$\Omega^*=(1,\beta_{1\ell}^{-1}\gamma_{1\ell},...,\beta_{s_\ell,\ell}^{-1}\gamma_{s_\ell,\ell})^T,$$

in which case the prior estimator has no effect on the posterior one. Therefore, the prior information matrix $\mathbf{G}(\mathcal{I})$ remains unchanged.

Example 4.3 (Redundant Prior Information)
Suppose that we have initial information on θ in the form

$$\mathcal{I}(\theta)=\{\mathbb{R}_+^m;\ \theta;\ \beta_1^{-1},...\beta_m^{-1}\},\quad \beta_\ell>0. \tag{4.8}$$

In this case

$$G_{\ell j}(\mathcal{I})=\begin{cases}\dfrac{2}{\beta_\ell^2}, & \ell=j,\\ \dfrac{1}{\beta_\ell\beta_j}, & \ell\neq j.\end{cases}$$

If we include additional prior information in (4.8) such as follows:

$$\mathcal{I}(\theta) = \left\{ \mathbb{R}_+^m;\ \theta, \log\theta_1, ..., \log\theta_m;\ \beta_1^{-1}, ..., \beta_m^{-1}; -\mathcal{K} - \log\beta_1, ..., -\mathcal{K} - \log\beta_m, \right\}, \tag{4.9}$$

where $\beta_\ell > 0$ and $\mathcal{K}(\approx 0.5772)$ is the Euler constant, then we have no change in $\mathbf{G}(\mathcal{I})$, so the added information is redundant.

Of course, additional prior information is not always redundant. Suppose now that we include, instead, additional prior information in (4.8) as follows:

$$\mathcal{I}(\theta) = \left\{ \mathbb{R}_+^m;\ \theta, \log\theta_1, ..., \log\theta_m;\ \beta_1^{-1} ..., \beta_m^{-1},\right.$$
$$\left. \psi(\alpha_1) - \log\alpha_1\beta_1, ..., \psi(\alpha_m) - \log\alpha_m\beta_m \right\},$$

where $\alpha_\ell, \beta_\ell > 0$, then the prior information matrix have the form

$$G_\ell(\mathcal{I}) = \begin{cases} \dfrac{1}{\beta_\ell^2}\left(1 + \dfrac{1}{\alpha_\ell}\right), \ell = j, \\ \dfrac{1}{\beta_\ell\beta_j}, \ell \neq j. \end{cases}$$

Example 4.4 (How sensitive is the prior information matrix?)
Suppose that we have initial information on θ and it is given by

$$\mathcal{I}(\theta) = \{\mathbb{R}_+^m;\ \theta;\ \beta_1^{-\frac{1}{\alpha_1}}\Gamma(1+\alpha_1^{-1}), ..., \beta_m^{-\frac{1}{\alpha_m}}\Gamma(1+\alpha_m^{-1})\}, \quad \alpha_\ell, \beta_\ell > 0. \tag{4.10}$$

Then the maximum entropy solution for $\pi^*(\theta_\ell)$ is

$$\pi^*(\theta_\ell) = \{\beta_\ell^{-\frac{1}{\alpha_\ell}}\Gamma(1+\alpha_\ell^{-1})\}^{-1} \exp\left\{-\left[\beta_\ell^{-\frac{1}{\alpha_\ell}}\Gamma(1+\alpha_\ell^{-1})\right]^{-1}\theta_\ell\right\}, \quad \theta_\ell > 0,$$

and matrix $\mathbf{G}(\mathcal{I})$ can be obtained by proceeding as in (4.9). If we include additional prior information on θ in (4.10), so that

$$\mathcal{I}(\theta) = \Big\{ \mathbb{R}_+^m;\ \theta, \log\theta_1, ..., \log\theta_m;\ \beta_1^{-\frac{1}{\alpha_1}}\Gamma(1+\alpha_1^{-1}), ...\beta_m^{-\frac{1}{\alpha_m}}\Gamma(1+\alpha_m^{-1}),$$
$$\psi[\Gamma(1+\alpha_1^{-1})] - \alpha_1^{-1}\log\beta_1, ..., \psi[\Gamma(1+\alpha_m^{-1})] - \alpha_m^{-1}\log\beta_m \Big\} \tag{4.11}$$

where $\alpha_\ell, \beta_\ell > 0$. Then the solution of the maximum entropy variational problem, $\pi^*(\theta_\ell)$, is the Gamma distribution with parameters $\beta_\ell^{\frac{1}{\alpha_\ell}}$, and $\Gamma(1+\alpha_\ell^{-1})$. The matrix $\mathbf{G}(\mathcal{I})$ can be obtained by proceeding as in (*ii*) in Examples 4.1.

Now, let us see what happens to $\mathbf{G}(\mathcal{I})$,if instead of (4.10) we set

$$\mathcal{I}(\theta) = \{\mathbb{R}_+^m;\ \theta, \log\theta_1, ..., \log\theta_m, \theta^{\alpha_1}, ..., \theta^{\alpha_m};$$
$$\beta^{-\frac{1}{\alpha}}\Gamma(1+\alpha_1^{-1}), ..., \beta^{-\frac{1}{\alpha}}\Gamma(1+\alpha_m^{-1}),$$
$$-\alpha_1^{-1}(\mathcal{K} - \log\beta_1), ..., -\alpha_m^{-1}(\mathcal{K} - \log\beta_m), \beta_1^{-1}, ..., \beta_m^{-1}\},\ \alpha_\ell, \beta_\ell > 0.$$

Then the maximum entropy density for θ_ℓ is the Weibull distribution with parameters α_ℓ and β_ℓ,that is

$$\pi^*(\theta_\ell) = \alpha_\ell\beta_\ell\theta_\ell^{\alpha_\ell - 1}\exp\{-\beta_\ell\theta_\ell^{\alpha_\ell}\}, \quad \theta_\ell > 0,$$

and the prior information matrix is

$$G_{\ell j}(\mathcal{I}) = \begin{cases} \left(\frac{1}{\beta_\ell}\right)^{\frac{2}{\alpha_\ell}}\Gamma\left(1+\frac{2}{\alpha_\ell}\right), & \ell = j, \\ \left(\frac{1}{\beta_\ell}\right)^{\frac{1}{\alpha_\ell}}\left(\frac{1}{\beta_m}\right)^{\frac{1}{\alpha_m}}\left[\Gamma(1+\alpha_\ell)\Gamma(1+\alpha_m)\right]^{-1}, & \ell \neq j. \end{cases}$$

This example shows how sensitive is the prior information matrix $\mathbf{G}(\mathcal{I})$ to changes when additional prior information is included.

5. A Stochastic Quasigradient Method with Prior Information

In this section we extend the stochastic quasigradient method to incorporate prior information. We also provide the convergence conditions for the extension. Our goal is to solve the problem:

$$\begin{cases} \text{Minimize} & g(\mathbf{X}), \\ \text{subject to} & \mathbf{X} \in \mathcal{C}, \end{cases} \tag{5.1}$$

where g is a convex, but not necessarily differentiable, function defined on the whole space $\mathbb{R}^m$. To solve (5.1) when there is prior information, we shall generate a random sequence of points $\mathbf{X}_k$, one at each iteration, such that

$$\mathrm{P}_\theta\{\lim_{k\to\infty} g(\mathbf{X}_k) = g(\mathbf{X}^*)\} = 1.$$

where $\mathbf{X}^*$ solves (5.1). At iteration k we define the random vector

$$\mathbf{X}_{k+1} = \mathbf{X}_k - \alpha_k \mathbf{G}(\mathcal{I}_k)^{-1}\xi(\mathbf{X}_k),$$

where $\mathbf{X}_1$ is an arbitrary initial point in $\mathbb{R}^m$.

Example 5.1 (Reconsidering Example 4.1)
(i). Consider (i) in Example 4.1, and suppose $\Sigma_k = [\sigma^2_{\ell k}]_{\ell=1}^m$ where $\sigma_{\ell k} \to \sigma_\ell$, as $k \to \infty$, for every ℓ. Then,

$$\mathbf{G}(\mathcal{I}_k)^{-1} \to \mathbf{G}(\mathcal{I})^{-1} \tag{5.2}$$

as $k \to \infty$.

(ii). In (ii) in Examples 4.1, consider sequences of positive numbers $\{\alpha_{\ell k}\}_{k=1}^\infty$ and $\{\beta_{\ell k}\}_{k=1}^\infty$ such that $\alpha_{\ell k} \to \alpha_\ell$ and $\beta_{\ell k} \to \beta_\ell$, as $k \to \infty$, for each $\ell = 1, ..., m$. Then (5.2) holds.

(iii). In Example 4.2, define $\Theta_{\ell k} = (b_{1\ell}^{(k)}, b_{s_\ell+1}^{(k)})$, where $b_{1\ell}^{(k)} \to b_{1\ell}$, and $b_{s_\ell+1}^{(k)} \to b_{s_\ell+1}$, as $k \to \infty$, for every $\ell = 1, ..., m$. Then (5.2) is satisfied.

In all of the above examples we have that the sequence $\{\|\mathbf{G}(\mathcal{I}_k)^{-1}\|\}_{k=1}^\infty$ is bounded. It is worthwhile to point out that, in general, from one iteration to the next we may find that g does not diminish, so further assumptions on the step parameter α_k have to be made. We shall be interested in the special case when $\alpha_k = k^{-1}$, $k = 1, 2, ...$

The following convergence conditions, although somehow strict, are easy to follow making the proof reasonably straightforward.

Theorem 5.1 *Consider the random sequence*

$$\mathbf{X}_{k+1} = \mathbf{X}_k - k^{-1}\mathbf{G}(\mathcal{I}_k)^{-1}\xi(\mathbf{X}_k), \quad k = 1, 2, ...$$

and suppose that:

(i). *There exists a constant A such that*

$$\|\mathbf{G}(\mathcal{I}_k)^{-1}\| \leq A, \quad \text{for all } k = 1, 2, ...$$

(ii). *$\mathcal{C}$ is a convex compact set in $\mathbb{R}^m$.*

(iii). *There exists a constant B such that*

$$\mathrm{E}\{\ \|\xi(\mathbf{X}_k)\|^2 \ \mid\ \mathbf{X}_k,\ \mathcal{I}_k\ \} \leq B, \quad \text{for all } k = 1, 2, ...$$

(iv). *There exists a constant C such that*

$$\|\mathbf{h}(\mathbf{X}_k, \mathcal{I}_k)\| \leq C, \quad \text{for all } k = 1, 2, ...$$

Then

$$\mathrm{P}_\theta\{\lim_{k\to\infty} g(\mathbf{X}_k) = g(\mathbf{X}^*)\} = 1,$$

where $\mathbf{X}^*$ *solves* (5.1).

Proof: Let $\mathbf{X}^*$ be an arbitrary solution of problem (5.1), then

$$\begin{aligned}\| \mathbf{X}^* - \mathbf{X}_{k+1} \|^2 &\leq \| \mathbf{X}^* - \mathbf{X}_k + k^{-1}\mathbf{G}(\mathcal{I}_k)^{-1}\xi(\mathbf{X}_k) \|^2 \\ &= \| \mathbf{X}^* - \mathbf{X}_k \|^2 + 2k^{-1}\langle \mathbf{G}(\mathcal{I}_k)^{-1}\xi(\mathbf{X}_k), \mathbf{X}^* - \mathbf{X}_k\rangle \\ &\quad + k^{-2} \| \mathbf{G}(\mathcal{I}_k)^{-1} \|^2 \| \xi(\mathbf{X}_k) \|^2 .\end{aligned}$$

Taking the conditional expectations at both sides of the above inequality we get

$$\begin{aligned}\mathrm{E}\{\| \mathbf{X}^* - \mathbf{X}_{k+1} \|^2 | \mathbf{X}_k, \mathcal{I}_k\} \leq &\| \mathbf{X}^* - \mathbf{X}_k \|^2 + 2k^{-1}\langle \widehat{\nabla} g(\mathbf{X}_k), \mathbf{X}^* - \mathbf{X}_k\rangle \\ &+ 2k^{-1}\langle \mathbf{h}(\mathbf{X}_k, \mathcal{I}_k), \mathbf{X}^* - \mathbf{X}_k\rangle \\ &+ k^{-2} \| \mathbf{G}(\mathcal{I}_k)^{-1} \|^2 \mathrm{E}\{\| \xi(\mathbf{X}_k) \|^2 | \mathbf{X}_k, \mathcal{I}_k\}.\end{aligned} \tag{5.3}$$

Notice now that from (2.1)

$$\langle \widehat{\nabla} g(\mathbf{X}^*), \mathbf{X}^* - \mathbf{X}_k\rangle \leq g(\mathbf{X}^*) - g(\mathbf{X}_k) \leq 0. \tag{5.4}$$

Hence, by using in (5.3) the Cauchy-Schwartz inequality, conditions (i), (iii), (iv) and (5.4), we obtain

$$\mathrm{E}\{\| \mathbf{X}^* - \mathbf{X}_{k+1} \|^2 \mid \mathbf{X}_k,\ \mathcal{I}_k\ \} \leq \| \mathbf{X}^* - \mathbf{X}_k \|^2 + 2C \| \mathbf{X}^* - \mathbf{X}_k \| + A^2 B.$$

From the compactness of $\mathcal{C}$, it follows that $\mathrm{E}\{\| \mathbf{X}^* - \mathbf{X}_{k+1} \|^2 | \mathbf{X}_k, \mathcal{I}_k\}$ is uniformly bounded. Furthermore, from (5.3)

$$\begin{aligned}\mathrm{E}\{\| \mathbf{X}^* - \mathbf{X}_{k+1} \|^2 | \mathbf{X}_k, \mathcal{I}_k\} \leq \mathrm{E} \| \mathbf{X}^* - \mathbf{X}_1 \|^2 + A^2 B \sum_{i=1}^{k} i^{-2} \\ + \sum_{i=1}^{k} i^{-1}\mathrm{E}\langle \widehat{\nabla} g(\mathbf{X}_i), \mathbf{X}^* - \mathbf{X}_i\rangle.\end{aligned}$$

Since $\mathrm{E}\{\| \mathbf{X}^* - \mathbf{X}_{k+1} \|^2 | \mathbf{X}_k, \mathcal{I}_k\}$ is uniformly bounded and $\sum_{i=1}^{\infty} i^{-1} = \infty$, we must have that

$$\sum_{i=1}^{\infty} i^{-1}\mathrm{E}\langle \widehat{\nabla} g(\mathbf{X}_i), \mathbf{X}^* - \mathbf{X}_i\rangle > -\infty.$$

Thus, $\mathrm{E}\langle \widehat{\nabla} g(\mathbf{X}_i), \mathbf{X}^* - \mathbf{X}_i\rangle \to 0$ as $i \to \infty$. Hence, there exists a subsequence i_ℓ such that $\langle \widehat{\nabla} g(\mathbf{X}_{i_\ell}), \mathbf{X}_{i_\ell} - \mathbf{X}^*\rangle \to 0$ with probability 1 as $\ell \to \infty$. Therefore, $\mathbf{X}_{i_\ell} \to \mathbf{X}_*$ with probability 1 as $\ell \to \infty$. By the continuity of g, we have that $g(\mathbf{X}_{i_\ell}) \to g(\mathbf{X}^*)$ with probability 1 as $\ell \to \infty$, which completes the proof. □

Corollary 5.1 *Under the hypothesis of theorem* 5.1, *instead of assuming condition* (*iii*), *we now assume that:*

(*iiia*). *There exists a constant* B_1 *such that*

$$\sum_{j=1}^{m} \mathrm{Var}\{\xi^j(\mathbf{X}_k) \mid \mathbf{X}_k,\ \mathcal{I}_k\ \} \leq B_1, \quad \textit{for all } k = 1, 2, ...$$

where $\xi^j(\mathbf{X}_k)$ *is the* j*-th component of* $\xi(\mathbf{X}_k)$.

(*iiib*). *There exists a constant* B_2 *such that*

$$\|\mathrm{E}\{\xi(\mathbf{X}_k) \mid\ \mathbf{X}_k,\ \mathcal{I}_k\ \}\| \leq B_2, \quad \textit{for all } k = 1, 2, ...$$

Then

$$\mathrm{P}_\theta\{\lim_{k\to\infty} g(\mathbf{X}_k) = g(\mathbf{X}^*)\} = 1,$$

where $\mathbf{X}^*$ *solves* (5.1).

Proof: It is enough to notice that

$$\mathrm{E}\{\|\xi(\mathbf{X}_k)\|^2 \mid \mathbf{X}_k, \mathcal{I}_k\ \} = \sum_{j=1}^{m} \mathrm{Var}\{\xi^j(\mathbf{X}_k) \mid \mathbf{X}_k,\ \mathcal{I}_k\ \} + \|\mathrm{E}\{\xi(\mathbf{X}_k) \mid\ \mathbf{X}_k,\ \mathcal{I}_k\ \}\|^2.$$

□

6. Asymptotic Expected Information

In this section, by using information-theoretical arguments, we state a limiting distribution representation for the expected information provided by the sequence of subgradient estimators generated by the proposed algorithm.

Let P_θ denote the distribution of $\xi(\mathbf{X})$. Assume that $\xi(\mathbf{X})$ has a density $f(\xi|\theta)$ (a Radon-Nikodym derivative) with respect to some fixed dominating σ-finite measure λ on $\mathbf{R}$, for all $\theta \in \Theta \subset \mathbf{R}^m$, that is, $\mathrm{d}P_\theta/\mathrm{d}\lambda = f(\xi|\theta)$, for all $\theta \in \Theta \subset \mathbb{R}^m$, so $P_\theta(A) = \int_A f(\xi|\theta)\mathrm{d}\lambda(\xi)$, for all Borel sets $A \in \mathbb{R}$. Accordingly, a random vector $\Psi_k = (\xi_1(\mathbf{X}_k),\ \xi_2(\mathbf{X}_k), ..., \xi_k(\mathbf{X}_k))^T$, with k independent observations from $\xi(\mathbf{X}_k)$,has density,

$$\mathrm{d}\mathcal{P}_\theta/\mathrm{d}\nu = f(\Psi_k|\theta) = \prod_{n=1}^{k} f(\xi_n(\mathbf{X}_k)|\theta),$$

for all $\theta \in \Theta \subseteq \mathbb{R}^m$,where

$$\mathcal{P}_\theta = \underbrace{P_\theta \otimes P_\theta \otimes \cdots \otimes P_\theta}_{k \text{ times}}, \quad \text{and} \quad \nu = \underbrace{\lambda \otimes \lambda \otimes \cdots \otimes \lambda}_{k \text{ times}} \quad \text{on} \quad \mathbb{R}^k.$$

Following Shannon (1948) and Lindley (1956), a measure of the expected information provided by Ψ_k when the prior distribution of θ is $\pi(\theta)$, is defined to be

$$\mathcal{E}(\Psi_k, \pi) = \int f(\Psi_k) \int f(\theta|\Psi_k) \log \frac{f(\theta|\Psi_k)}{\pi(\theta)} \mathrm{d}\mu(\theta)\mathrm{d}\nu(\mathbf{x}), \tag{6.1}$$

where

$$f(\Psi_k) = \int f(\Psi_k|\theta)\pi(\theta)\mathrm{d}\mu(\theta), \qquad f(\theta|\Psi_k) = \frac{f(\Psi_k|\theta)\pi(\theta)}{f(\Psi_k)}. \tag{6.2}$$

In order to obtain an asymptotic representation for (6.1), which includes Fisher's information function and the Gaussian distribution, we need to prove some limit theorems concerning its integrand and justify the passage of the limit under the integral signs. We start by writing (6.1) as:

$$\mathcal{E}(\Psi_k, \pi) = H_\pi(\theta) + \log\sqrt{k} - \int\int \mathcal{L}(\omega) f(\Psi_k|\theta)\pi(\theta)\mathrm{d}\nu(\Psi_k)d\mu(\theta), \tag{6.3}$$

where

$$\mathcal{L}(\omega) = \log\left(\int T_k(\omega)W_k(\omega)\mathrm{d}\mu(\omega)\right),$$

$$H_\pi(\theta) = -\int \pi(\theta)\log\pi(\theta)\mathrm{d}\mu(\theta), \tag{6.4}$$

$$T_k(\omega) = \frac{f(\Psi_k|\theta + \frac{\omega}{\sqrt{k}})}{f(\Psi_k|\theta)}, \tag{6.5}$$

and

$$W_k(\omega) = \frac{\pi(\theta + \frac{\omega}{\sqrt{k}})}{\pi(\theta)}\ . \tag{6.6}$$

Expression (6.4) is known as *Shannon's information measure* about θ when its density is $\pi(\theta)$.

We now state a general limit theorem, which will lead us to the normal convergence of the stochastic process (6.5). We restrict ourselves, for the sake of simplicity, to $m = s = \rho = 1$, that is, $\theta \in \Theta \subseteq \mathbb{R}$ and $\mathbf{X}$, $\xi(\mathbf{X}) \in \mathbb{R}$. Everywhere throughout, both λ and μ will stand for the Lebesgue measure on $\mathbb{R}$. Also, we shall always assume that all densities involved are Lebesgue measurable in both arguments, $\xi(\mathbf{X})$ and θ.

Theorem 6.1 *Assume that $f_{kn}(\xi|\theta)$, $n = 1, 2, ..., k$, $k = 1, 2, ...$, are the corresponding densities of the elements of a triangular array $\xi_{kn} = \xi_n(\mathbf{X}_k)$, $n = 1, 2, ..., k$, $k = 1, 2...$, where each row consists of independent random variables. Assume also that the corresponding distributions $P_{kn,\theta}$ are absolutely continuous with respect to λ for all $\theta \in \Theta$, and that the following set of conditions are satisfied:*

(i). Θ is an open interval in $\mathbb{R}$.

For every k and $n = 1, 2, ..., k$, we have

(ii). $\{x|f_{kn}(\xi|\theta) > 0\}$ is independent of θ.

(iii). If $\theta, \theta' \in \Theta$, then $\theta \neq \theta'$ implies $\lambda\{\xi|f_{kn}(\xi|\theta) \neq f_{kn}(\xi|\theta')\} > 0$.

(iv). For all $\delta > 0$, *and all* $\theta \in \Theta$

$$\max_{1\leq n\leq k} \int_{B_{kn,\delta}(\frac{\omega}{\sqrt{k}})} \left(\sqrt{f_{kn}(\xi|\theta + \tfrac{\omega}{\sqrt{k}})} - \sqrt{f_{kn}(\xi|\theta)} \right)^2 \mathrm{d}\lambda(\xi) = o(\tfrac{1}{k}),$$

where $B_{kn,\delta}(\frac{\omega}{\sqrt{k}}) = \left\{ \xi : \left| \sqrt{f_{kn}(\xi|\theta + \frac{\omega}{\sqrt{k}})} - \sqrt{f_{kn}(\xi|\theta)} \right| \geq \delta \sqrt{f_{kn}(\xi|\theta)} \right\}$.

(v) There exists a random variable X *with density* $f_X(\xi|\theta)$ *such that*

$$\{\xi | f_X(\xi|\theta) > 0\}$$

is independent of θ, $\frac{\partial}{\partial\theta} \log f_X(\xi|\theta)$ *exists for all* $\theta \in \Theta$ *and every* ξ, *and*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{n=1}^{k} \int \left(\frac{\sqrt{f_{kn}(\xi|\theta+\frac{\omega}{\sqrt{k}})} - \sqrt{f_{kn}(\xi|\theta)}}{\frac{\omega}{\sqrt{k}}} \right)^2 \mathrm{d}\lambda(\xi) = \int \left(\tfrac{\partial}{\partial\theta} \sqrt{f_X(\xi|\theta)} \right)^2 \mathrm{d}\lambda(\xi),$$

for all $\theta \in \Theta$, *where* $4 \int (\frac{\partial}{\partial\theta} \sqrt{f_X(\xi|\theta)})^2 \mathrm{d}\lambda(\xi) = \mathcal{I}(\theta) > 0$ *is a bounded function of* θ. *Let*

$$Y_{kn}(\omega) = \log\{f_{kn}(\xi_{kn}|\theta + \tfrac{\omega}{\sqrt{k}}) / f_{kn}(\xi_{kn}|\theta)\},$$

and define $F_{kn}(y) = P_{kn,\theta}\{Y_{kn}(\omega) \leq y\}$. *Then, the following conditions hold: For all* $\epsilon > 0$, *as* $k \to \infty$

$$\sum_{n=1}^{k} \int_{|y|\geq\epsilon} \mathrm{d}F_{kn}(y) = o(1); \tag{6.7}$$

For some $\gamma > 0$, *as* $k \to \infty$, *we have both*

$$\sum_{n=1}^{k} \int_{|y|<\gamma} y \ \mathrm{d}F_{kn}(y) = -\tfrac{1}{2}\omega^2 \mathcal{I}(\theta) + o(1) \tag{6.8}$$

and

$$\sum_{n=1}^{k} \int_{|y|<\gamma} y^2 \mathrm{d}F_{kn}(y) = \omega^2 \mathcal{I}(\theta) + o(1). \tag{6.9}$$

Comments: Conditions (i)-(iii) are quite standard, and conditions (iv)-(v) are bounded-variance conditions. The function $\mathcal{I}(\theta)$ appearing in (v) is usually referred as Fisher's information measure about θ provided by a single observation of X with density $f_X(\xi|\theta)$. We also observe that the triangular array $\{\xi_{kn}\}$, $n = 1, 2, ..., k$, $k = 1, 2, ...$, induce a triangular array $\{Y_{kn}(\omega)\}$, $n = 1, 2, ..., k$, $k = 1, 2, ...$, where each row also consists of independent random terms. The general ideas of the proof of Theorem 6.1 can be found in Venegas (1990).

Conditions (6.7)-(6.9) satisfy the Kolmogorov three-series criterion for normal convergence (see, for instance, Loève, ch. VI),implying

$$\log T_k(\omega) = \sum_{n=1}^{k} \log \frac{f_{kn}(\xi_{kn}|\theta + \frac{\omega}{\sqrt{k}})}{f_{kn}(\xi_{kn}|\theta)} \xrightarrow{\mathcal{L}} U(\omega,\theta)\big[Z - \tfrac{1}{2}U(\omega,\theta)\big]$$

as $k \to \infty$. □

Next, we shall state the limit conditions, so that the passage of the limit under the integral signs in (6.1) holds. From the notation in (6.5) and (6.6), it will be convenient to introduce the random variable $U_k = \int T_k(\omega)W_k(\omega)\mathrm{d}\mu(\omega)$.

Theorem 6.2 *Under the conditions of Theorem* 6.1, *assuming also that:*
(vii) There exist $c > 0$ *and* $\tau > 0$ *such that*

$$\int \big|\, \pi(\theta + u) - \pi(\theta) \,\big| \mathrm{d}\mu(\theta) \leq c|u|^{\tau}; \tag{6.10}$$

(viii) For all $\beta > 0$, *as* $k \to \infty$

$$\int_{|\omega|>k^{\beta}} \Big(T_k(\omega)W_k(\omega) - T_k(\omega)\Big)\mathrm{d}\mu(\omega) \xrightarrow{\mathrm{P}} 0; \tag{6.11}$$

(ix) The sequence of random variables $\{\log U_k\}_{n=1}^{\infty}$ *satisfies*

$$\lim_{\alpha\to\infty} \sup_{n\geq 1} \int_{|\log U_k|\geq\alpha} |\log U_k| \mathrm{d}P = 0, \tag{6.12}$$

where $P\{\Psi_k \in \mathbf{A},\ \theta \in B\} = \int_B \pi(\theta) \int_{\mathbf{A}} f(\Psi_k|\theta)\mathrm{d}\nu(\Psi_k)d\mu(\theta)$. *Then, as* $k \to \infty$

$$\mathcal{E}(\Psi_k, \pi) + H(Z) = -\int \pi(\theta) \log \frac{\pi(\theta)}{\sqrt{\mathcal{I}_k(\theta)}} \mathrm{d}\mu(\theta) + o(1), \tag{6.13}$$

where $H(Z)$ *is Shannon's information measure of a random variable* $Z \sim \mathcal{N}(0,1)$ *and* $\mathcal{I}_k(\theta) = k\mathcal{I}(\theta)$ *is Fisher's information measure about* θ.

Comments: Condition (vii) is a smoothness condition, and (viii) is a convergence condition. Condition (ix) simply says that the sequence of random variables $\{\log U_k\}_{k=1}^{\infty}$ is uniformly integrable with respect to P.
Proof: Let $\epsilon > 0$ be arbitrary, then from Tchebyshev's inequality and (6.10)

$$\begin{aligned}
&P\left\{\left|\int_{-k^{\beta}}^{k^{\beta}} \Big(T_k(\omega)W_k(\omega) - T_k(\omega)\Big)d\mu(\omega)\right| > \epsilon\right\} \\
&\leq \frac{1}{\epsilon}\int\int_{-k^{\beta}}^{k^{\beta}} \left(\int T_k(\omega) f(\Psi_k|\theta)d\nu(\Psi_k)\right) \big|\, \pi(\theta + \tfrac{\omega}{\sqrt{k}}) - \pi(\theta) \,\big| d\mu(\omega)d\mu(\theta) \\
&= \frac{1}{\epsilon}\int_{-k^{\beta}}^{k^{\beta}} \left(\int \big|\, \pi(\theta + \tfrac{\omega}{\sqrt{k}}) - \pi(\theta) \,\big| d\mu(\theta)\right) d\mu(\omega) \leq \frac{c}{\epsilon k^{\frac{\tau}{2}}} \int_{-k^{\beta}}^{k^{\beta}} |\omega|^{\tau} d\mu(\omega) \\
&\leq \frac{2ck^{(\tau+1)\beta}}{\epsilon k^{\frac{\tau}{2}}}.
\end{aligned} \tag{6.14}$$

Choosing $\beta > 0$ such that $(1 + \tau^{-1})\beta < \frac{1}{2}$ and letting $k \to \infty$ in (6.14), we conclude

$$\int_{-k^\beta}^{k^\beta} \Big(T_k(\omega) W_k(\omega) - T_k(\omega) \Big) d\mu(\omega) \xrightarrow{\mathrm{P}} 0. \tag{6.15}$$

Hence, from (6.11) and (6.15), we obtain

$$\int T_k(\omega) W_k(\omega) d\mu(\omega) - \int T_k(\omega) d\mu(\omega) \xrightarrow{\mathrm{P}} 0 \tag{6.16}$$

as $k \to \infty$.

On the other hand, from Theorem 6.1, as $k \to \infty$

$$\int T_k(\omega) d\mu(\omega) \xrightarrow{\mathcal{L}} \int \exp\big\{ U(\omega, \theta) \big[Z - \tfrac{1}{2} U(\omega, \theta) \big] \big\} d\mu(\omega) = \sqrt{\tfrac{2\pi}{\mathcal{I}(\theta)}}\, e^{\frac{1}{2} Z^2}. \tag{6.17}$$

From (6.16) and (6.17), and the continuity of the logarithmic function on $(0, \infty)$, we arrive at

$$\log U_k = \log \int T_k(\omega) W_k(\omega) d\mu(\omega) \xrightarrow{\mathcal{L}} \log \sqrt{\tfrac{2\pi}{\mathcal{I}(\theta)}} + \tfrac{1}{2} Z^2. \tag{6.18}$$

Finally, from (6.12) and (6.18) we obtain (see, for instance, Billingsley, ch. 1), as $k \to \infty$

$$\begin{aligned} &\int \int (\log U_k) f(\Psi_k | \theta) \pi(\theta) d\nu(\Psi_k) d\mu(\theta) \\ &= \int \int \left(\log \sqrt{\tfrac{2\pi}{\mathcal{I}(\theta)}} + \tfrac{1}{2} z^2 \right) d\Phi(z) \pi(\theta) d\mu(\theta) + o(1) \\ &= H(Z) - \int \pi(\theta) \log \sqrt{\mathcal{I}(\theta)} d\mu(\theta) + o(1), \end{aligned} \tag{6.19}$$

where Φ is the distribution of a random variable $Z \sim \mathcal{N}(0, 1)$. The last equality in (6.19) transforms (6.1) into (6.13). □

7. Summary and Conclusions

We have extended the stochastic quasigradient method to incorporate prior information. A number of prior information patterns has been studied. The convergence conditions, although somehow strict, are easy to follow making the proofs reasonably straightforward. We have obtained a limiting distribution representation for the expected information provided by the sequence of subgradients generated by the algorithm. Our work misses, however, to analyze the case when θ consists of dependent random variables. More work has to be done in this direction.

References

Billingsley, P., (1968), Convergence of Probability Measures, New York: Wiley.

de Haan, L., (1981), Estimation of the Minimum of a Function using order Statistics, JASA, Theory and Methods, Vol. 79, pp. 467-469

Dorea, C. C. Y., (1991), Effort Associated with a Class of Random Optimization Methods, Mathematical Programming, Vol. 50, pp. 91-98.

Dorea, C. C. Y., (1987), Estimation of Extreme Values and the Extreme Points, Annals of Inst. Stat. Math. Vol. 39, pp. 37-48.

Galambos, A., (1978), The asymptotic Theory of Extreme Order Statistics, John Wiley and Sons, New York.

Gradshteyn, I. S. and Ryzhik, I. M., (1980), Table of Integrals, Series, and Products, New york: Academic Press.

Jaynes, E. T., (1957), Information Theory and Statistical Mechanics I, Phys. Rev., Vol. 106, pp. 620-630.

Kullback, S., (1956), Information Theory and Statistics. New York, Wiley.

Lindley, D. V., (1956), On a Measure of Information Provided by an Experiment, Ann. Math. Statist., Vol. 27, pp. 986-1005.

Loève, M., (1977), Probability Theory, Vol I, New York, Springer-Verlag.

Shannon, C. E., (1948), A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379-423. Table of Integrals,

Shore, J. E. and Johnson R. W., (1980), Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross Entropy, IEEE Trans. Inform. Theory, Vol. IT-26, pp. 26-37.

Shore, J. E. and Johnson R. W., (1981), Properties of Cross-Entropy Minimization, IEEE Trans. Inform. Theory, Vol. IT-27, pp.472-482.

Venegas F., (1992), Entropy Maximization and Cross-Entropy Minimization. A Matrix Approach, Agrociencia, Serie Matemáticas Aplicadas, Estadística y Computación, Vol. 2, 3, pp. 71-76.

Venegas F., (1990), On Regularity and Optimality Conditions for Maximum Entropy Priors, The Brazilian Journal of Probability and Statistics (Rebrape), Vol. 4, pp. 105-136.

Venegas F., (1990a), Información Suplementaria a Priori, Aspectos Computacionales y Clasificación, Interamerican Statistical Institute (Estadística, IASI), Vol. 42, No. 139, pp. 64-80.

Venegas F., (1990b), Supplementary Information, Contributions to Probability and Mathematical Statistics, Vol. 4, pp.228-237.