The Annals of Statistics
1988, Vol. 16, No. 3, 1069–1112

LARGE SAMPLE THEORY OF EMPIRICAL DISTRIBUTIONS IN BIASED SAMPLING MODELS

BY RICHARD D. GILL, YEHUDA VARDI AND JON A. WELLNER

Centrum voor Wiskunde en Informatica, AT & T Bell Laboratories and University of Washington

Vardi (1985a) introduced an s-sample model for biased sampling, gave conditions which guarantee the existence and uniqueness of the nonparametric maximum likelihood estimator $\mathbb{G}_n$ of the common underlying distribution G and discussed numerical methods for calculating the estimator.

Here we examine the large sample behavior of the NPMLE $\mathbb{G}_n$, including results on uniform consistency of $\mathbb{G}_n$, convergence of $\sqrt{n}(\mathbb{G}_n - G)$ to a Gaussian process and asymptotic efficiency of $\mathbb{G}_n$ as an estimator of G. The proofs are based upon recent results for empirical processes indexed by sets and functions and convexity arguments. We also give a careful proof of identifiability of the underlying distribution G under connectedness of a certain graph $\mathbf{G}$.

Examples and applications include length-biased sampling, stratified sampling, "enriched" stratified sampling, "choice-based" sampling in econometrics and "case-control" studies in biostatistics.

A final section discusses design issues and further problems.

1. Introduction: Biased sampling models and Vardi's nonparametric MLE.

Biased sampling models. Let G be an unknown distribution function (df) of a real-valued random variable Y. If iid observations $Y_1, \dots, Y_n$ are available, then the classical nonparametric maximum likelihood estimator of G is simply the empirical df $n^{-1}\sum_{i=1}^n 1_{(-\infty, x]}(Y_i)$ of the Y_i's.

In a biased sampling situation we do not observe Y_i's iid G, but instead we observe $X_1, \dots, X_n$ iid F where F is a distribution resulting from biased sampling of G according to some known biasing or weight function w. If w is a nonnegative function, then the "w-biased" distribution function F is

$$F(x) \equiv \frac{\int_{-\infty}^x w(y)\,dG(y)}{\int_{-\infty}^\infty w(y)\,dG(y)} \equiv \int_{-\infty}^x \frac{w(y)}{W}\,dG(y) \tag{1.1}$$

for $-\infty < x < \infty$ with

$$W \equiv W(G) \equiv \int w(y)\,dG(y).$$

The one-sample biased sampling problem is to estimate G on the basis of an iid

Received July 1985; revised January 1988.

AMS 1980 *subject classifications.* Primary 62G05, 60F05; secondary 62G30, 60G44.

Key words and phrases. Asymptotic theory, case-control studies, choice based sampling, empirical processes, enriched stratified sampling, graphs, length-biased sampling, Neyman allocation, nonparametric maximum likelihood, selection bias models, stratified sampling, Vardi's estimator.

sample $X_1, \ldots, X_n$ from F.

EXAMPLE 1.1 (Length-biased sampling). Suppose that G is a df on $R^+ \equiv [0, \infty)$ with positive finite mean $\mu \equiv \int_0^\infty y\,dG(y)$ and let $w = x$ for $x \geq 0$. Then $W = \mu$ and F in (1.1) becomes

$$F(x) = \frac{1}{\mu}\int_0^x y\,dG(y) \quad \text{for } x \geq 0. \tag{1.2}$$

This is the *length-biased* distribution corresponding to G and is well known as the limiting distribution of the "total life" in renewal theory; see, e.g., Feller [(1966), page 371]. In this case it follows easily from (1.2) that if $G(0) = 0$,

$$G(x) = \frac{\int_0^x y^{-1}\,dF(y)}{\int_0^\infty y^{-1}\,dF(y)},$$

a relationship which was used by Cox (1969) to discuss estimation of G. We will return to this example in Section 4.

Note that in the one-sample biased sampling model (1.1) we can really only hope to estimate the conditional distribution

$$G^+(x) \equiv G((-\infty, x] \cap [w > 0])/G([w > 0])$$

which reduces to G if w is strictly positive. This follows from the homogeneity of degree 0 of F as a function of G: If $F_G \equiv F$ in (1.1), then $F_{cG} = F_G$ for any $c > 0$.

Vardi (1982, 1985a) generalized the one-sample biased sampling model (1.1) to allow for s different biased samples as follows: Suppose that $w_1, \ldots, w_s$ are given nonnegative weight functions and that G is an (unknown) df. The corresponding biased distributions are

$$F_i(x) \equiv \frac{\int_{-\infty}^x w_i(y)\,dG(y)}{\int_{-\infty}^\infty w_i(y)\,dG(y)} \equiv \int_{-\infty}^x \frac{w_i(y)}{W_i}\,dG(y) \tag{1.3}$$

for $-\infty < x < \infty$ and $i = 1, \ldots, s$ with

$$W_i \equiv W_i(G) \equiv \int w_i(y)\,dG(y) < \infty.$$

In the s-sample biased sampling model, we observe s different independent samples

$$X_{i1}, \ldots, X_{in_i} \text{ iid } F_i, \qquad i = 1, \ldots, s. \tag{1.4}$$

The problem now is to use all of the $n \equiv n_1 + \cdots + n_s$ observations in the s independent samples to (efficiently) estimate the underlying df G. Or to put it another way, we want to find a *bias-corrected estimator* which untangles (corrects for) the biasing involved in the distributions F_i.

A necessary and sufficient condition for existence of a unique nonparametric maximum likelihood estimator $\mathbb{G}_n$ of G in the s-sample biased sampling model given by (1.3) and (1.4) was given by Vardi (1985a). For a definition of a nonparametric maximum likelihood estimator in this setting, see Vardi (1985a) or, in general, Kiefer and Wolfowitz (1956); for further discussion see Gill (1988). As will be seen, the asymptotic version of Vardi's condition assures the identifiability of G based on the biased distributions $F_1, \ldots, F_s$.

Our main goal in this paper is to give a thorough treatment of the asymptotic distribution theory of Vardi's (1985a) nonparametric maximum likelihood estimator $\mathbb{G}_n$ under minimal assumptions. Since the results of Vardi (1985a) are valid not only for real-valued random variables X as already introduced, but for any X in a general sample space $\mathbf{X}$ (in particular for $\mathbf{X} = R^k$), we study $\mathbb{G}_n$ as an empirical measure using the asymptotic theory of empirical measures and processes. This project was carried out independently by Gill and Wellner (1985) for a general sample space $\mathbf{X}$ and by Vardi (1985b) in the case $\mathbf{X} = R^1$. The present paper has resulted from combining those two earlier efforts. Simplification of some of the earlier arguments in Gill and Wellner (1985) has resulted from conversations with Ya'acov Ritov.

The paper is organized as follows: The remainder of this section is devoted to identifiability issues and to a heuristic introduction to Vardi's nonparametric maximum likelihood estimator. The key identifiability Proposition 1.1 is basic. The main results giving consistency and asymptotic normality are stated in Section 2 and optimality of the estimator is established in Section 3. Examples and applications are developed further in Section 4. The proofs for Section 2 are deferred to Section 5. Finally some remaining problems are discussed briefly in Section 6.

Identifiability. The first order of business is identifiability. To begin with, we face the same difficulty as in the one-sample biased sampling model: We cannot hope to estimate G off the set

$$\mathbf{X}^+ \equiv \bigcup_{i=1}^{s} [x\colon w_i(x) > 0] = \left[x\colon w_i(x) > 0 \text{ for some } i = 1, \ldots, s\right]. \tag{1.5}$$

Hence we simplify notation by assuming throughout:

ASSUMPTION S (Support). $\mathbf{X}^+ \equiv \mathbf{X}$.

If Assumption S fails, so that $\mathbf{X}^+$ is a proper subset of $\mathbf{X}$, then we must replace G by $G^+ \equiv G(\cdot\,|\mathbf{X}^+)$.

However there is now a further difficulty: If the w_i's have disjoint supports, then we can only estimate $G_i \equiv G(\cdot\,|\mathbf{X}_i)$ with $\mathbf{X}_i \equiv [x\colon w_i(x) > 0]$. Thus G itself is not identifiable in general under Assumption S; see, e.g., Example 3a. Fortunately, however, G is identifiable if a simple graph condition holds. To state this condition, consider the graph $\mathbf{G}$ on the s vertices $i = 1, \ldots, s$ defined as follows: $i \leftrightarrow j$ if and only if

$$\int 1_{[w_i > 0]} 1_{[w_j > 0]}\, dG > 0$$

or equivalently if

$$\int 1_{[w_i > 0]} w_j \, dG / W_j = \int 1_{[w_i > 0]} \, dF_j > 0.$$

To say that the graph **G** is connected just means that every pair of i, j is connected by a path. Since connectedness of the graph **G** is the key condition of this paper, we label it here as

ASSUMPTION C (Connectedness). The graph **G** is connected.

PROPOSITION 1.1 (Identifiability of G). *The distribution function G is identifiable if and only if the graph* **G** *is connected* (*Assumption* C *holds*). *That is, the map from G to the collection of distributions $F_1, \dots, F_s$ is one-to-one if and only if* **G** *is connected.*

Proposition 1.1 has a straightforward generalization to the case of a disconnected graph **G**.

PROPOSITION 1.2. *If $\mathbf{X}_1, \dots, \mathbf{X}_r$ for some $1 \le r \le s$ are the unions of the supports of the w_j's in the each of the r connected subgraphs of the graph* **G** (*so the $\mathbf{X}_i$'s are disjoint a.s. G*), *then the identifiable parameters are the r conditional distributions $G_i \equiv G(\cdot \mid \mathbf{X}_i)$.*

We defer the proof of Proposition 1.1 to the end of this section. Proposition 1.1 is closely related to the results of Cosslett (1981) and Vardi (1985a). A condition equivalent to the connectedness assumption C was apparently first introduced and used by Cosslett (1981). Vardi (1985a) showed that if **G** is connected, then with probability 1 a unique nonparametric maximum likelihood estimator exists asymptotically as $\min n_i \to \infty$. Vardi [(1985a), pages 196–197] also conjectured that Assumption C (connectedness of the graph **G**) implies identifiability. This conjecture was first proved by Gill and Wellner [(1985), Proposition 2.2]. Ya'acov Ritov suggested an alternative proof of the complete result to us at Oberwolfach in March 1987. We present Ritov's proof in the following text.

The following two examples illustrate the s-sample biased sampling model and the key connectedness assumption C.

EXAMPLE 1.2 (Ordinary and length-biased sampling). Suppose that G is a df on R^+ with mean $\mu = \int_0^\infty x \, dG(x) < \infty$ as in Example 1, but now suppose $s = 2$ and that $w_1(x) = 1$ and $w_2(x) = x$ for $x \ge 0$. Thus the first sample is from G and the second sample is from the length biased distribution corresponding to G. This special case of the model (1.3) and (1.4) was studied by Vardi (1982); we will consider it further in Section 4. Note that the key connectedness assumption C needed for identifiability is easily satisfied here since

$$\int_0^\infty 1_{[w_1(x)>0]} 1_{[w_2(x)>0]}\, dG(x) = \int_0^\infty dG(x) = G(0,\infty) > 0;$$

of course G is clearly identifiable in this model since $F_1 = G$.

Now we give both (a) a simple example which illustrates failure of the key connectedness assumption C (and hence also identifiability of G) and (b) a related example which repairs the difficulty.

EXAMPLE 1.3a (Stratified sampling). Suppose that for some s (≥ 3) the collection of (measurable) sets $\{D_1, \dots, D_s\}$ is a partition of the sample space $\mathbf{X} = R^1$: $\bigcup_{i=1}^s D_i = \mathbf{X} = R^1$ and $D_i \cap D_j = \varnothing$ for $i \neq j$. If the weight functions in the biased sampling model are $w_i(x) = 1_{D_i}(x)$ for $i = 1, \dots, s$, then

$$F_i(x) = G(x|D_i) \equiv G((-\infty, x] \cap D_i)/G(D_i)$$

is just the conditional distribution given the event D_i. This is just stratified sampling from the strata $D_1, \dots, D_s$. It is clear that estimation of G itself is not possible without knowledge of the stratum probabilities $G(D_i)$; we can estimate relative (conditional) probabilities within each separate stratum, but we have no way to relate the conditional probabilities without knowledge of the $W_i = G(D_i)$. Note that this special case of the biased sampling model corresponds to stratified sampling in the survey (finite population) sampling literature: There the standard assumption is that the stratum sizes N_i, $i = 1, \dots, s$, are known [corresponding to known $G(D_i)$]; see, e.g., Cochran [(1963), page 87]. In terms of the key connectedness assumption C, we have

$$\int 1_{[w_i>0]} 1_{[w_j>0]}\, dG = \int_{D_i \cap D_j} dG(x) = 0$$

for all pairs $i \neq j$ since $\{D_i\}$ is a partition. Thus Assumption C fails completely. Of course this difficulty is easily remedied by simply sampling with some positive sampling fraction from the entire distribution G as follows.

EXAMPLE 1.3b ("Enriched" stratified sampling). Suppose that $D_1, \dots, D_s$ is a partition of $\mathbf{X} = R^1$ as in Example 1.3a, but now suppose we have iid samples from the $s+1$ distributions $(F_1, \dots, F_{s+1}) = (F_1, \dots, F_s, G)$ corresponding to the weight functions $(w_1, \dots, w_{s+1}) = (1_{D_1}, \dots, 1_{D_s}, 1)$. (Note that $s+1$ now plays the role of s.) With this sampling scheme we can use the $(s+1)$st sample to estimate the stratum probabilities $G(D_i)$, then combine appropriately to estimate G. Note that we now have

$$\int 1_{[w_i(x)>0]} 1_{[w_{s+1}(x)>0]}\, dG(x) = \int 1_{D_i}(x)\, dG(x) = G(D_i) > 0$$

for $i = 1, \dots, s$ so that assumption C holds: The graph $\mathbf{G}$ is connected via $s+1$.

How to combine: A heuristic approach to Vardi's nonparametric MLE. We now give a naive approach to Vardi's (1985a) nonparametric MLE $\mathbb{G}_n$ of G. We first consider the case $\mathbf{X} = R^1$ and then extend to a general $\mathbf{X}$. We will not prove that the estimator is the MLE; our approach here is instead completely heuristic.

First, some notation: Let $n \equiv n_1 + \cdots + n_s$ denote the total sample size and write $\lambda_{ni} \equiv n_i/n$ for the sampling fraction from F_i for $i = 1, \ldots, s$. We also set

$$\bar{F}_n \equiv \lambda_{n1} F_1 + \cdots + \lambda_{ns} F_s; \tag{1.6}$$

$\bar{F}_n$ is the "average df." Note that by (1.3) we have

$$\begin{aligned} \bar{F}_n(x) &= \sum_{i=1}^{s} \lambda_{ni} \int_{-\infty}^{x} \frac{w_i(y)}{W_i}\, dG(y) \\ &= \int_{-\infty}^{x} \left[\sum_{i=1}^{s} \frac{\lambda_{ni} w_i(y)}{W_i} \right] dG(y). \end{aligned} \tag{1.7}$$

It follows immediately from (1.7) and assumption S that

$$G(x) = \int_{-\infty}^{x} \left[\sum_{i=1}^{s} \frac{\lambda_{ni} w_i(y)}{W_i} \right]^{-1} d\bar{F}_n(y). \tag{1.8}$$

From (1.8) it is clear that we can easily estimate G if the $W_i = W_i(G)$ are *known* since the empirical df $\mathbb{F}_n$ of all the observations "estimates" the average df $\bar{F}_n$ of (1.5). We write

$$\mathbb{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} 1_{(-\infty, x]}(X_{ij}) = \sum_{i=1}^{s} \lambda_{ni} \mathbb{F}_{ni}(x), \tag{1.9}$$

where

$$\mathbb{F}_{ni}(x) \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} 1_{(-\infty, x]}(X_{ij}). \tag{1.10}$$

Replacing $\bar{F}_n$ on the right side of (1.8) by the empirical df yields a nondecreasing function $\mathbb{D}_n$ which has $\mathbb{D}_n(\infty) \neq 1$ in general. To get an estimate of G (really a "pseudoestimate" since the W_i's are unknown) we could use

$$\mathbb{G}_n^0(x) \equiv \frac{\mathbb{D}_n(x)}{\mathbb{D}_n(\infty)} = \frac{\int_{-\infty}^{x} \left[\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/W_i\right]^{-1} d\mathbb{F}_n(y)}{\int_{-\infty}^{\infty} \left[\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/W_i\right]^{-1} d\mathbb{F}_n(y)}. \tag{1.11}$$

Now the question is: How can we estimate the $W_i = W_i(G)$'s? Once we have estimates of the W_i's, we can use them on the right side of (1.11) to obtain an honest estimate of G. First note that since $W_i = W_i(G) = \int w_i(y)\, dG(y)$ (> 0 without loss of generality), it follows easily from (1.8) that $W_1, \ldots, W_s$ satisfy the equations

$$\begin{aligned} 1 &= \frac{1}{W_i} \int w_i(y)\, dG(y) \\ &= \frac{1}{W_i} \int \frac{w_i(y)}{\sum_{j=1}^{s} (\lambda_{nj} w_j(y))/W_j}\, d\bar{F}_n(y) \\ &\equiv H_{ni}(W_1, \ldots, W_s) \end{aligned} \tag{1.12}$$

for $i = 1, \dots, s$, where

$$H_{ni}(u_1, \dots, u_s) \equiv \frac{1}{u_i} \int \frac{w_i(y)}{\sum_{j=1}^s (\lambda_{nj} w_j(y))/u_j} \, d\bar{F}_n(y) \tag{1.13}$$

for $i = 1, \dots, s$. Since the functions $H_{ni}(u_1, \dots, u_s)$ can be estimated from the data by simply replacing $\bar{F}_n$ by the empirical df $\mathbb{F}_n$, (1.12) suggests that we *define* estimates $(\mathbb{W}_{n1}, \dots, \mathbb{W}_{ns})$ of $(W_1, \dots, W_s)$ as a solution (if it exists) of the system of s equations

$$1 = \frac{1}{\mathbb{W}_{ni}} \int \frac{w_i(y)}{\sum_{j=1}^s (\lambda_{nj} w_j(y))/\mathbb{W}_{nj}} \, d\mathbb{F}_n(y) \equiv \mathbb{H}_{ni}(\mathbb{W}_{n1}, \dots, \mathbb{W}_{ns}) \tag{1.14}$$

for $i = 1, \dots, s$, where

$$\mathbb{H}_{ni}(u_1, \dots, u_s) \equiv \frac{1}{u_i} \int \frac{w_i(y)}{\sum_{j=1}^s (\lambda_{nj} w_j(y))/u_j} \, d\mathbb{F}_n(y). \tag{1.15}$$

We note immediately, however, that the functions $H_{ni}(u_1, \dots, u_s)$ in (1.13) and their empirical counterparts in (1.15) are homogeneous of degree 0 in $\underline{u} \equiv (u_1, \dots, u_s)$, i.e.,

$$H_{ni}(cu_1, \dots, cu_s) = H_{ni}(u_1, \dots, u_s) \quad \text{for all } c \neq 0.$$

Therefore (1.12) and (1.14) only determine $\underline{W} \equiv (W_1, \dots, W_s)$ and $\underline{\mathbb{W}}_n \equiv (\mathbb{W}_{n1}, \dots, \mathbb{W}_{ns})$, respectively, up to a constant multiple. Fortunately this is enough, since the right side of (1.11) is also homogeneous of degree 0 in $\underline{W}$.

We can therefore take the following tack: Let $V_i \equiv W_i/W_s$ and $\mathbb{V}_{ni} \equiv \mathbb{W}_{ni}/\mathbb{W}_{ns}$ for $i = 1, \dots, s$ so that $V_s \equiv \mathbb{V}_{ns} \equiv 1$. Then solve the $s - 1$ equations

$$1 = \mathbb{H}_{ni}(\mathbb{V}_{n1}, \dots, \mathbb{V}_{n,s-1}, 1), \qquad i = 1, \dots, s-1, \tag{1.16}$$

for $\mathbb{V}_{n1}, \dots, \mathbb{V}_{n,s-1}$. Then, letting $\mathbb{G}_n^0(x; W_1, \dots, W_s)$ denote the right-hand side of (1.11), the resulting estimator $\mathbb{G}_n$ of G is simply

$$\begin{aligned} \mathbb{G}_n(x) &\equiv \mathbb{G}_n^0(x; \mathbb{V}_{n1}, \dots, \mathbb{V}_{n,s-1}, 1) \\ &= \frac{\int_{-\infty}^x [\sum_{i=1}^s \lambda_{ni}(w_i(y))/\mathbb{V}_{ni}]^{-1} \, d\mathbb{F}_n(y)}{\int_{-\infty}^\infty [\sum_{i=1}^s \lambda_{ni}(w_i(y))/\mathbb{V}_{ni}]^{-1} \, d\mathbb{F}_n(y)}. \end{aligned} \tag{1.17}$$

In fact, as Vardi (1985a) shows, the equations (1.16) have a unique solution with probability 1 as $n \to \infty$ if the connectedness assumption C holds. Then $\mathbb{G}_n$ of (1.17) is the nonparametric MLE of G. Finally, note that with $\mathbb{G}_n$ as in (1.17), we can estimate $\underline{\mathbb{W}}_n \equiv (\mathbb{W}_{n1}, \dots, \mathbb{W}_{ns})$ by

$$\begin{aligned} \mathbb{W}_{ni} &= \int w_i \, d\mathbb{G}_n = \frac{\mathbb{V}_{ni} \mathbb{H}_{ni}(\mathbb{V}_{n1}, \dots, \mathbb{V}_{n,s-1}, 1)}{\int_{-\infty}^\infty [\sum_{i=1}^s \lambda_{ni}(w_i(y))/\mathbb{V}_{ni}]^{-1} \, d\mathbb{F}_n(y)} \\ &= \mathbb{V}_{ni} \mathbb{W}_{ns} \quad [\text{by } (1.16)] \end{aligned} \tag{1.18}$$

with

$$\mathbb{W}_{ns} \equiv \frac{1}{\int_{-\infty}^\infty [\sum_{i=1}^s \lambda_{ni}(w_i(y))/\mathbb{V}_{ni}]^{-1} \, d\mathbb{F}_n(y)}. \tag{1.19}$$

To build more feeling for the equations (1.6), we again consider Example 1.3b.

EXAMPLE 1.3b (continued). In this example $s+1$ plays the role of s and the integrand in the definition of H_{ni} in (1.15) becomes, at the point $(u_1, \dots, u_s, 1)$,

$$1_{D_i}(x) \sum_{j=1}^{s} \left[\lambda_{n,s+1} + (\lambda_{ni}/u_j)\right]^{-1} 1_{D_j}(x)$$

(since the denominator equals $[\lambda_{n,s+1} + (\lambda_{nj}/u_j)]$ on D_j), and hence the function H_{ni} is given by

$$\begin{aligned} \mathbb{H}_{ni}(u_1, \dots, u_s, 1) &= (1/u_i)\left[\lambda_{n,s+1} + (\lambda_{n,i}/u_i)\right]^{-1} \mathbb{F}_n(D_i) \\ &= \left[u_i \lambda_{n,s+1} + \lambda_{n,i}\right]^{-1} \mathbb{F}_n(D_i) \end{aligned}$$

for $i = 1, \dots, s$. But

$$n\mathbb{F}_n(D_i) = n_i + n_{s+1}\mathbb{F}_{n,s+1}(D_i),$$

or

$$\mathbb{F}_n(D_i) = \lambda_{ni} + \lambda_{n,s+1}\mathbb{F}_{n,s+1}(D_i).$$

Hence in this case the equations (1.16) become

$$1 = \frac{\lambda_{ni} + \lambda_{n,s+1}\mathbb{F}_{n,s+1}(D_i)}{\lambda_{ni} + \lambda_{n,s+1}\mathbb{V}_{ni}} \quad \text{for } i = 1, \dots, s,$$

which yields the intuitive result $\mathbb{V}_{ni} = F_{n,s+1}(D_i)$, $i = 1, \dots, s$. After a little calculation, (1.17) reduces in this special case to

$$\mathbb{G}_n(x) = \sum_{i=1}^{s} \mathbb{F}_{n,s+1}(D_i) \frac{\mathbb{F}_n(D_i \cap (-\infty, x])}{\mathbb{F}_n(D_i)},$$

which is just the estimator that should be expected on intuitive grounds or by analogy with the familiar finite sampling model for stratified sampling. Note that $\mathbb{W}_{n,s+1} = \mathbb{G}_n(1) = 1$ so that $\mathbb{V}_{ni} = \mathbb{W}_{ni}$ for $i = 1, \dots, s$ in this example.

The biased sampling model (1.3) and (1.4) and the preceding approach to the nonparametric MLE for the case $\mathbf{X} = R^1$ extends easily to a general sample space $\mathbf{X}$ with σ-field of subsets $\mathbf{B}$. If G is an unknown distribution (i.e., probability measure) on $(\mathbf{X}, \mathbf{B})$, and $w_1, \dots, w_s$ are nonnegative (measurable) weight functions defined on $\mathbf{X}$, then the corresponding biased distributions (probability measures) $F_1, \dots, F_s$ are given by

$$(1.20) \qquad \begin{aligned} F_i(A) &\equiv \frac{\int_A w_i(y)\, dG(y)}{\int_{\mathbf{X}} w_i(y)\, dG(y)} \equiv \int_A \frac{w_i(y)}{W_i}\, dG(y) \\ &\equiv \frac{G(1_A w_i)}{G(w_i)} \quad \text{for } A \in \mathbf{B} \text{ and } i = 1, \dots, s; \end{aligned}$$

here and frequently in the following, we use the functional or de Finetti notation $G(f) \equiv \int f\, dG$ to simplify expressions for integrals. Then in the general s-sample biased sampling model we observe

$$(1.21) \qquad X_{i1}, \dots, X_{in_i} \text{ iid } F_i, \qquad i = 1, \dots, s.$$

[In (1.21) all the X_{ij}'s take values in $\mathbf{X}$; e.g., if $\mathbf{X} = R^k$, then the X_{ij}'s are random vectors.]

To extend our development of the estimator (1.17), we introduce the empirical measures

$$\mathbb{F}_n(A) \equiv \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \delta_{X_{ij}}(A) = \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} 1_A(X_{ij}) \tag{1.22}$$

and

$$\mathbb{F}_{ni}(A) \equiv \frac{1}{n_i} \sum_{j=1}^{n} \delta_{X_{ij}}(A) \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} 1_A(X_{ij}), \qquad i = 1, \dots, s, \tag{1.23}$$

for $A \in \mathbf{B}$. Now the argument in (1.6)–(1.8) holds in general with df's replaced by measures and the set $(-\infty, x]$ replaced by a general set $A \in \mathbf{B}$. Thus (1.11) becomes

$$\mathbb{G}_n^0(A) \equiv \frac{\mathbb{D}_n(A)}{\mathbb{D}_n(\mathbf{X})} = \frac{\int_A [\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/W_i]^{-1} \, d\mathbb{F}_n(y)}{\int_{\mathbf{X}} [\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/W_i]^{-1} \, d\mathbb{F}_n(y)} \tag{1.24}$$

for $A \in \mathbf{B}$, while the argument in (1.12)–(1.17) works with df's replaced by measures throughout and yields, in place of (1.17), the estimator $\mathbb{G}_n$ of the measure G given by

$$\begin{aligned} \mathbb{G}_n(A) &\equiv \mathbb{G}_n^0(A; \mathbb{V}_{n1}, \dots, \mathbb{V}_{n,s-1}, 1) \\ &= \frac{\int_A [\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/\mathbb{V}_{ni}]^{-1} \, d\mathbb{F}_n(y)}{\int_{\mathbf{X}} [\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/\mathbb{V}_{ni}]^{-1} \, d\mathbb{F}_n(y)} \end{aligned} \tag{1.25}$$

for $A \in \mathbf{B}$. This is the estimator that we study in Section 2.

An equivalent more symmetric formulation of the definition of $\mathbb{G}_n$ and $\underline{\mathbb{W}}_n$ which avoids the $\mathbb{V}_{ni}$'s is as follows: Let $\mathbb{G}_n$ and $\underline{\mathbb{W}}_n$ be a solution with all $\mathbb{W}_{ni} > 0$ of the equations

$$\mathbb{G}_n(A) = \frac{\int_A [\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/\mathbb{W}_{ni}]^{-1} \, d\mathbb{F}_n(y)}{\int_{\mathbf{X}} [\sum_{i=1}^{s} \lambda_{ni}(w_i(y))/\mathbb{W}_{ni}]^{-1} \, d\mathbb{F}_n(y)} \quad \text{for } A \in \mathbf{B}, \tag{1.26}$$

$$1 = \mathbb{H}_{ni}(\mathbb{W}_{n1}, \dots, \mathbb{W}_{ns}), \qquad i = 1, \dots, s, \tag{1.27}$$

and

$$\begin{aligned} (\mathbb{W}_{n1}, \dots, \mathbb{W}_{ns}) &= (\mathbb{G}_n(w_1), \dots, \mathbb{G}_n(w_s)) \\ &\equiv \left(\int w_1 \, d\mathbb{G}_n, \dots, \int w_s \, d\mathbb{G}_n \right). \end{aligned} \tag{1.28}$$

Equation (1.28) can be viewed as a "self-consistency" equation and will be particularly useful in our formulation of the limit theory for $\underline{\mathbb{W}}_n$ in Section 2.

Here is a quick proof that (1.26)–(1.28) are equivalent to (1.16) and (1.25). First suppose that (1.26)–(1.28) hold with all $\mathbb{W}_{ni} > 0$. Then by homogeneity of degree 0 in $\underline{\mathbb{W}}_n$ of the right sides of (1.26) and (1.27), (1.25) and (1.16) hold with

$\mathbb{V}_{ni} \equiv \mathbb{W}_{ni}/\mathbb{W}_{ns}$, $i = 1,\dots, s-1$ and $\mathbb{V}_{ns} = 1$. Now suppose (1.16) and (1.25) hold and define $\mathbb{W}_{ns}$ as in (1.19):

$$\mathbb{W}_{ns} \equiv \left\{ \int_{\mathbf{X}} \left[\sum_{i=1}^{s} \lambda_{ni} \frac{w_i(y)}{\mathbb{V}_{ni}} \right]^{-1} d\mathbb{F}_n(y) \right\}^{-1}. \tag{1.29}$$

Then by homogeneity of degree 0 of the right sides of (1.16) and (1.25) in the V_{ni}'s, (1.26) and (1.27) hold with $\mathbb{W}_{ni} \equiv \mathbb{W}_{ns}\mathbb{V}_{ni}$ for $i = 1,\dots, s-1$. But since (1.16) holds for $i = 1,\dots, s-1$ and $\sum_{i=1}^{s} \lambda_{ni}\mathbb{H}_{ni}(u_1,\dots, u_s) = 1$, (1.16) also holds for $i = s$. Hence (1.27) holds for $i = s$ and for $\mathbb{G}_n$ of (1.26) we have, by (1.29) and (1.16),

$$\begin{aligned} \mathbb{G}_n(w_i; \underline{\mathbb{W}}_n) &= \mathbb{G}_n(w_i; \underline{\mathbb{V}}_n) \\ &= \mathbb{V}_{ni}\mathbb{H}_{ni}(\mathbb{V}_{n1},\dots, \mathbb{V}_{n,s-1}, 1)\mathbb{W}_{ns} \\ &= \mathbb{V}_{ni}\mathbb{W}_{ns} = \mathbb{W}_{ni} \end{aligned}$$

for $i = 1,\dots, s$, so (1.28) holds.

We now give a statement of Vardi's (1985a) theorem. Define a directed graph $\mathbf{G}^*$ on s vertices $\{1,\dots, s\}$ by

$$i \to j \quad \text{if and only if} \quad \int 1_{[w_i > 0]}\, d\mathbb{F}_{nj} > 0.$$

Then call $\mathbf{G}^*$ *strongly connected* if for any two vertices i and j there is a directed path from i to j and from j to i. It is well known that $\mathbf{G}^*$ is strongly connected if and only if the matrix $A = (a_{ij})$ with elements $a_{ij} = \int 1_{[w_i > 0]}\, d\mathbb{F}_{nj}$ is irreducible [see, e.g., Berman and Plemmons (1979), pages 27–30].

THEOREM 1.1 (Vardi, 1985a). *The equations* (1.16) *have a unique solution if and only if* $\mathbf{G}^*$ *is strongly connected. Then* $\mathbb{G}_n$ *defined by* (1.25) *is the unique nonparametric maximum likelihood estimate of* G. *Equivalently,* $\mathbf{G}^*$ *is strongly connected if and only if the system of equations* (1.26)–(1.28) *has a unique solution.*

If $W_j \equiv G(w_j) = \int w_j\, dG < \infty$ for all $1 \le j \le s$, then by the strong law of large numbers

$$\int 1_{[w_i > 0]}\, d\mathbb{F}_{nj} \to_{\text{a.s.}} \int 1_{[w_i > 0]}\, dF_j = \frac{1}{W_j} \int 1_{[w_i > 0]} w_j\, dG < \infty.$$

This proves the following corollary relating connectedness of $\mathbf{G}$ to strong connectedness of $\mathbf{G}^*$.

COROLLARY 1.1. *If the graph* $\mathbf{G}$ *is connected and* $W_i \equiv G(w_i) < \infty$ *for* $1 \le i \le s$, *then with probability* 1 *for* $n \ge$ *some* N_ω *the graph* $\mathbf{G}^*$ *is strongly connected and the conclusions of Theorem* 1 *hold.*

PROOF OF PROPOSITION 1.1. From (1.8) it follows that

$$G(x) = \int_{-\infty}^{x} \left[\sum_{i=1}^{s} \frac{\lambda_{ni} w_i(y)}{W_i} \right]^{-1} d\bar{F}_n(y) \tag{a}$$

$$= \frac{\int_{-\infty}^{x} [\Sigma_{i=1}^{s} (\lambda_{ni} w_i(y))/W_i]^{-1} d\bar{F}_n(y)}{\int_{-\infty}^{\infty} [\Sigma_{i=1}^{s} (\lambda_{ni} w_i(y))/W_i]^{-1} d\bar{F}_n(y)}$$

$$= \frac{\int_{-\infty}^{x} [\Sigma_{i=1}^{s} (\lambda_{ni} w_i(y))/V_i]^{-1} d\bar{F}_n(y)}{\int_{-\infty}^{\infty} [\Sigma_{i=1}^{s} (\lambda_{ni} w_i(y))/V_i]^{-1} d\bar{F}_n(y)}, \tag{b}$$

where $V_i \equiv W_i/W_s$, $i = 1, \ldots, s$, and it remains only to show that we can determine the V_i's as a function of the F_i's.

Consider the system of $s - 1$ equations suggested by (1.12) and the discussion following it:

$$H_{ni}(V_1, \ldots, V_{s-1}, 1) = 1, \qquad i = 1, \ldots, s-1, \tag{1.30}$$

where

$$H_{ni}(u_1, \ldots, u_s) \equiv \frac{1}{u_i} \int \frac{w_i(y)}{\Sigma_{j=1}^{s} (\lambda_{nj} w_j(y))/u_j} d\bar{F}_n(y). \tag{c}$$

We reparametrize: Let $\exp(z_j) \equiv (\lambda_{nj}/u_j)$ and set

$$K_{ni}(z_1, \ldots, z_s) \equiv \lambda_{ni} H_{ni}(\lambda_{n1} e^{-z_1}, \ldots, \lambda_{ns} e^{-z_s}) - \lambda_{ni},$$

$i = 1, \ldots, s$. Then (1.30) becomes

$$K_{ni}(Z_1, \ldots, Z_{s-1}, \log(\lambda_{ns})) = 0, \qquad i = 1, \ldots, s-1. \tag{d}$$

We want to prove that the system of $s - 1$ equations in (d) has a unique solution if $\mathbf{G}$ is connected. We show that $\underline{K}_n$ is the gradient of a convex function which is strictly convex if $\mathbf{G}$ is connected. Hence *if* a solution exists (which we know is true), it must be unique.

First note that $\underline{K}_n(\underline{z}) = \nabla D_n(\underline{z})$, where D_n: $R^s \to R^1$ is given by

$$D_n(\underline{z}) = \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(y) \right] d\bar{F}_n(y) - \sum_{i=1}^{s} \lambda_{ni} z_i. \tag{e}$$

Also, the Hessian of D_n, D_n'', is given by

$$D_n''(\underline{z})_{ij} = \int \left\{ \frac{e^{z_i} w_i(y) \delta_{ij}}{B(y)} - \frac{e^{z_i} w_i(y) e^{z_j} w_j(y)}{B^2(y)} \right\} d\bar{F}_n(y), \tag{f}$$

where $B(y) \equiv \sum_{j=1}^{s} e^{z_j} w_j(y)$. Hence for $\underline{a} \in R^s$ we have

$$\underline{a}^T D_n'' \underline{a} = \int \left\{ \frac{\sum_{i=1}^{s} a_i^2 e^{z_i} w_i(y)}{\sum_{i=1}^{s} e^{z_i} w_i(y)} - \left[\frac{\sum_{i=1}^{s} a_i e^{z_i} w_i(y)}{\sum_{i=1}^{s} e^{z_i} w_i(y)} \right]^2 \right\} d\overline{F}_n(y)$$

$$\equiv \int \left\{ \sum a_i^2 p_i(y) - \left[\sum a_i p_i(y) \right]^2 \right\} d\overline{F}_n(y)$$

$$\left[\text{where } p_i(y) \equiv \frac{e^{z_i} w_i(y)}{B(y)}\right]$$

$$= \int \mathrm{Var}_y[a_I] \, d\overline{F}_n(y) \quad [\text{where } I \sim p.(y)]$$

$$\text{(g)} \qquad \geq 0 \quad \text{for all } \underline{a} \in R^s.$$

Thus to show that the system (d) has a unique solution, it suffices to show that the upper left $(s-1) \times (s-1)$ submatrix of $D_n''(\underline{z})$ is positive definite. To do this we argue that if the graph **G** is connected, then strict inequality holds in (g) for all $\underline{a} \neq c\underline{1}$ for some $c \neq 0$.

Suppose that equality holds in (g). Since $\mathrm{Var}_y[a_I] \geq 0$ for all y, equality implies that

$$\text{(h)} \qquad \mathrm{Var}_y(a_I) = 0 \quad \text{a.e. } (\overline{F}_n) y;$$

that is, $\overline{F}_n(A) = 1$ where

$$\text{(i)} \qquad A \equiv \{ y \colon \mathrm{Var}_y[a_I] = 0 \}.$$

Let $1 \leq i, j \leq s$ with $i \neq j$. We want to show that (h) implies $a_i = a_j$ if the graph **G** is connected. By connectedness of **G** there exists a path

$$\text{(j)} \qquad i \equiv i_1 \leftrightarrow i_2 \leftrightarrow \cdots \leftrightarrow i_{m-1} \leftrightarrow i_m \equiv j$$

connecting i to j. Thus

$$\text{(k)} \qquad \int 1_{[w_{i_{k-1}} > 0]} 1_{[w_{i_k} > 0]} \, dG > 0 \quad \text{for } k = 2, \dots, m.$$

Let

$$\text{(l)} \qquad B_k \equiv \{ y \colon w_{i_{k-1}}(y) w_{i_k}(y) > 0 \}, \qquad k = 2, \dots, m.$$

Then (k) implies $G(B_k) > 0$, $k = 2, \dots, m$, and hence $\overline{F}_n(B_k) > 0$. Let $B_k^* \equiv A \cap B_k$, $k = 2, \dots, m$. Choose $y \in B_2^*$. Then $\mathrm{Var}_y[a_I] = 0$ and $p_{i_1}(y), p_{i_2}(y) > 0$ which forces $a_i \equiv a_{i_1} = a_{i_2}$ since a must be constant for the coordinates with $p_k(y) > 0$. Now choose $y \in B_3^*$. Then $\mathrm{Var}_y[a_I] = 0$ and $p_{i_2}(y), p_{i_3}(y) > 0$, which forces $a_{i_2} = a_{i_3}$. Continuing in this way yields

$$a_i = a_{i_1} = \cdots = a_{i_m} = a_j.$$

Since the same argument holds for any pair i, j, it follows that (h) implies that $\underline{a} = c\underline{1}$ for some c if **G** is connected. Hence D_n'' has rank $s-1$ and the upper left $(s-1) \times (s-1)$ submatrix is nonsingular for all $\underline{z}$ if **G** is connected. Hence the solution $V_i = W_i/W_s$ of (d) is unique; see, e.g., Ortega and Rheinboldt (1970), 4.1.4, page 94, and 4.2.9, page 101. □

Remark 1.1. Suppose that $D_n(\underline{z})$ is defined by

$$\text{(1.31)} \qquad \mathbb{D}_n(\underline{z}) = \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(y) \right] d\mathbb{F}_n(y) - \sum_{i=1}^{s} \lambda_{ni} z_i.$$

Then the argument of the preceding proof shows that $D_n(\underline{z})$ is a convex function of $\underline{z}$.

In his discussion of existence and uniqueness of the nonparametric maximum likelihood estimate under an empirical version of assumption C, Vardi (1985a) also uses the reparametrization $z_j = \log(\lambda_{nj}/u_j)$ and convexity arguments. In fact (the empirical counterpart of) (e) turns out to be the profile of a linear function of several variables on a convex region (i.e., the maximum over some of the arguments, considered as a function of the others). The reparametrization puts him into a situation formally equivalent to calculating the MLE in an exponential family.

2. Main results: Asymptotic theory for $\mathbb{G}_n$.

Consistency. The first task is to establish consistency of the estimators $\mathbb{G}_n$ of G and $\underline{\mathbb{W}}_n$ of $\underline{W}$ given by (1.25) and (1.18), respectively. The key to our consistency proofs is consistency of $\underline{\mathbb{V}}_n$, and this in turn depends on the strict convexity of D_n under the connectedness assumption C [and the resulting uniqueness of $\underline{V}$ as a solution of (1.30)] established in the proof of Proposition 1.1.

PROPOSITION 2.1 (Consistency of $\underline{\mathbb{V}}_n$ and $\underline{\mathbb{W}}_n$). *Suppose that* $\mathbf{G}$ *is connected (Assumption* C *holds) and* $W_i \equiv G(w_i) = \int w_i \, dG < \infty$, $i = 1, \dots, s$. *Then the equations* (1.16) *have (with probability* 1 *as* $n \to \infty$) *the unique solution* $\underline{\mathbb{V}}_n = (\mathbb{V}_{n1}, \dots, \mathbb{V}_{n,s-1}, 1)$ *which satisfies*

$$\underline{\mathbb{V}}_n \to_{a.s.} \underline{V} = \underline{W}/W_s \quad \text{as } n \to \infty. \tag{2.1}$$

Furthermore, $\underline{\mathbb{W}}_n$ *of* (1.18) *satisfies*

$$\underline{\mathbb{W}}_n \to_{a.s.} \underline{W} \quad \text{as } n \to \infty. \tag{2.2}$$

Now let $\mathbf{C}$ be a *Vapnik–Chervonenkis class* of subsets of $\mathbf{X}$, let h_e be a fixed nonnegative G-integrable function $[G(h_e) = \int h_e \, dG < \infty]$ and consider the collection of functions

$$\mathbf{H} \equiv \{h_e 1_C \colon C \in \mathbf{C}\}. \tag{2.3}$$

The following theorem gives consistency of both $\mathbb{G}_n$ and $\mathbb{G}_n^0$ [from (1.24) when the W_i's are known] as estimators of G uniformly over $\mathbf{H}$.

THEOREM 2.1 (Consistency of $\mathbb{G}_n$ and $\mathbb{G}_n^0$). *Suppose that* $\mathbf{G}$ *is connected (Assumption* C *holds), that* $G(w_i) = \int w_i \, dG < \infty$, $i = 1, \dots, s$, *and that* $\mathbf{H}$ *is as defined in* (2.3) *with* $G(h_e) < \infty$. *Then*

$$\|\mathbb{G}_n - G\|_{\mathbf{H}} \equiv \sup\{|\mathbb{G}_n(h) - G(h)| \colon h \in \mathbf{H}\} \to_{a.s.} 0 \tag{2.4}$$

as $n \to \infty$.

Theorem 2.1 has two straightforward corollaries.

COROLLARY 2.1. *Suppose that* $\mathbf{X} = R^d$ *and* $\mathbf{C}$ *is a Vapnik–Chervonenkis class of subsets of* $\mathbf{X}$ *(e.g., the class of all lower left orthants, or of all* *rectangles, or the class of all half spaces). If* $\mathbf{G}$ *is connected and* $G(w_i) < \infty$, $i = 1, \dots, s$, *then*

$$\|\mathbb{G}_n - G\|_{\mathbf{C}} \equiv \sup\{|\mathbb{G}_n(C) - G(C)|: C \in \mathbf{C}\} \to_{a.s.} 0$$

as $n \to \infty$.

COROLLARY 2.2. *Suppose that* $G(|h|) < \infty$, *that* $\mathbf{G}$ *is connected and that* $G(w_i) < \infty$ *for* $i = 1, \dots, s$. *Then* $\mathbb{G}_n(h) \to_{a.s.} G(h)$ *as* $n \to \infty$.

Empirical processes and notation. It will be convenient to formulate the asymptotic distribution theory for $\sqrt{n}(\underline{\mathbb{V}}_n - \underline{V})$, $\sqrt{n}(\underline{\mathbb{W}}_n - \underline{W})$ and $\sqrt{n}(\mathbb{G}_n - G)$ in terms of the s-sample empirical process

$$\mathbb{X}_n^* \equiv \sqrt{n}\left(\mathbb{F}_n - \overline{F}_n\right). \tag{2.5}$$

Here $\mathbb{F}_n$ is the empirical measure defined in (1.22) and

$$\overline{F}_n \equiv \sum_{i=1}^{s} \lambda_{ni} F_i \to \sum_{i=1}^{s} \lambda_i F_i \equiv \overline{F} \tag{2.6}$$

since we assume that

$$\lambda_{ni} \equiv \frac{n_i}{n} \to \lambda_i > 0 \quad \text{for } i = 1, \dots, s.$$

Note that

$$\mathbb{X}_n^* = \sum_{i=1}^{s} \sqrt{\lambda_{ni}}\sqrt{n_i}\left(\mathbb{F}_{ni} - F_i\right). \tag{2.7}$$

Hence if $\mathbf{F}$ is a *Donsker class* for all F_i, $i = 1, \dots, s$, there is a special construction of $\mathbf{X}_n^*$ and a (sequence of) Gaussian process(es) on a common probability space $(\Omega, \mathbf{A}, P)$ as in Dudley and Philipp (1983), satisfying

$$\|\mathbb{X}_n^* - \mathbb{X}^*\|_{\mathbf{F}} \equiv \sup_{f \in \mathbf{F}} |\mathbb{X}_n^*(f) - \mathbb{X}^*(f)| \to_p 0 \quad \text{as } n \to \infty, \tag{2.8}$$

where the mean zero Gaussian process $\mathbb{X}^*$ has covariance function

$$\begin{aligned} \operatorname{Cov}(\mathbb{X}^*(h_1), \mathbb{X}^*(h_2)) &= \sum_{i=1}^{s} \lambda_i \{F_i(h_1 h_2) - F_i(h_1) F_i(h_2)\} \\ &= \overline{F}(h_1 h_2) - \underline{F}(h_1)^T \underline{\lambda} \underline{F}(h_2) \qquad (2.9) \\ &= G\left(r^{-1} h_1 h_2\right) - G\left(h_1 \underline{\tilde{w}}^T\right) \underline{\lambda} G\left(h_2 \underline{\tilde{w}}\right) \qquad (2.10) \end{aligned}$$

and where $\tilde{w}_i \equiv w_i / W_i$, $i = 1, \dots, s$, and $r(x)^{-1} \equiv \sum_{i=1}^{s} \lambda_i \tilde{w}_i(x)$. This follows from the one-sample results of Dudley and Philipp (1983), or from other central limit theorems for the empirical process in the one-sample case such as Pollard (1982) or Ossiander (1987), by straightforward calculation.

The following additional notation will be used to state our limit theorems. We write, as in (2.10),

$$\tilde{w}_i(x) \equiv \frac{w_i(x)}{W_i}, \qquad i = 1, \dots, s, \ x \in \mathbf{X}, \tag{2.11}$$

and

$$r(x) \equiv \left[\sum_{i=1}^{s} \lambda_i \frac{w_i(x)}{W_i} \right]^{-1} = \left[\sum_{i=1}^{s} \lambda_i \tilde{w}_i(x) \right]^{-1}. \tag{2.12}$$

Note that r is the (Radon–Nikodym) derivative $dG/d\overline{F}$: $\int h\, dG = \int hr\, d\overline{F}$ for $h \in L_1(G)$. The following matrices and vectors occur frequently in the sequel. First, let

$$A \equiv \int r^2 \underline{\tilde{w}}\underline{\tilde{w}}^T\, d\overline{F} = \int r\underline{\tilde{w}}\underline{\tilde{w}}^T\, dG \qquad (s \times s) \tag{2.13}$$

[note that $w_i(x)r(x) \le W_i/\lambda_i$ implies that the integrals in (2.13) always exist] and let

$$M \equiv \underline{\underline{\lambda}}^{-1} - A \qquad (s \times s), \tag{2.14}$$

where, for a vector $\underline{u} \in R^s$, $\underline{\underline{u}}$ denotes the $s \times s$ diagonal matrix with entries u_i on the diagonal, $\underline{\underline{u}} \equiv \text{diag}\{u_i\}$. As will be seen in the proof [see (5.11) and (5.12)], $\underline{\underline{\lambda}} M \underline{\underline{\lambda}}$ is the matrix $D_n''(\underline{Z})$ of Section 1. Note that

$$M\underline{\lambda} = \underline{1} - \int r\underline{\tilde{w}} r^{-1}\, dG = \underline{1} - \int \underline{\tilde{w}}\, dG = \underline{1} - \underline{1} = \underline{0},$$

so M is singular. When the graph $\mathbf{G}$ is connected, it follows from the proof of Proposition 1.1 that M has rank $s - 1$; see Lemma 5.1. We let M^- denote any $\{1,2\}$-generalized inverse of M: i.e., $MM^-M = M$ and $M^-MM^- = M^-$; see Lemma 5.2 for the facts we need about generalized inverses and see Remark 5.1 for more on terminology and further references.

Asymptotic distributions. Now we can formulate our limit theorem for the *biased sampling empirical process*

$$\mathbb{Z}_n \equiv \sqrt{n}\,(\mathbb{G}_n - G), \tag{2.15}$$

regarded as a process indexed by a collection of functions $\mathbf{H} \subset L_2(G)$: Thus for $h \in \mathbf{H}$,

$$\mathbb{Z}_n(h) = \int h\, d\mathbb{Z}_n = \sqrt{n} \int h d(\mathbb{G}_n - G).$$

(Note that $\mathbb{Z}_n$ deserves this name because it plays the same role in the biased sampling context as does the usual empirical process in the context of random sampling.) At the end of the section we will give a corresponding theorem for the process

$$\mathbb{Z}_n^0 \equiv \sqrt{n}\left(\mathbb{G}_n^0 - G\right) \tag{2.16}$$

based on the estimator $\mathbb{G}_n^0$ in (1.24); recall that this estimator can be used only if the W_i's are *known*.

The appropriate (sequences of) limiting Gaussian processes $\mathbb{Z}$ and $\mathbb{Z}^0$ are described as follows: For h such that $h\sqrt{r} \in L_2(G)$, define $Vh \in L_2(G)$ by

$$Vh(x) \equiv r(x)h(x) + G\left(r\underline{\tilde{w}}^T h\right)M^- \underline{\tilde{w}}(x) r(x) \quad \text{for } x \in \mathbf{X}. \tag{2.17}$$

Then define the process $\mathbb{Z}$: $\mathbf{H} \to R^1$ by

$$\mathbb{Z}(h) = \mathbb{X}^*(V(h - G(h))), \tag{2.18}$$

where $\mathbb{X}^*$ is the mean zero Gaussian process given in (2.8) and (2.9). The representation (2.17) and (2.18) of $\mathbb{Z}$ is valid for any $\{1,2\}$-generalized inverse M^- of M; see Lemma 5.2. Note that the second term in (2.17) gives the contribution to the limit process $\mathbb{Z}$ from estimation of the W_i's. The process $\mathbb{Z}$ has covariance function

$$\begin{aligned} K(h_1, h_2) &\equiv \operatorname{Cov}[\mathbb{Z}(h_1), \mathbb{Z}(h_2)] \\ &= G(r[h_1 - G(h_1)][h_2 - G(h_2)]) \\ &\quad + G([h_1 - G(h_1)] r\underline{\tilde{w}}^T) M^- G([h_2 - G(h_2)] r\underline{\tilde{w}}). \end{aligned} \tag{2.19}$$

The covariance formula (2.19) will be proved in Section 5.

THEOREM 2.2 (CLT for $\mathbb{Z}_n$). *Suppose that the graph* $\mathbf{G}$ *is connected, that* $G(w_i) < \infty$, $i = 1, \dots, s$, *and* $\mathbf{H}$ *is a collection of functions with envelope function* h_e ($|h| \le h_e$ *for all* $h \in \mathbf{H}$) *such that* (2.8) *holds for the class of functions* $\mathbf{F} \equiv \{hr\colon h \in \mathbf{H}\}$ *and* $G(h_e^2 r) = \overline{F}(h_e^2 r^2) < \infty$. *Then, for the same special construction of* (2.8),

$$\|\mathbb{Z}_n - \mathbb{Z}\|_{\mathbf{H}} \equiv \sup_{h \in \mathbf{H}} |\mathbb{Z}_n(h) - \mathbb{Z}(h)| \to_p 0 \quad \text{as } n \to \infty, \tag{2.20}$$

where $\mathbb{Z}$ *is the* (*sequence of*) *mean zero Gaussian process*(*es*) (2.18) *with covariance function* (2.19).

REMARK 2.1. We have ignored measurability problems in our statement of Theorem 2.2. The theorem is true as stated if $\mathbf{H}$ is a countable collection, but for an uncountable collection $\mathbf{H}$ the supremum in (2.20) may not be measurable. This can be handled via measurable covering functions as in Dudley and Philipp (1983).

There are many sufficient conditions which imply that (2.8) holds for $\mathbf{F} \equiv \{hr\colon h \in \mathbf{H}\}$; see, e.g., Dudley and Philipp (1983), Pollard (1984) or Ossiander (1987). The important special case of $\mathbf{H} = \{1_C\colon C \in \mathbf{C}\}$, where $\mathbf{C}$ is a Vapnik–Chervonenkis class of subsets of $\mathbf{X}$, which follows as a consequence of Pollard's (1982) Theorem 9, is singled out in the following corollary.

COROLLARY 2.3. *Suppose that* $\mathbf{X} = R^d$ *and* $\mathbf{H}$ *is the collection of all indicator functions of lower left orthants, or of all rectangles, or of all balls, or the class of all half spaces. If* $G(r) = \overline{F}(r^2) < \infty$, *the graph* $\mathbf{G}$ *is connected, and* $G(w_i) < \infty$ *for* $i = 1, \dots, s$, *then* $\mathbb{Z}_n$ *satisfies* (2.20).

The proof of Theorem 2.2 involves first establishing asymptotic normality of $\sqrt{n}(\underline{\mathbb{W}}_n - \underline{W})$, but the cleanest expressions for the limiting random vector and its limiting covariance matrix result from the self-consistency equations (1.26)–(1.28) and Theorem 2.2 itself.

PROPOSITION 2.2 (Asymptotic normality of $\underline{\mathbb{W}}_n$). *If* $\mathbf{G}$ *is connected, then*

$$\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W}) \to_d K^{-}\mathbb{X}^*(r\underline{\tilde{w}}) + \underline{W}\mathbb{Z}_\alpha \tag{2.21}$$

$$= \mathbb{X}^*(r(\underline{w} - \underline{W})) + G\big(r(\underline{w} - \underline{W})\underline{\tilde{w}}^T\big)M^{-}\mathbb{X}^*(r\underline{\tilde{w}}) \tag{2.22}$$

$$\simeq N_s(0, \Sigma), \tag{2.23}$$

where

$$\Sigma \equiv G\big(r(\underline{w} - \underline{W})(\underline{w} - \underline{W})^T\big) + G\big(r(\underline{w} - \underline{W})\underline{\tilde{w}}^T\big)M^{-}G\big(r\underline{\tilde{w}}(\underline{w} - \underline{W})^T\big). \tag{2.24}$$

In (2.21), $K^{-} \equiv \underline{\underline{W}}\,\underline{\underline{\lambda}}^{-1}M^{-}$ *and* $\mathbb{Z}_\alpha$ *is a random variable which can be specified in terms of* $\mathbb{X}^*$ *later; see Remark* 2.4. *In* (2.24), M^{-} *is any* $\{1,2\}$-*generalized inverse of* M; *see Lemma* 5.2. *Since* M *is symmetric we take* M^{-} *symmetric too.*

REMARK 2.2. Note that the integrability hypothesis of Theorem 2.2 is automatically satisfied for w_i: Just as in the definition of the matrix A,

$$G\big(w_i^2 r\big) = \overline{F}\big(w_i^2 r^2\big) \le (W_i/\lambda_i)^2 < \infty$$

since $(\lambda_i w_i(x)/W_i)r(x) \le 1$.

REMARK 2.3. Note that the random variable in (2.22), the second way of expressing the limiting form of $\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W})$, is just $\mathbb{Z}(\underline{w})$, and that the formula (2.19) for its variance–covariance matrix is just $K(\underline{w}, \underline{w}^T)$ with K as defined in (2.19). These formulas result from Theorem 2.2 together with the self-consistency equations (1.26)–(1.28).

Asymptotic normality of $\sqrt{n}\,(\underline{\mathbb{V}}_n - \underline{V})$ can also be easily derived, even though this is usually not of primary interest and requires some additional notation. To state the result, we set

$$K \equiv M\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1} \qquad (s \times s), \tag{2.25}$$

$$K_\sharp \equiv J^T K J \qquad (s - 1 \times s - 1), \tag{2.26}$$

where J is the $s \times s - 1$ matrix

$$J \equiv \begin{pmatrix} I_{s-1 \times s-1} \\ 0 \end{pmatrix} \qquad (s \times s - 1). \tag{2.27}$$

Thus $K_\sharp \equiv J^T K J$ gives the upper left $s - 1 \times s - 1$ submatrix of K (for any $s \times s$ matrix K) and $u_\sharp \equiv J^T u$ gives the first $s - 1$ coordinates of $\underline{u} \in R^s$. We also set

$$C \equiv \operatorname{Cov}[\mathbb{X}^*(r\underline{\tilde{w}}), \mathbb{X}^*(r\underline{\tilde{w}})] = A - A\underline{\underline{\lambda}}A \qquad (s \times s) \tag{2.28}$$

and

$$C_\sharp \equiv J^T C J \qquad (s - 1 \times s - 1). \tag{2.29}$$

PROPOSITION 2.3 (Asymptotic normality of $\underline{\mathbb{V}}_n$). *If* $\mathbf{G}$ *is connected and* $G(w_i) < \infty$, $i = 1, \ldots, s$, *then*

$$\begin{aligned} J^T\sqrt{n}\,(\underline{\mathbb{V}}_n - \underline{V}) &\to_d W_s^{-1}K_{\#}^{-1}J^T\mathbb{X}^*(r\underline{\tilde{w}}) \\ &\simeq N_{s-1}\Big(0, K_{\#}^{-1}C_{\#}\big(K_{\#}^{-1}\big)^T/W_s^2\Big). \end{aligned} \tag{2.30}$$

Now consider estimation of the biased distributions $F_1, \ldots, F_s$. Since we know the nonparametric maximum likelihood estimate $\mathbb{G}_n$ of G, the nonparametric MLE $\hat{\mathbb{F}}_{ni}$ of an F_i is given by

$$\hat{\mathbb{F}}_{ni}(h) = \frac{\int h w_i \, d\mathbb{G}_n}{\int w_i \, d\mathbb{G}_n} = \frac{\mathbb{G}_n(h w_i)}{\mathbb{G}_n(w_i)} \tag{2.31}$$

for $h \in \mathbf{H}$, $i = 1, \ldots, s$. We let

$$\mathbb{Y}_{ni} \equiv \sqrt{n}\left(\hat{\mathbb{F}}_{ni} - F_i\right) \quad \text{for } i = 1, \ldots, s \tag{2.32}$$

and set

$$\underline{\mathbb{Y}}_n \equiv (\mathbb{Y}_{n1}, \ldots, \mathbb{Y}_{ns})^T,$$

regarded as a vector of processes indexed by a collection of functions $\mathbf{H} \subset L_2(\bar{F})$.

The appropriate (sequence of) limiting Gaussian processes $\underline{\mathbb{Y}}$ is described as follows: Define $\underline{\mathbb{Y}}\colon \mathbf{H} \to R^s$ by

$$\underline{\mathbb{Y}}(h) = \mathbb{X}^*(hr\underline{\tilde{w}}) - \left[G\left(h\underline{\tilde{w}}\right)\underline{\underline{\lambda}}^{-1} - G\left(hr\underline{\tilde{w}}\underline{\tilde{w}}^T\right)\right]M^{-}\mathbb{X}^*(r\underline{\tilde{w}}). \tag{2.33}$$

THEOREM 2.3 (CLT for $\underline{\mathbb{Y}}_n$). *Suppose that the graph* $\mathbf{G}$ *is connected, that* $G(w_i) < \infty$, $i = 1, \ldots, s$, *and* $\mathbf{H}$ *is a collection of functions with envelope* h_e *such that* (2.8) *holds for* $\mathbf{F} \equiv \bigcup_{i=1}^{s}\{hrw_i\colon h \in \mathbf{H}\}$ *and* $\bar{F}(h_e^2) < \infty$. *Then for the same special construction of* (2.8),

$$\max_{1 \le i \le s} \|\mathbb{Y}_{ni} - \mathbb{Y}_i\|_{\mathbf{H}} \to_p 0 \quad \text{as } n \to \infty, \tag{2.34}$$

where $\mathbb{Y}$ *is the* (*sequence of*) *vectors of mean zero Gaussian processes* (2.33).

REMARK 2.4. The two different expressions (2.26) and (2.27) for the limiting form of $\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W})$ are both of use. The first form (2.26) is the easiest to prove; it will be used in our proof of Theorem 2.2. The second form (2.27) connects the limit of $\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W})$ with the self-consistency equations (1.26)–(1.28) and our result for the process $\mathbb{Z}_n$ given in Theorem 2.2. It also yields the simplest expression (2.29) for the covariance matrix Σ; see Remark 2.3.

REMARK 2.5. Note that when $\mathbf{G}$ is connected, the asymptotic normal distribution of $\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W})$ is singular if and only if there exist constants $c_1, \ldots, c_s$ such that

$$\sum_{i=1}^{s} c_i w_i = \text{constant} \quad \text{a.e. } G. \tag{2.35}$$

For example, if $w_s \equiv 1$, then the limit distribution of $\sqrt{n}(\underline{\mathbb{W}}_n - \underline{W})$ is singular, but of course (2.35) can happen in other ways and the limit normal distribution is nonsingular whenever (2.35) fails.

REMARK 2.6. It is possible to express both the operator V in (2.22) and the covariance (2.24) of the limiting process $\mathbb{Z}$ in a less symmetric way which avoids introduction of the generalized inverse M^-. Because the resulting formulas are considerably more complicated, and because of our feeling that the generalized inverse M^- presents no real difficulty, we have not given them here. See the examples and applications in Section 4 and the facts about generalized inverses summarized in Lemma 5.2.

REMARK 2.7. The asymptotic covariance of the biased sampling process $\mathbb{Z}_n$ defined in (2.15) can be easily estimated. The sample analogue of (2.19) (with G and λ replaced throughout—i.e., also in M, $\underline{\tilde{w}}$ and r—by $\mathbb{G}_n$ and λ_{ni}) is easily shown to converge in probability or even a.s. to (2.19) as $n \to \infty$ under natural conditions. In fact, the covariance estimator is also obtained by formal likelihood calculations [inversion of the $(n-1) \times (n-1)$ matrix of second derivatives of the log likelihood] continuing the derivation of $\mathbb{G}_n$ itself as a maximum likelihood estimator in the model where G is discrete, with mass at the actual observations only.

Results for $\mathbb{G}_n^0$ and $\mathbb{Z}_n^0$. We conclude this section with a brief statement of some corresponding results for the estimator $\mathbb{G}_n^0$ which can be used when the W_i's are known (whether the graph $\mathbf{G}$ is connected or not).

THEOREM 2.4 (Consistency of $\mathbb{G}_n^0$). *Suppose that $G(w_i) = \int w_i \, dG < \infty$, $i = 1, \dots, s$, and that $\mathbf{H}$ is as defined in (2.3) with $G(h_e) < \infty$. Then (whether $\mathbf{G}$ is connected or not)*

$$\|\mathbb{G}_n^0 - G\|_{\mathbf{H}} \equiv \sup\left\{\left|\mathbb{G}_n^0(h) - G(h)\right| : h \in \mathbf{H}\right\} \to_{a.s.} 0 \tag{2.36}$$

as $n \to \infty$.

Now consider the process $\mathbb{Z}_n^0$ defined in (2.21). The appropriate (sequences of) limiting Gaussian process(es) are described as follows: For h such that $h\sqrt{r} \in L_2(G)$, define $V^0 h \in L_2(G)$ by

$$V^0 h(x) \equiv r(x) h(x). \tag{2.37}$$

Then define the process(es) $\mathbb{Z}^0 \colon \mathbf{H} \to R^1$ by

$$\mathbb{Z}^0(h) \equiv \mathbb{X}^*(V^0(h - G(h))) = \mathbb{X}^*(r(h - G(h))), \tag{2.38}$$

where $\mathbb{X}^*$ is the mean zero Gaussian process given in (2.8) and (2.9). The process $\mathbb{Z}^0$ has covariance function

$$\begin{aligned} K^0(h_1, h_2) &\equiv \operatorname{Cov}[\mathbb{Z}^0(h_1), \mathbb{Z}^0(h_2)] \\ &= G(r[h_1 - G(h_1)][h_2 - G(h_2)]) \\ &\quad - G([h_1 - G(h_1)] r \underline{\tilde{w}}^T) \underline{\underline{\lambda}} G([h_2 - G(h_2)] r \underline{\tilde{w}}). \end{aligned} \tag{2.39}$$

This covariance formula follows directly from the definition (2.38) and the covariance formula (2.10) for $\mathbb{X}^*$.

THEOREM 2.5 (CLT for $\mathbb{Z}_n^0$). *Suppose that* $G(w_i) < \infty$, $i = 1, \ldots, s$, *and that* $\mathbf{H}$ *is a collection of functions with envelope function* h_e ($|h| \le h_e$ *for all* $h \in \mathbf{H}$) *such that* (2.8) *holds for the class of functions* $\mathbf{F} \equiv \{hr\colon h \in \mathbf{H}\}$ *and* $G(h_e^2 r) = \bar{F}(h_e^2 r^2) < \infty$. *Then for the same special construction of* (2.8), *whether the graph* $\mathbf{G}$ *is connected or not,*

$$(2.40) \qquad \|\mathbb{Z}_n^0 - \mathbb{Z}^0\|_{\mathbf{H}} \equiv \sup_{h \in \mathbf{H}} \left|\mathbb{Z}_n^0(h) - \mathbb{Z}^0(h)\right| \to_p 0 \quad \text{as } n \to \infty,$$

where $\mathbb{Z}^0$ *is the (sequence of) mean zero Gaussian process(es)* (2.38) *with covariance function* K^0 *given by* (2.39).

3. Optimality of the nonparametric MLE $\mathbb{G}_n$. It is well known that the usual empirical df is an optimal estimator of the underlying df in the usual iid sampling situation; results of this type are due to Dvoretzky, Kiefer and Wolfowitz (1956), Kiefer and Wolfowitz (1959), Beran (1977), Levit (1978) and Millar (1979). For more recent results concerning the optimality of empirical measures more generally, see Millar (1985).

Despite recent progress on the optimality of nonparametric maximum likelihood estimates [see, e.g., Gill (1988), Bickel, Klaassen, Ritov and Wellner (1989) or van der Vaart (1988)], there does not yet exist a complete theory guaranteeing that they are in fact efficient estimates in general. Therefore our goal in this section is to give a convolution theorem for regular estimates of G which shows that Vardi's nonparametric maximum likelihood estimator for the biased sampling problem is, in fact, optimal. Our calculations follow the approach of Begun, Hall, Huang, and Wellner (1983).

Suppose that G has density g with respect to the (sigma-finite) measure ν on $\mathbf{X}$. We say that $\hat{G}_n$ is a *regular estimator* of G if, under $P_n \equiv P_{g_n}$ with $\sqrt{n}(g_n^{1/2} - g^{1/2}) \to \beta$ in $L_2(\nu)$, it follows that

$$(3.1) \qquad \sqrt{n}\left(\hat{G}_n - G\right) \Rightarrow \mathbb{S},$$

where the distribution of $\mathbb{S}$ does not depend on β.

Now let $V\colon L_2(G) \to L_2(G)$ be defined by (2.33), i.e.,

$$(3.2) \qquad Vh(x) \equiv r(x)h(x) + \langle rh, \underline{\tilde{w}}^T\rangle_G M^- \underline{\tilde{w}}(x) r(x)$$

for $x \in \mathbf{X}$, where $\langle\ ,\ \rangle_G$ denotes the usual inner product in $L_2(G)$. Also let K be the covariance function

$$(3.3) \qquad \begin{aligned} K(h_1, h_2) &\equiv \langle h_1 - G(h_1), V(h_2 - G(h_2))\rangle_G \\ &= G(r[h_1 - G(h_1)][h_2 - G(h_2)]) \\ &\quad + G([h_1 - G(h_1)]r\underline{\tilde{w}}^T)M^+ G([h_2 - G(h_2)]r\underline{\tilde{w}}) \end{aligned}$$

as in (2.28) with M^- a $\{1,2\}$-generalized inverse of M and let $\mathbb{Z}$ be a mean zero Gaussian process with covariance function K. By Section 2 we know that $\mathbb{Z}$ can be related to $\mathbb{X}^*$ by (2.24) and (2.26).

THEOREM 3.1 (Convolution theorem for regular estimates of G). *Suppose that the graph* $\mathbf{G}$ *is connected, that* r *and* $1/r$ *are bounded functions of* $x \in \mathbf{X}$ *and* $\lambda_{ni} \to \lambda_i > 0$ *for* $i = 1, \dots, s$. *Then the limit process* $\mathbb{S}$ *for any regular estimator of* G *in the biased sampling model* (1.21) *can be represented as*

$$\mathbb{S} =_d \mathbb{Z} + \mathbb{W}, \tag{3.4}$$

where $\mathbb{Z}$ *has the covariance function* K *given in* (3.3) *and the process* $\mathbb{W}$ *is independent of the process* $\mathbb{Z}$.

REMARK 3.1. Since the nonparametric MLE $\mathbb{G}_n$ given in (1.25) has limit process $\mathbb{Z}$ by Theorem 2.2, the assertion (3.4) of Theorem 3.1 is that for a *given, fixed* biased sampling problem as in (1.21), any regular estimator $\hat{\mathbb{G}}_n$ of G has a limit process $\mathbb{S}$ which is "at least as dispersed as" the limit process $\mathbb{Z}$ of $\mathbb{G}_n$ in the sense of (3.4). This assumes that the number of samples $s \geq 1$, the biasing functions $w_1, \dots, w_s$ and the sampling fractions $\lambda_1, \dots, \lambda_s$ are all fixed and given. If we have control of one or more of these elements of the problem, then we can consider *designing* the biased sampling in order to optimize some given criteria. (We will return briefly to questions of this kind in Section 6.) Theorem 3.1 says that once these design elements have been fixed, then the nonparametric MLE $\mathbb{G}_n$ is (asymptotically) optimal.

PROOF OF THEOREM 3.1. Since $G \ll \nu$ with density g, we can write

$$f_i = \frac{w_i g}{W_i} \equiv \tilde{w}_i g, \tag{a}$$

with

$$W_i = \int w_i g \, d\nu = G(w_i) = \langle w_i, 1 \rangle_G.$$

We also let

$$r = \left(\sum_{i=1}^{s} \lambda_i \tilde{w}_i \right)^{-1} = \left(\underline{\lambda}^T \underline{\tilde{w}} \right)^{-1} \tag{b}$$

and

$$f = \sum_{i=1}^{s} \lambda_i f_i = r^{-1} g. \tag{c}$$

It is easily verified that each f_i is Hellinger differentiable with respect to g with derivative

$$A_i \beta = f_i^{1/2} \left(\frac{\beta}{g^{1/2}} - \int \tilde{w}_i \beta g^{1/2} \, d\nu \right), \qquad i = 1, \dots, s, \tag{d}$$

so that

$$A_i^T A_i \beta = \frac{f_i \beta}{g} - \frac{f_i}{g^{1/2}} \int \beta f_i g^{-1/2} \, d\nu, \qquad i = 1, \dots, s, \tag{e}$$

and hence

$$
\begin{aligned}
U\beta &\equiv \sum_{i=1}^{s} \lambda_i A_i^T A_i \beta \\
&= \sum_{i=1}^{s} \lambda_i \tilde{w}_i \beta - \sum_{i=1}^{s} \lambda_i \tilde{w}_i g^{1/2} \langle \tilde{w}_i g^{1/2}, \beta \rangle_\nu \\
&= r^{-1}\beta - \underline{\tilde{w}}^T \underline{\lambda} g^{1/2} \langle \underline{\tilde{w}} g^{1/2}, \beta \rangle_\nu. \qquad \text{(f)}
\end{aligned}
$$

The operator U maps $L^2(\nu)$ to $L^2(\nu)$ (since r^{-1} is bounded) and is the s-sample analogue of the (one-sample) information operator A^TA of Begun, Hall, Huang and Wellner (1983). It is often convenient to work instead with U^* mapping $L^2(G)$ to $L^2(G)$ defined by

$$U^*\beta^* \equiv g^{-1/2} U(g^{1/2}\beta^*) \quad \text{for } \beta^* \in L^2(G).$$

Thus

$$U^*\beta^* = r^{-1}\beta^* - \underline{\tilde{w}}^T \underline{\lambda} \langle \underline{\tilde{w}}, \beta^* \rangle_G. \qquad \text{(g)}$$

To carry out the calculations of Begun, Hall, Huang and Wellner (1983) we need to calculate the inverse operator U^{-1} of U, or, equivalently, the inverse operator U^{*-1} of the operator U^*. This would be trivial without the second term in (g); to account for the second term, we must express $\langle \underline{\tilde{w}}, \beta^* \rangle$ in terms of $U^*\beta^*$. To do this we take inner products across (g) with the vector of functions $\underline{\tilde{w}}r$ to obtain

$$
\begin{aligned}
\langle \underline{\tilde{w}}, rU^*\beta^* \rangle_G &= \left(I - \langle \underline{\tilde{w}}, r\underline{\tilde{w}}^T \rangle_G \underline{\lambda}\right) \langle \underline{\tilde{w}}, \beta^* \rangle_G \\
&= \left(\underline{\lambda}^{-1} - \langle \underline{\tilde{w}}, r\underline{\tilde{w}}^T \rangle_G\right) \underline{\lambda} \langle \underline{\tilde{w}}, \beta^* \rangle_G \\
&= M \underline{\lambda} \langle \underline{\tilde{w}}, \beta^* \rangle_G, \qquad \text{(h)}
\end{aligned}
$$

where

$$M \equiv \underline{\lambda}^{-1} - \langle \underline{\tilde{w}}, r\underline{\tilde{w}}^T \rangle_G. \qquad \text{(i)}$$

Thus, by (5.17) of Lemma 5.2, (h) can be inverted to yield

$$\underline{\lambda} \langle \underline{\tilde{w}}, \beta^* \rangle_G = M^- \langle \underline{\tilde{w}}, rU^*\beta^* \rangle_G - \alpha \underline{\lambda}, \qquad \text{(j)}$$

where the constant $\alpha = \alpha(\beta^*)$ is still to be determined. Substitution of (j) into (g) yields

$$rU^*\beta^* \equiv \beta^* - r\underline{\tilde{w}}^T M^- \langle \underline{\tilde{w}}, rU^*\beta^* \rangle_G - \alpha, \qquad \text{(k)}$$

which implies

$$
\begin{aligned}
U^{*-1}\beta^* &= r\beta^* + r\underline{\tilde{w}}^T M^- \langle \underline{\tilde{w}}, r\beta^* \rangle_G + \alpha \qquad \text{(l)} \\
&\equiv V\beta^* + \alpha,
\end{aligned}
$$

where

$$V\beta^* = r\beta^* + r\underline{\tilde{w}}^T M^- \langle \underline{\tilde{w}}, r\beta^* \rangle_G. \qquad \text{(m)}$$

Now we want $U^{*-1}\beta^* \perp 1$ in $L^2(G)$; therefore

$$\alpha = \alpha(\beta^*) = -\langle V\beta^*, 1\rangle_G$$

and

$$\begin{aligned} U^{*-1}\beta^* &= V\beta^* - \langle V\beta^*, 1\rangle_G \\ &= V\beta^* - G(V\beta^*). \end{aligned} \tag{n}$$

Thus it follows that

$$\begin{aligned} K(h_1, h_2) &\equiv \langle g^{1/2}(h_1 - G(h_1)), U^{-1}g^{1/2}(h_2 - G(h_2))\rangle_\nu \\ &= \langle h_1 - G(h_1), U^{*-1}(h_2 - G(h_2))\rangle_G \\ &= \langle h_1 - G(h_1), V(h_1 - G(h_2)) - G(V(h_2 - G(h_2)))\rangle_G \\ &= \langle h_1 - G(h_1), V(h_2 - G(h_2))\rangle_G, \end{aligned} \tag{o}$$

where V is given in (m) and M is given in (i). The proof is completed by an application of (the s-sample version of) Theorem 4.1 of Begun, Hall, Huang and Wellner (1983). □

4. Examples and applications. The following examples illustrate the results obtained in Sections 1–3. As mentioned in Section 3, when considering examples it is useful to keep in mind the distinction between situations in which we have some control over the design elements of the biased sampling (choice of s, choice of λ_i's, choice of w_i's), and those situations in which the design is fixed and given. While the examples here are presented from the latter perspective, we may often, in fact, be interested in the former perspective (e.g., stratified sampling); see Problem 6.1 in Section 6 for a bit of this.

EXAMPLE 4.1 (Example 1.2 continued; ordinary and length-biased sampling). Let $\mathbf{X} = R^+ = (0, \infty)$ and suppose $0 < \mu \equiv \int x\, dG(x) < \infty$. Let $w_1(x) \equiv 1$ so that the first sample is from G itself and let $w_2(x) \equiv x$ so that the second sample is from the "length-biased" distribution $F_2(x) = \mu^{-1}\int_0^x y\, dG(y)$. Thus the graph $\mathbf{G}$ is easily seen to be connected (as discussed in Example 1.2). Furthermore $\underline{W} = (1, \mu)^T$ and letting $\lambda_1 \equiv \lambda \in (0,1)$, $\lambda_2 = 1 - \lambda \equiv \bar{\lambda}$,

$$r(x) = \left(\lambda + \bar{\lambda}\frac{x}{\mu}\right)^{-1}.$$

Hence

$$M = \begin{pmatrix} 1/\lambda - \bar{F}(r^2) & -\bar{F}(r^2\tilde{w}_2) \\ -\bar{F}(r^2\tilde{w}_2) & 1/\bar{\lambda} - \bar{F}(r^2\tilde{w}_2^2) \end{pmatrix} = \begin{pmatrix} \bar{\lambda}K/\lambda & -K \\ -K & \lambda K/\bar{\lambda} \end{pmatrix}$$

since $M\underline{\lambda} = \underline{0}$, where

$$K \equiv \bar{F}(r^2\tilde{w}_2) = G(r\tilde{w}_2) = \int_0^\infty \frac{x/\mu}{\lambda + \bar{\lambda}x/\mu}\, dG(x). \tag{4.1}$$

Also define

$$K(x) \equiv \int_0^x \tilde{w}_2 r\, dG = G(1_{[0,x]}\underline{\tilde{w}}_2 r) \tag{4.2}$$

and note that

$$G(1_{[0,x]}r) = \frac{1}{\lambda}(G(x) - \bar{\lambda}K(x)). \tag{4.3}$$

Then with $\mathbb{Z}(t) \equiv \mathbb{Z}(1_{[0,t]})$, $K(s,t) \equiv E[\mathbb{Z}(s)\mathbb{Z}(t)]$ and using the $\{1,2\}$-inverse (see Lemma 5.2)

$$M^{-} = \begin{pmatrix} \lambda/\bar{\lambda}K & 0 \\ 0 & 0 \end{pmatrix}$$

of M, it follows from (2.28) that

$$\begin{aligned} K(s,t) &= G\big(r(1_{[0,s]} - G(s))(1_{[0,t]} - G(t))\big) \\ &\quad + G\big(r(1_{[0,s]} - G(s))\big)\frac{\lambda}{\bar{\lambda}K}G\big(r(1_{[0,t]} - G(t))\big) \\ &= \frac{1}{\lambda}\{G(s \wedge t) - G(s)G(t)\} \\ &\quad - \frac{\bar{\lambda}}{\lambda}K\left\{\frac{K(s \wedge t)}{K} - \frac{K(s)}{K}\frac{K(t)}{K}\right\}, \end{aligned} \tag{4.4}$$

in agreement with equation (3.6) of Vardi (1982).

Moreover, by tedious but straightforward calculation from Proposition 2.2 the limit rv in (2.22) is

$$\mathbb{X}^*\begin{pmatrix} 0 \\ \dfrac{x}{\lambda K}r \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbb{X}^*\left(\dfrac{x}{\lambda K}r\right) \end{pmatrix}$$

and the covariance matrix $\Sigma \equiv (\sigma_{ij})$ in (2.24) is all zeros except for

$$\sigma_{22} \equiv \mathrm{Var}\left[\mathbb{X}^*\left(\frac{x}{\lambda K}r\right)\right] = \frac{\mu^2(1-K)}{\lambda\bar{\lambda}K}, \tag{4.5}$$

which agrees with (3.3) of Vardi (1982). An interesting robustness feature of this application of Proposition 2.2 is that only $\mu = \int x\,dG(x) < \infty$ was assumed and not $\int x^2\,dG(x) < \infty$. [Note that $K = G(r\tilde{w}_2)$ is always finite and in fact $0 < K < 1$ if G is not degenerate at 0 or μ; this is easily seen from the fact that $0 \le G(r(x-\mu)^2) = \mu^2(1-K)/\lambda\bar{\lambda}$.] Thus, in combined ordinary and length-biased sampling the nonparametric maximum likelihood estimator $\hat{\mu}_n = \mathbb{W}_{n2} = \int x\,d\mathbb{G}_n(x)$ of the mean $\mu = W = \int x\,dG(x)$ of G is asymptotically normal with asymptotic variance as in (4.5) even if $\int x^2\,dG(x) = \infty$.

EXAMPLE 4.2 ($s = 1$; biasing with only one stratum). For a general sample space $\mathbf{X}$ suppose that $s = 1$ and that $w_1 = w$ satisfies both $G(w) < \infty$ and $W_{-1} \equiv G(w^{-1}) < \infty$. Then the graph $\mathbf{G}$ is trivially connected and, letting $W \equiv G(w)$, $r = W/w = \tilde{w}^{-1}$, $\bar{F}(r) = G(\mathbf{X}^+)$ may be equal to 1 or < 1 (if assumption S fails) depending on w and G. Let $G^+ \equiv G(\cdot \cap \mathbf{X}^+)/G(\mathbf{X}^+)$. Since $M = 0$ and $r\tilde{w} = 1$ (an extension of) Proposition 2.2 [to this case of $\bar{F}(r) < 1$]

yields

$$\sqrt{n}(\mathbb{W}_n - W) \to_d -\mathbb{X}^*(r/\overline{F}^2(r))W \tag{4.6}$$
$$\simeq N\left(0, \frac{W^2}{\overline{F}(r)^4}(W_{-1}W - 1)\right).$$

Furthermore, the process $\mathbb{Z}^*$ of Theorem 2.2 reduces to

$$\mathbb{Z}^*(h) = \mathbb{X}^*(r(h - G^+(h))/\overline{F}(r)) \tag{4.7}$$
$$= \mathbb{X}^*(r(h - G(h))), \quad \text{when } \overline{F}(r) = 1,$$

with covariance function

$$\begin{aligned} K(h_1, h_2) &= G^+(r(h_1 - G^+(h_1))(h_2 - G^+(h_2)))/\overline{F}(r) \\ &= G(r(h_1 - G(h_1))(h_2 - G(h_2))) \quad [\text{when } \overline{F}(r) = 1] \\ &= W_{-1}W\left\{\left(\frac{G(h_1h_2/w)}{W_{-1}} - \frac{G(h_1)G(h_2/w)}{W_{-1}}\right)\right. \\ &\qquad \left. + \left(G(h_1h_2) - \frac{G(h_1/w)G(h_2)}{W_{-1}}\right)\right\}. \end{aligned} \tag{4.8}$$

These formulas agree with the results of Vardi [(1985a), Section 7(ii)] in the case $\overline{F}(r) = 1$. The further special case $w = x$ (which is also the special case $\lambda = 0$ in Example 1) was considered by Cox (1969).

EXAMPLE 4.3 (Truncated sampling or restricted measurement). This is a further special case of Example 4.2. For a general sample space $\mathbf{X}$ suppose that $s = 1$ and that $w_1(x) = 1_C(x)$ where $C \in \mathbf{B}$, $C \neq \mathbf{X}$ and $G(C) < 1$. Then $\mathbf{X}^+ = C$, $W_1 = G(C) < 1$, $r(x) = G(C)1_C(x)$ and $G^+ = G/G(C)$ is simply the conditional distribution $G(\cdot\,|C)$. Thus

$$G^+(A) = \frac{G(A \cap C)}{G(C)} = \overline{F}(A) \quad \text{for } A \in \mathbf{B} \cap C,$$

and the estimator $\mathbb{G}_n^+$ of G^+ is simply $\mathbb{F}_n$. Note that $W_1 \equiv G(C)$ is not identifiable in this situation.

EXAMPLE 4.4 (Choice-based sampling in econometrics; case-control studies in biostatistics). Suppose that $\mathbf{X} = (Y, Z)$ where Y takes values in $\{1, \dots, M\}$ and $Z \simeq H$ with density h with respect to μ is a covariate vector with values in $\mathbf{Z} \subset$ some R^p. The basic (unbiased or prospective) model G has density

$$g(y, z) = p_\theta(y|z)h(z) \tag{4.9}$$

so that

$$G(\{y\} \times A) = \int_A p_\theta(y|z)\, dH(z) \tag{4.10}$$

for $y = 1, \dots, M$ and $A \in \mathbf{B}(R^p)$, where $p_\theta(y|z) \equiv P_\theta(Y = y|Z = z)$ is a parametric (finite-dimensional) model. A frequent choice is the logistic regression model

$$(4.11) \qquad p_\theta(y|z) = \frac{\exp\left(\alpha_y + \beta_y^T z\right)}{\sum_{y'=1}^M \exp\left(\alpha_{y'} + \beta_{y'}^T z\right)}$$

with $\theta = (\alpha, \beta) \in R^{(p+1)M}$.

The biased (or retrospective) sampling model F is obtained from G via the weight functions $w_i(x) = w_i(y) = 1_{D_i}(y)$ where $D_i \subset \{1, \dots, M\}$ for $i = 1, \dots, s$. This again yields a semiparametric submodel, since only distributions G of the form (4.10) are considered.

One case of particular interest is that of "pure choice-based sampling" in the terminology of Cosslett (1981). In this sampling scheme, the strata D_i are taken to be just $D_i = \{i\}$, $i = 1, \dots, s \equiv M$. In this case the graph $\mathbf{G}$ of Section 1 is *not* connected: In fact $G(w_i w_j) = 0$ for all $i \neq j$ and hence there is no unique nonparametric MLE $\mathbb{G}_n$ of G for this sampling scheme. Manski and Lerman (1977) avoid this difficulty of pure choice-based sampling by assuming that the "aggregate shares"

$$G(y) \equiv G(\{y\} \times \mathbf{Z}) = \int p_\theta(y|z)\, dH(z), \qquad y = 1, \dots, M,$$

are known. Note that for this biasing system we can view F as a biased distribution derived from H with new biasing (weight) functions $w_y^\#(z; \theta) \equiv p_\theta(y|z)$, $y = 1, \dots, M$, depending on the unknown parameter θ. Then $W_y^\# = G(y)$, typically the graph $\mathbf{G}^\#$ for these $w^\#$'s will be connected and, if θ is known, the methods of the preceding sections yield estimates of H together with the asymptotic behavior of the estimates.

This same pure choice-based sampling design is also frequently used in case-control studies in biostatistics where the y's often denote different disease categories. In the biostatistics applications interest centers on odds ratios which can be estimated from purely choice-based sampling in spite of the fact that G itself cannot be estimated; see, e.g., Prentice and Pyke (1979), who examine the case of (4.11), and Breslow and Day (1980). If the "pure choice-based" design is enriched by taking $s = M + 1$, $\lambda_M + 1 = 0$ and choosing $w_{M+1}(x) = 1_{\{1, \dots, M\}}(y)$, then the graph $\mathbf{G}$ is connected and the nonparametric MLE $\mathbb{G}_n$ of G exists (a.s. for $n \geq$ some N_ω) and is unique. See Example 4.5.

For general D_i's the biased distribution F has density

$$(4.12) \qquad f(y, z, i) = \lambda_i \frac{1_{D_i}(y) p_\theta(y|z) h(z)}{\int \sum_{y'=1}^M 1_{D_i}(y') p_\theta(y'|z') h(z')\, d\mu(z')}$$

and the connectedness assumption C for existence of a unique solution is precisely Cosslett's (1981) Assumption 10:

$$(4.13) \qquad \left\{ \bigcup_{i \in B} D_i \right\} \cap \left\{ \bigcup_{i \in B^c} D_i \right\} \neq \varnothing$$

for every proper subset B of $\{1, \dots, s\}$;

see Vardi [(1985a), Sections 2 and 8].

For known θ, efficient estimates of H and their asymptotic behavior via the preceding sections can be obtained as follows: The marginal distribution of (Z, I) is

$$(4.14)\qquad \begin{aligned} f(z,i) &= \lambda_i \frac{\{\sum_{y'=1}^{M} 1_{D_i}(y') p_\theta(y'|z)\} h(z)}{\int \{\sum_{y'=1}^{M} 1_{D_i}(y') p_\theta(y'|z')\} h(z')\, d\mu(z')} \\ &\equiv \lambda_i \frac{w_i^\#(z) h(z)}{\int w_i^\#(z')\, dH(z')}, \end{aligned}$$

where the new biasing functions $w_i^\#(z) = w_i^\#(z;\theta)$ depend on θ. Thus if θ is known, the methods of Vardi (1985a) and the preceding sections apply to yield efficient estimates of H, which can in turn be used to construct efficient estimates of θ. This method is implicit in Cosslett [(1981), Section 4] and will be discussed in more detail in Bickel, Klaassen, Ritov and Wellner (1989). See also Morgenthaler and Vardi (1986) for a nonparametric discussion of this problem.

EXAMPLE 4.5 (Example 1.3b continued; enriched stratified sampling). Let $\mathbf{X}$ be a general sample space and suppose that $D_1, \dots, D_s$ form a (measurable) partition of $\mathbf{X}$: $D_i \cap D_j = \varnothing$ for $i \neq j$ and $\bigcup_{i=1}^{s} D_i = \mathbf{X}$. Let $w_i(x) = 1_{D_i}(x)$ for $i = 1, \dots, s$ and suppose that $w_{s+1}(x) = 1$. Thus the stratified sample from $D_1, \dots, D_s$ is enriched by sampling from all of $\mathbf{X}$ with sampling fraction $\lambda_{s+1} > 0$; this terminology is that of Cosslett (1981).

For this sampling scheme the graph $\mathbf{G}$ is connected as discussed in Example 1.3b [assuming without loss of generality that $G(D_i) > 0$, $i = 1, \dots, s$], $\overline{F}(r) = 1$ and we have

$$W_i = G(D_i), \qquad i = 1, \dots, s,$$

$$r(x) = \sum_{i=1}^{s} \frac{1}{\lambda_{s+1} + \lambda_i / G(D_i)} 1_{D_i}(x),$$

so that the upper left $s \times s$ submatrix of $\overline{F}(r^2 \underline{\tilde{w}}\underline{\tilde{w}}^T)$ is diagonal with elements

$$\overline{F}(r\tilde{w}_i^2) = \frac{1}{\lambda_{s+1} G(D_i) + \lambda_i}.$$

Hence the upper left $s \times s$ submatrix of M is diagonal with elements

$$M_{ii} = \frac{1}{\lambda_i} - \frac{1}{\lambda_{s+1} G(D_i) + \lambda_i} = \frac{\lambda_{s+1} G(D_i)}{\lambda_i(\lambda_{s+1} G(D_i) + \lambda_i)},$$

for $i = 1, \dots, s$, and a $\{1,2\}$-inverse M^- of M is given by the diagonal matrix with the last row and column containing all zeros and having diagonal elements M_{ii}^{-1}; see Lemma 5.2. Thus, $K^- = \underline{\underline{W}}\underline{\underline{\lambda}}^{-1} M^-$ is also diagonal with the last row and column all zero and diagonal entries

$$(4.15)\qquad K_{ii}^- = \frac{1}{\lambda_{s+1}} \{\lambda_{s+1} G(D_i) + \lambda_i\} \equiv \frac{1}{\lambda_{s+1}} a_i$$

for $i = 1, \dots, s$. Similarly,

$$G(r\tilde{w}_i) = G(D_i)/a_i$$

and hence $M^{-}G(r\underline{\tilde{w}}) = (\lambda_1/\lambda_{s+1}, \dots, \lambda_s/\lambda_{s+1}, 0)$. Therefore

$$\left(1 + \underline{\tilde{w}}^T M^{-} G(r\underline{\tilde{w}})\right)\underline{W}r = \left(1 + \sum_{j=1}^{s} \frac{1_{D_i}}{G(D_i)} \frac{\lambda_i}{\lambda_{s+1}}\right) r\underline{W}$$
$$= \frac{1}{\lambda_{s+1}} r^{-1} r\underline{W} = \frac{1}{\lambda_{s+1}} \underline{W},$$

which is constant in x, and hence the limiting random vector in Proposition 2.2 becomes $\mathbb{K}^{-}\mathbb{X}^*(r\underline{\tilde{w}})$, the last element of which is 0 by the form of K^{-}, and where the first s elements of $\mathbb{X}^*(r\underline{\tilde{w}})$ have an $s \times s$ covariance matrix

$$\begin{aligned} C &= \underline{a} - \left(\underline{a}, \underline{a}G(\underline{D})\right)\underline{\lambda}\left(\underline{a}, \underline{a}G(\underline{D})\right)^T \\ &= \underline{a}\left(\underline{a}^{-1} - \left(I, \underline{p}\right)\underline{\lambda}\left(I, \underline{p}\right)^T\right)\underline{a}, \end{aligned} \tag{4.16}$$

where $\underline{p} \equiv G(\underline{D}) \equiv (G(D_1), \dots, G(D_s))^T$. Thus by Proposition 2.2 and straightforward calculation using (4.15) and (4.16), it follows that

$$\sqrt{n}\left(\mathbb{G}_n(\underline{D}) - G(\underline{D})\right) \to_d N_s\left(0, \frac{1}{\lambda_{s+1}}\left(G\left(\underline{\underline{D}}\right) - G(\underline{D})G(\underline{D})^T\right)\right).$$

Alternatively, from the discussion of this example in Section 1 (Example 1.3b), we have $\mathbb{V}_{ni} = \mathbb{W}_{ni} = F_{n, s+1}(D_i)$ for $i = 1, \dots, s$, and hence, since $F_{s+1} = G$,

$$\begin{aligned} \sqrt{n}\left(\mathbb{G}_n(\underline{D}) - G(\underline{D})\right) &= \sqrt{n}\left(\mathbb{F}_{n, s+1}(\underline{D}) - F_{s+1}(\underline{D})\right) \\ &= \sqrt{\frac{n}{n_{s+1}}}\sqrt{n_{s+1}}\left(\mathbb{F}_{n, s+1}(\underline{D}) - F_{s+1}(\underline{D})\right) \\ &\to_d N_s\left(0, \frac{1}{\lambda_{s+1}}\left(G\left(\underline{\underline{D}}\right) - G(\underline{D})G(\underline{D})^T\right)\right), \end{aligned}$$

in complete agreement. Note that this is just the covariance matrix for the usual (multinomial) estimate of $G(\underline{D})$ from a random sample of size $n\lambda_{s+1}$ from all of $\mathbf{X}$. In other words, sampling *within* the strata D_i does not help in estimating the strata probabilities $G(\underline{D})$.

EXAMPLE 4.6 (Stratified or truncated regression). This interesting and rich family of semiparametric submodels of the general biased sampling model begins with ordinary linear regression with unknown error distribution G_0 as the basic (unbiased) model: Suppose that $X = (Y, Z) \simeq G$, where $Y = \theta^T Z + \varepsilon$ with $\varepsilon \simeq G_0$ with density g_0 with respect to Lebesgue measure and $Z \simeq H$ independent of ε with density h with respect to μ. Thus G has density

$$g(y, z) = g_0\left(y - \theta^T z\right)h(z).$$

The biased sampling model is typically determined by weight functions $w_i(x) \equiv w_i(y) \equiv 1_{D_i}(y)$, $i = 1, \dots, s$, where the D_i's are disjoint subintervals of R^1. The case of $s = 1$ and $D_1 = (-\infty, y_0]$, which is also a special case of

Example 3, has been considered by Bhattacharya, Chernoff and Yang (1983). Jewell (1985) considers the case $s = 2$ and $D_1 = (-\infty, y_0]$, $D_2 = (y_0, \infty)$ in which the connectedness assumption C fails, so a unique completely nonparametric estimator of G does not exist in view of Vardi's Theorem 1.1. Nevertheless the parameters θ, G_0 and H are identifiable in this model and for known θ the methods of Vardi (1985a) can be applied iteratively by first regarding G_0 as known and absorbing it into the biasing functions and estimating H, and then by treating H as known and absorbing it into the biasing functions and estimating G_0 and so forth. This type of semiparametric submodel of the biased sampling model will be treated by Bickel, Klaassen, Ritov and Wellner (1989).

5. Proofs.

Notation and basic lemmas. We now introduce the additional notation used in the proofs of the results stated in Section 2. If $\underline{w} \equiv (w_1, \dots, w_s)^T$ is the vector of biasing functions and

$$\underline{u} \in R^{+s} = \{\underline{u} \in R^s\colon u_i > 0,\ i = 1, \dots, s\},$$

let

$$\tilde{\underline{w}}(\underline{u}) \equiv \underline{\underline{u}}^{-1}\underline{w}, \qquad r(\underline{u}) \equiv \left(\underline{\lambda}^T\tilde{\underline{w}}(\underline{u})\right)^{-1}, \qquad r_n(\underline{u}) \equiv \left(\underline{\lambda}_n^T\tilde{\underline{w}}(\underline{u})\right)^{-1}. \tag{5.1}$$

Thus $\tilde{\underline{w}}$, r and r_n of Section 2 are given by

$$\tilde{\underline{w}} \equiv \tilde{\underline{w}}(\underline{W}) = \underline{\underline{W}}^{-1}\underline{w}, \qquad r \equiv \left(\underline{\lambda}^T\tilde{\underline{w}}\right)^{-1}, \qquad r_n \equiv \left(\underline{\lambda}_n^T\tilde{\underline{w}}\right)^{-1}. \tag{5.2}$$

Also, $\bar{F} = \sum_{i=1}^s \lambda_i F_i$ and $\bar{F}_n = \sum_{i=1}^s \lambda_{ni} F_i$ are related to G by

$$d\bar{F} = \left(\underline{\lambda}^T\tilde{\underline{w}}\right) dG = r^{-1}\, dG \tag{5.3}$$

and

$$d\bar{F}_n = \left(\underline{\lambda}_n^T\tilde{\underline{w}}\right) dG = r_n^{-1}\, dG \tag{5.4}$$

or

$$G(A) = \int_A r(x)\, d\bar{F}(x) \quad \text{for } A \in \mathbf{B} \tag{5.5}$$

and

$$G(A) = \int_A r_n(x)\, d\bar{F}_n(x) \quad \text{for } A \in \mathbf{B}. \tag{5.6}$$

We also define

$$\begin{aligned} \underline{H}(\underline{u}) &\equiv \int r(\underline{u})\tilde{\underline{w}}(\underline{u})\, d\bar{F} = \bar{F}((r\tilde{\underline{w}})(\underline{u})), \\ \underline{H}_n(\underline{u}) &\equiv \int r_n(\underline{u})\tilde{\underline{w}}(\underline{u})\, d\bar{F}_n = \bar{F}_n((r_n\tilde{\underline{w}})(\underline{u})), \\ \underline{\mathbb{H}}_n(\underline{u}) &\equiv \int r_n(\underline{u})\tilde{\underline{w}}(\underline{u})\, d\mathbb{F}_n = \mathbb{F}_n((r_n\tilde{\underline{w}})(\underline{u})) \end{aligned} \tag{5.7}$$

and similarly,

$$
\begin{aligned}
D(\underset{=}{z}) &= \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(y)\right] d\bar{F}(y) - \sum_{i=1}^{s} \lambda_i z_i, \\
D_n(\underset{=}{z}) &= \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(y)\right] d\bar{F}_n(y) - \sum_{i=1}^{s} \lambda_{ni} z_i, \qquad (5.8)\\
\mathbb{D}_n(\underset{=}{z}) &= \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(y)\right] d\mathbb{F}_n(y) - \sum_{i=1}^{s} \lambda_{ni} z_i.
\end{aligned}
$$

Note that $\underline{H}(\underline{u})$ and $\underline{H}_n(\underline{u})$ are homogeneous of degree 0 in $\underline{u}$: $\underline{H}(c\underline{u}) = \underline{H}(\underline{u})$ for all $c > 0$. Also note that $\underline{H}(\underline{W}) = \underline{H}_n(\underline{W}) = \underline{1}$ and in fact $\underline{H}(c\underline{W}) = \underline{1}$ for all $c > 0$ by homogeneity.

The derivative matrix $\underline{\nabla}\underline{H}^T$ of the matrix of functions $\underline{H}(\underline{u})$ defined in (5.7) plays an important role in the arguments which follow. By straightforward differentiation under the integral sign (easily justified by monotone convergence)

$$
\begin{aligned}
(\nabla H)(\underline{u}) &\equiv \left(\underline{\nabla}\underline{H}^T(\underline{u})\right)^T \\
&= -\left\{\underset{=}{H}(\underline{u})\underset{=}{\lambda}^{-1} - \bar{F}\left(r^2(\underline{u})\underline{\tilde{w}}(\underline{u})\underline{\tilde{w}}^T(\underline{u})\right)\right\}\underset{=}{\lambda}\underset{=}{u}^{-1}. \qquad (5.9)
\end{aligned}
$$

Hence for $\underline{u} = c\underline{W}$ with $c > 0$, since $\underline{H}(c\underline{W}) = \underline{1}$,

$$
\begin{aligned}
\nabla H(c\underline{W}) &= -c^{-1}\left\{\underset{=}{\lambda}^{-1} - \bar{F}\left(r^2\underline{\tilde{w}}\underline{\tilde{w}}^T\right)\right\}\underset{=}{\lambda}\,\underset{=}{W}^{-1} \\
&\equiv -c^{-1}M\underset{=}{\lambda}\,\underset{=}{W}^{-1} \qquad (5.10)\\
&\equiv -c^{-1}K
\end{aligned}
$$

with A, M and K as in (2.13), (2.14) and (2.15).

Of course ∇H is closely related to (the limit of) the matrix D_n'' of (f) of the Proof of Proposition 1.1. To make the connection, we let $D''(\underline{z})$ denote $D_n''(\underline{z})$ with λ_{ni} replaced by λ_i and $\bar{F}_n$ replaced by $\bar{F}$: Thus

$$
D''(\underline{z})_{ij} = \int \left\{\frac{e^{z_i} w_i(y)\delta_{ij}}{B(y)} - \frac{e^{z_i} w_i(y) e^{z_j} w_j(y)}{B^2(y)}\right\} d\bar{F}(y),
$$

where $B(y) \equiv \sum_{j=1}^{s} e^{z_j} w_j(y)$ and $z_i = \log(\lambda_i/u_i)$, $i = 1, \dots, s$. By easy calculation it follows that

$$
D''(\underline{z}) = -\underset{=}{\lambda}(\nabla H(\underline{u}))\underset{=}{u}, \qquad (5.11)
$$

and, in particular, with $Z_i \equiv \log(\lambda_i/V_i)$,

$$
D''(\underline{Z}) = \underset{=}{\lambda}\left\{\underset{=}{\lambda}^{-1} - A\right\}\underset{=}{\lambda} = \underset{=}{\lambda} M \underset{=}{\lambda} \equiv D''. \qquad (5.12)
$$

The following lemma is therefore a consequence of our proof of Proposition 1.1.

Lemma 5.1. *If the graph* G *is connected, then the matrix M defined in* (5.9) *and* (2.14) *has rank $s - 1$. In fact, every principal proper submatrix of M is nonsingular and the same holds for $(\nabla H)(\underline{u})$. In particular, if the ith row and* *column are deleted, the resulting* $(s-1)\times(s-1)$ *matrix is of full rank.*

Note that

$$(\nabla H(\underline{u}))\cdot\underline{u}=\underline{0}, \tag{5.13}$$

$$M\underline{\lambda}=\underline{0}, \qquad K\underline{W}=0, \tag{5.14}$$

and

$$D''\underline{1}=\underline{0}. \tag{5.15}$$

Since M is not of full rank, it does not have an inverse and we must use a generalized inverse. The following lemma summarizes the facts about generalized inverses of M which we will need, using the terminology of Berman and Plemmons (1979). For connections with terminology of Rao (1973) or Rao and Mitra (1971), see Remark 5.1.

LEMMA 5.2. (i) *If the graph* $\mathbf{G}$ *is connected, then the matrix* M *has a* $\{1,2\}$*-generalized inverse* M^-; *thus* M^- *satisfies*

$$MM^-M=M \quad \text{and} \quad M^-MM^-=M^-. \tag{5.16}$$

(ii) *For any such* $\{1,2\}$*-inverse* M^-,

$$\underline{y}=M\underline{x} \quad \text{implies} \quad \underline{x}=M^-\underline{y}+c\underline{\lambda} \quad \text{for some } c, \tag{5.17}$$

where $\underline{\lambda}$ *is the unique eigenvector of* M *with eigenvalue* 0.

(iii) *If, in addition,* M^- *is the* $\{1,2,3,4\}$ *or Moore–Penrose generalized inverse of* M [*which satisfies* (5.16) *and* $MM^-=(MM^-)^T$, $M^-M=(M^-M)^T$], *then*

$$MM^-=1-\underline{\theta}\underline{\theta}^T, \tag{5.18}$$

where $\underline{\theta}\equiv\underline{\lambda}/|\underline{\lambda}|$.

(iv) $\underline{a}^TM^-\underline{b}$ *is independent of the choice of generalized inverse* M^- *for all* $\underline{a},\underline{b}\in\text{Range}(M)=\{\underline{x}\colon \lambda^T\underline{x}=0\}$.

REMARK 5.1. A $\{1\}$-inverse of M in the terminology of Berman and Plemmons (1979) is a matrix M^- satisfying $MM^-M=M$. This is a g-inverse in the terminology of Rao and Mitra (1971) or Rao (1973). A $\{1,2\}$-inverse of M in the terminology of Berman and Plemmons (1979) also satisfies $M^-MM^-=M^-$; this is what Rao and Mitra (1971) and Rao (1973) call a reflexive g-inverse. See Rao [(1973), pages 24–27] and Rao and Mitra [(1971), Lemma 2.2.4, Theorem 2.2.1 and Lemma 2.5.1].

PROOF OF LEMMA 5.2. (i) By Lemma 5.1, deleting any row and column, in particular the last row and column, from M yields an $(s-1)\times(s-1)$ matrix M_{11} of rank $s-1$. Thus a $\{1,2\}$-inverse M^- of M is given by

$$M^-=\begin{pmatrix} M_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}; \tag{a}$$

see, e.g., Seber [(1977), page 76] and Berman and Plemmons [(1979), page 117].

(iii) Since M is symmetric, another way to get a $\{1,2\}$-inverse M^- satisfying (5.16) is to use the decomposition $M = P\underline{\underline{d}}P^T$ where the columns of P are the normalized eigenvectors of M and $\underline{\underline{d}}$ is the diagonal matrix of the eigenvalues $d_1 \geq d_2 \geq \cdots \geq d_{s-1} > d_s = 0$ of M. Then

$$\text{(b)} \qquad M^- = P\underline{\underline{d}}^- P^T,$$

where $\underline{\underline{d}}^-$ is the diagonal matrix with $\underline{\underline{d}}_{ii}^- = 1/d_i$ for $i = 1, \ldots, s-1$ and $\underline{\underline{d}}_{ss}^- = 0$ is in fact the Moore–Penrose (or $\{1,2,3,4\}$) generalized inverse of M; see Berman and Plemmons [(1979), page 117]. The proof of (5.18) proceeds by direct computation using $M = P\underline{\underline{d}}P^T$ and (b).

(ii) Now for any $\{1,2\}$-inverse M^- [an inverse satisfying (5.16)], M^-M is the projector on $R(M^-)$ along $N(M)$; see Berman and Plemmons [(1979), page 118]. Thus if $\underline{x} = \underline{x}_0 + c\underline{\lambda}$ and $\underline{y} = M\underline{x}$, it follows that $M^-\underline{y} = M^-M\underline{x} = \underline{x}_0 = \underline{x} - c\underline{\lambda}$ and hence (5.17) holds.

(iv) Suppose that $\underline{a}, \underline{b} \in \text{Range}(M)$ so that $\underline{a} = M\underline{x}$ and $\underline{b} = M\underline{y}$ for some $\underline{x}$, $\underline{y}$. Then it follows immediately from (5.16) that

$$\underline{a}^T M^- \underline{b} = \underline{x}^T M M^- M \underline{y} = \underline{x}^T M \underline{y},$$

which is clearly independent of the choice of M^-. □

Consistency proofs. Our consistency proofs depend on the uniqueness of $\underline{\mathbb{V}}_n$ and $\underline{V}$ as solutions of the systems of equations (1.16) and (1.30) together with convexity of D_n and D as established in the Proof of Proposition 1.1. With $D_n(\underline{z})$ as defined in (1.31) and (5.8), set

$$\begin{aligned}\tilde{\mathbb{D}}_n(\underline{z}) &\equiv \mathbb{D}_n(\underline{z}) - \mathbb{D}_n(\underline{Z}) \\ &= \int \log\left\{\frac{\sum_{i=1}^s e^{z_i} w_i(x)}{\sum_{i=1}^s e^{Z_i} w_i(x)}\right\} d\mathbb{F}_n(x) - \sum_{i=1}^s \lambda_{ni}(z_i - Z_i)\end{aligned} \tag{5.19}$$

and similarly

$$\tilde{D}(\underline{z}) \equiv D(\underline{z}) - D(\underline{Z}), \tag{5.20}$$

where $\underline{Z} \equiv (Z_1, \ldots, Z_s)$, $Z_i \equiv \log(\lambda_i/V_i)$. Note that a minimizer of $D_n(\underline{z})$ is a minimizer of $\tilde{D}_n(\underline{z})$ and conversely, and since $D_n(\underline{z})$ is a convex function of $\underline{z}$, so is $\tilde{D}_n$. Subtraction of the term $D_n(\underline{Z})$ yields a function $\tilde{D}_n$ which converges a.s. to the corresponding population function $\tilde{D}(\underline{z})$ under no additional hypotheses on the w_i's.

LEMMA 5.3. *The function $\tilde{D}_n(\underline{z})$ defined in* (5.19) *satisfies*

$$\tilde{\mathbb{D}}_n(\underline{z}) \to_{a.s.} \tilde{D}(\underline{z}) \quad \text{as } n \to \infty \tag{5.21}$$

for each fixed $\underline{z} \in R^s$. Since $\tilde{D}_n(\underline{z})$ is convex, it follows that

$$\sup_{\underline{z} \in C} \left|\tilde{\mathbb{D}}_n(\underline{z}) - \tilde{D}(\underline{z})\right| \to_{a.s.} 0 \quad \text{as } n \to \infty \tag{5.22}$$

for any compact subset $C \subset R^s$.

PROOF. Let

$$q(\underline{z}, x) \equiv \log\left\{\frac{\sum_{i=1}^{s} e^{z_i} w_i(x)}{\sum_{i=1}^{s} e^{Z_i} w_i(x)}\right\} \tag{a}$$

denote the function in the integral in (5.19). Since $q(\underline{z}, x)$ is a bounded function of x, it follows from the strong law of large numbers and $\lambda_{ni} \to \lambda_i$ that

$$\begin{aligned}
\tilde{\mathbb{D}}_n(\underline{z}) &= \int q(\underline{z}, x)\, d\mathbb{F}_n(x) - \sum_{i=1}^{s} \lambda_{ni}(z_i - \mathbb{Z}_i) \\
&= \sum_{j=1}^{s} \lambda_{nj} \int q(\underline{z}, x)\, d\mathbb{F}_{nj}(x) - \sum_{i=1}^{s} \lambda_{ni}(z_i - Z_i) \\
&\to_{\text{a.s.}} \sum_{j=1}^{s} \lambda_j \int q(\underline{z}, x)\, dF_j(x) - \sum_{i=1}^{s} \lambda_i(z_i - Z_i) \\
&= \tilde{D}(\underline{z}).
\end{aligned}$$

Thus (5.21) holds.

But $D_n(\underline{z})$ is a convex function of $\underline{z}$ by Remark 1.1; recall the Proof of Proposition 1.1. Therefore $\tilde{D}_n(\underline{z})$ is also a convex function of $\underline{z}$. Thus (5.22) follows from (5.21) and the convexity by Theorem 10.8 of Rockafellar (1970). □

PROOF OF PROPOSITION 2.1. If the graph $\mathbf{G}$ is connected, then by Proposition 1.1, $\underline{V} = \underline{W}/W_s$ is the unique solution of (1.30). Hence $\underline{Z} = (Z_1, \dots, Z_s)$ given by $Z_i = \log(\lambda_i/V_i)$ is the unique minimizer of $D(\underline{z})$, and hence also of $\tilde{D}(\underline{z})$, subject to $z_s = \log(\lambda_s)$. Now let $\underline{\mathbb{Z}}_n$ denote a minimizer of $D_n(\underline{z})$ and hence also a minimizer of $\tilde{D}_n(\underline{z})$ subject to $z_s = \log(\lambda_{ns})$. Then for any compact set C with $\underline{Z} \in C^0$, the interior of C, it follows from Lemma 5.3 and the definition of $\tilde{D}(\underline{z})$ and $\underline{Z}$ that

$$\inf_{z \in \partial C} \tilde{\mathbb{D}}_n(\underline{z}) \to_{\text{a.s.}} \inf_{z \in \partial C} \tilde{D}(\underline{z}) > 0$$

while

$$\tilde{\mathbb{D}}_n(\underline{Z}) \to_{\text{a.s.}} \tilde{D}(\underline{Z}) = 0.$$

Since $\tilde{\mathbb{D}}_n(\underline{z})$ is convex, it follows that $\underline{\mathbb{Z}}_n \in C$ for $n \geq$ some N_ω with probability 1, and since C can be made arbitrarily small, this implies [as in Appendix II of Andersen and Gill (1982)] that $\underline{\mathbb{Z}}_n \to_{\text{a.s.}} \underline{Z}$. But $\underline{\mathbb{Z}}_n = (\mathbb{Z}_{n1}, \dots, \mathbb{Z}_{ns-1}, \log \lambda_{ns})$ with $\mathbb{Z}_{ni} = \log(\lambda_{ni}/\mathbb{V}_{ni})$, where $\underline{\mathbb{V}}_n = (\mathbb{V}_{n1}, \dots, \mathbb{V}_{ns})$ is the solution of (1.16) (which exists and is unique for $n \geq$ some n_ω with probability 1) by an easy calculation. Hence $\mathbb{V}_n \to_{\text{a.s.}} \underline{V} = \underline{W}/W_s$, i.e., (2.1) holds.

Now (2.1) implies that $\mathbb{W}_{ns}$ given by (1.19) converges a.s. to

$$\frac{1}{\int_{\mathbf{X}} \left[\sum_{i=1}^{s} \lambda_i (w_i(y)/V_i)\right]^{-1} d\bar{F}(y)} = \frac{W_s}{\int_{\mathbf{X}} r\, d\bar{F}} = W_s$$

by (5.5) and assumption S. Hence

$$\mathbb{W}_{ni} = \mathbb{V}_{ni}\mathbb{W}_{ns} \to_{\text{a.s.}} V_i W_s = W_i \quad \text{as } n \to \infty,$$

for $i = 1, \dots, s$, so (2.2) holds. □

PROOF OF THEOREMS 2.1 AND 2.4. For a fixed function h we write

$$\text{(a)} \qquad \begin{aligned} &|\mathbb{F}_n(hr_n(\underline{\mathbb{V}}_n)) - \bar{F}(hr)| \\ &\leq |\mathbb{F}_n(h(r_n(\underline{\mathbb{V}}_n) - r))| + |\mathbb{F}_n(hr) - \bar{F}(hr)| \\ &\leq \left\| \frac{r_n(\mathbb{V}_n)}{r} - 1 \right\|_\infty |\mathbb{F}_n(hr)| + |\mathbb{F}_n(hr) - \bar{F}(hr)| \\ &\to_{\text{a.s.}} 0 \quad \text{as } n \to \infty \end{aligned}$$

for any fixed function h with $\bar{F}(hr) = G(h) < \infty$ by the strong law of large numbers, and by continuity and boundedness of $r(u)/r$ (as a function of x) together with $\underline{\mathbb{V}}_n \to_{\text{a.s.}} \underline{V}$ by Proposition 2.1. Thus by Pollard's Glivenko–Cantelli theorem [see Dudley (1984), Theorems 11.1.2 and 11.1.6] or $\mathbf{H}$ as in (2.3),

$$\text{(b)} \qquad \begin{aligned} &\sup_{h\in\mathbf{H}} |\mathbb{F}_n(hr_n(\underline{\mathbb{V}}_n)) - \bar{F}(hr)| \\ &\leq \left\| \frac{r_n(\mathbb{V}_n)}{r} - 1 \right\|_\infty \sup_{h\in\mathbf{H}} |\mathbb{F}_n(hr)| + \sup_{h\in\mathbf{H}} |\mathbb{F}_n(hr) - \bar{F}(hr)| \\ &\leq \left\| \frac{r_n(\mathbb{V}_n)}{r} - 1 \right\|_\infty \mathbb{F}_n(h_e r) + \sup_{h\in\mathbf{H}} |\mathbb{F}_n(hr) - \bar{F}(hr)| \\ &\to_{\text{a.s.}} 0 \cdot \bar{F}(h_e r) + 0 = 0. \end{aligned}$$

But

$$\begin{aligned} |\mathbb{G}_n(h) - G(h)| &= \left| \frac{\mathbb{F}_n(hr_n(\mathbb{V}_n))}{\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n))} - \frac{\bar{F}(hr)}{\bar{F}(r)} \right| \\ &\leq \frac{|\mathbb{F}_n(hr_n(\mathbb{V}_n)) - \bar{F}(hr)|}{\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n))} + \frac{|\bar{F}(hr)|}{\bar{F}(r)} \frac{|\mathbb{F}_n(r_n(\mathbb{V}_n)) - \bar{F}(r)|}{\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n))}, \end{aligned}$$

so (2.4) follows from (a) with $h = 1$ and (b).

The proof of consistency of $\mathbb{G}_n^0$ stated in (2.36) is similar, but does not use Proposition 2.1 since $\mathbb{G}_n^0$ depends only on the W_i's and not the $\mathbb{W}_{ni}$'s; we therefore omit it. □

Asymptotic normality proofs. The first step in our proofs is to expand $\underline{\mathbb{H}}_n(\underline{\mathbb{V}}_n)$ or $\underline{\mathbb{H}}_n(\underline{\mathbb{W}}_n)$ about $\underline{V}$ or $\underline{W}$, respectively.

PROOF OF PROPOSITION 2.3. Let $\underline{1} = (1, \dots, 1)^T \in R^{+s}$ and recall the definition (2.17) of the $s \times s - 1$ matrix J. Since

$$\begin{aligned} J^T\underline{1} &= J^T\underline{\mathbb{H}}_n(\underline{\mathbb{V}}_n) = J^T\underline{H}_n(\underline{V}) \\ \text{(a)} \qquad &= J^T\{\underline{\mathbb{H}}_n(\underline{V}) + \nabla\mathbb{H}_n(\underline{\mathbb{V}}_n^*)(\underline{\mathbb{V}}_n - \underline{V})\}, \end{aligned}$$

where $\underline{\mathbb{V}}_n^*$ lies on the line segment between $\underline{\mathbb{V}}_n$ and $\underline{V}$ (with a different $\mathbb{V}_n^*$ for each of the $s-1$ equations), it follows that

$$\text{(b)} \quad -J^T \nabla \mathbb{H}_n(\underline{\mathbb{V}}_n^*)\sqrt{n}(\underline{\mathbb{V}}_n - \underline{V}) = \sqrt{n}(\underline{\mathbb{H}}_n(\underline{V}) - \underline{H}_n(\underline{V})) = \mathbb{X}_n^*(r_n\underline{\tilde{w}}).$$

Hence, letting $n \to \infty$ and writing $\underline{\mathbb{Z}}_V \equiv \lim_n \sqrt{n}(\underline{\mathbb{V}}_n - \underline{V})$, it follows from continuity of $\nabla H(\underline{u})$ and (2.1) that

$$\text{(c)} \qquad \mathbb{X}^*(r\underline{\tilde{w}}) = -\nabla H(\underline{V})\underline{\mathbb{Z}}_V = W_s K_\# J^T \underline{\mathbb{Z}}_V$$

or, since $K_\#^{-1}$ exists by Lemma 5.1,

$$J^T \underline{\mathbb{Z}}_V = W_s^{-1} K_\#^{-1} J^T \mathbb{X}^*(r\underline{\tilde{w}})$$

$$\text{(d)} \qquad \simeq N_{s-1}\Big(0, K_\#^{-1} C_\# \big(K_\#^{-1}\big)^T / W_s^2\Big),$$

by a straightforward covariance calculation and by the definition of the matrices C and $C_\#$ given in (2.28) and (2.29). This completes the proof of (2.30). □

Proof of Proposition 2.2. To prove (2.21), we again expand, but now we work with the equations (1.27) involving all s coordinates. Since

$$\underline{1} = \underline{\mathbb{H}}_n(\underline{\mathbb{W}}_n) = \underline{H}_n(\underline{W})$$

$$\text{(a)} \qquad = \underline{\mathbb{H}}_n(\underline{W}) + \nabla \mathbb{H}_n(\underline{\mathbb{W}}_n^*)(\underline{\mathbb{W}}_n - \underline{W}),$$

where $\underline{\mathbb{W}}_n^*$ lies on the line segment between $\underline{\mathbb{W}}_n$ and $\underline{W}$ (actually there is a different $\underline{\mathbb{W}}_n^*$ for each of the s equations), it follows that

$$\text{(b)} \quad -\nabla \mathbb{H}_n(\underline{\mathbb{W}}_n^*)\sqrt{n}(\underline{\mathbb{W}}_n - \underline{W}) = \sqrt{n}(\underline{\mathbb{H}}_n(\underline{W}) - \underline{H}_n(\underline{W})) = \mathbb{X}_n^*(r_n\underline{\tilde{w}})$$

and hence, letting $n \to \infty$ and writing $\underline{\mathbb{Z}}_W \equiv \lim_n \sqrt{n}(\underline{\mathbb{W}}_n - \underline{W})$, it follows from continuity of $\nabla H(\underline{u})$ and (2.2) that

$$\text{(c)} \qquad \mathbb{X}^*(r\underline{\tilde{w}}) = -\nabla H(\underline{W})\underline{\mathbb{Z}}_W = M\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1}\underline{\mathbb{Z}}_W$$

or, by (5.17) in Lemma 5.2,

$$\underline{\mathbb{Z}}_W = \underline{\underline{W}}\underline{\underline{\lambda}}^{-1} M^- \mathbb{X}^*(r\underline{\tilde{w}}) + \underline{\underline{W}}\underline{\underline{\lambda}}^{-1} \mathbb{Z}_\alpha \underline{\lambda}$$

$$\text{(d)} \qquad = K^- \mathbb{X}^*(r\underline{\tilde{w}}) + \mathbb{Z}_\alpha \underline{W}.$$

This completes the proof of (2.21). We postpone identification of $\mathbb{Z}_\alpha$ as given in (2.22) since this step requires the convergence of $\mathbb{Z}_n$ which will be established in Theorem 2.2. In proving Theorem 2.2. we will use only (2.21) and *not* (2.22). □

Proof of Theorems 2.2 and 2.5. First consider $\mathbb{Z}_n(h)$ for fixed h with $\bar{F}(h^2 r^2) = G(h^2 r) < \infty$. Since $\mathbb{Z}_n(1) = 0$, we may assume without loss of generality that $G(h) = \int h\,dG = 0$. Also, note that by (1.19), $\mathbb{W}_{ni} = \mathbb{V}_{ni}\mathbb{W}_{ns}$ and

Assumption S, $\mathbb{F}_n(r_n(\underline{\mathbb{W}}_n)) = \bar{F}_n(r_n) = \bar{F}(r) = 1$. Hence

$$\text{(a)}\quad \mathbb{Z}_n(h) \equiv \sqrt{n}\,(\mathbb{G}_n(h) - G(h))$$

$$= \sqrt{n}\left\{\frac{\mathbb{F}_n(hr_n(\mathbb{W}_n))}{\mathbb{F}_n(r_n(\underline{\mathbb{W}}_n))} - \frac{\bar{F}_n(hr_n(W))}{\bar{F}_n(r_n(\underline{W}))}\right\}$$

$$= \sqrt{n}\left\{\mathbb{F}_n(hr_n(\underline{\mathbb{W}}_n)) - \bar{F}_n(hr_n(\underline{W}))\right\}$$

$$= \sqrt{n}\left\{\mathbb{F}_n(hr_n(\underline{W})) - \bar{F}_n(hr_n(\underline{W}))\right\} + \sqrt{n}\,\mathbb{F}_n\left[h(r_n(\underline{\mathbb{W}}_n) - r_n(\underline{W}))\right]$$

$$= \mathbb{X}_n^*(hr_n) + \sqrt{n}\,\mathbb{F}_n\left(hr_nr_n(\underline{\mathbb{W}}_n)\sum_{i=1}^{s}\frac{\lambda_{ni}}{W_i\mathbb{W}_{ni}}(\mathbb{W}_{ni} - W_i)\right)$$

$$\text{(b)}\qquad = \mathbb{X}_n^*(hr_n) + \mathbb{F}_n\left(hr_nr_n(\underline{\mathbb{W}}_n)\underline{\tilde{w}}^T\right)\underline{\underline{\lambda}}_n\underline{\underline{\mathbb{W}}}_n^{-1}\sqrt{n}\,(\underline{\mathbb{W}}_n - W)$$

$$\text{(c)}\qquad \to_p \mathbb{X}^*(hr) + \bar{F}\left(hr^2\underline{\tilde{w}}^T\right)\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1}\mathbb{Z}_W$$

[by (2.8), (2.2) and Theorem 2.1]

$$\text{(d)}\qquad = \mathbb{X}^*(hr) + G\left(hr\underline{\tilde{w}}^T\right)\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1}\mathbb{Z}_W.$$

But by (2.21) of Proposition 2.2 and $K \equiv M\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1}$ so that $K^- = \underline{\underline{W}}\,\underline{\underline{\lambda}}^{-1}M^-$, the second term in (d) equals

$$G\left(hr\underline{\tilde{w}}^T\right)\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1}\left\{\underline{\underline{W}}\,\underline{\underline{\lambda}}^{-1}M^-\mathbb{X}^*(r\underline{\tilde{w}}) + \underline{W}\mathbb{Z}_\alpha\right\}$$

$$= G\left(hr\underline{\tilde{w}}^T\right)M^-\mathbb{X}^*(r\underline{\tilde{w}}) + G\left(hr\underline{\tilde{w}}^T\right)\underline{\lambda}\mathbb{Z}_\alpha$$

$$\text{(e)}\qquad = G\left(hr\underline{\tilde{w}}^T\right)M^-\mathbb{X}^*(r\underline{\tilde{w}}),$$

since $G(hr\underline{\tilde{w}}^T)\underline{\lambda} = G(hrr^{-1}) = G(h) = 0$. Combining (d) and (e) yields

$$\text{(f)}\qquad \mathbb{Z}_n(h) \to_p \mathbb{X}^*(hr) + G\left(hr\underline{\tilde{w}}^T\right)M^-\mathbb{X}^*(r\underline{\tilde{w}})$$

$$= \mathbb{X}^*\left(rh + G\left(hr\underline{\tilde{w}}^T\right)M^-r\underline{\tilde{w}}\right)$$

$$= \mathbb{X}^*(Vh) \equiv \mathbb{Z}^*(h)$$

with Vh as defined in (2.17).

It remains only to establish that the convergence in (b) holds uniformly in $h \in \mathbf{H}$. By comparison of (b) and (c) and by the asymptotic normality of $\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W})$ established in Proposition 2.2, it clearly suffices to show

$$\text{(g)}\qquad \sup_{h\in\mathbf{H}}\left|\mathbb{X}_n^*(hr_n) - \mathbb{X}^*(hr)\right| \to_p 0$$

and

$$\text{(h)}\qquad \sup_{h\in\mathbf{H}}\left|\mathbb{F}_n\left(hr_nr_n(\underline{\mathbb{W}}_n)\underline{\tilde{w}}^T\right)\underline{\underline{\lambda}}_n\underline{\underline{\mathbb{W}}}_n^{-1} - \bar{F}\left(hr^2\underline{\tilde{w}}^T\right)\underline{\underline{\lambda}}\,\underline{\underline{W}}^{-1}\right| \to_p 0.$$

Now, since $\|\underline{\mathbb{W}}_n - \underline{W}\| \to_{\text{a.s.}} 0$ by Proposition 2.1 and $\bar{F}(h_e^2r^2) < \infty$, (h) follows from a Glivenko–Cantelli theorem for $\mathbb{F}_n$ as in the Proof of Theorem 2.1.

To prove (g), note that the left side is bounded by

(i) $$\sup_{h \in \mathbf{H}} |\mathbb{X}_n^*(hr_n) - \mathbb{X}^*(hr_n)| + \sup_{h \in \mathbf{H}} |\mathbb{X}^*(hr_n) - \mathbb{X}^*(hr)| \equiv \mathrm{I} + \mathrm{II}.$$

Then, since

$$\mathrm{II} = \sup_{h \in \mathbf{H}} |\mathbb{X}^*(h(r_n - r))|,$$

where

(j) $$\|h(r_n - r)\|_{L_2(\bar{F})} \le \left\| \frac{r_n}{r} - 1 \right\|_\infty \bar{F}(h_e^2 r^2)^{1/2} \to 0$$

uniformly in $h \in \mathbf{H}$, II $\to_p 0$ by uniform continuity of $\mathbb{X}^*$. To handle I, note that

$$\begin{aligned}
\mathrm{I} &\le \sup_{h \in \mathbf{H}} |\mathbb{X}_n^*(hr_n) - \mathbb{X}^*(hr_n)| \\
&\le \sup_{h \in \mathbf{H}} \left|\mathbb{X}_n^*\left(hr\left(\frac{r_n}{r} - 1\right)\right)\right| + \sup_{h \in \mathbf{H}} |\mathbb{X}_n^*(hr) - \mathbb{X}^*(hr)| \\
&\quad + \sup_{h \in \mathbf{H}} \left|\mathbb{X}^*\left(hr\left(1 - \frac{r_n}{r}\right)\right)\right| \\
&\le \left\| \frac{r_n}{r} - 1 \right\|_\infty \sup_{h \in \mathbf{H}} |\mathbb{X}_n^*(hr)| + \|\mathbb{X}_n^*(\cdot r) - \mathbb{X}^*(\cdot r)\|_{\mathbf{H}} \\
&\quad + \left\| \frac{r_n}{r} - 1 \right\|_\infty \sup_{h \in \mathbf{H}} |\mathbb{X}^*(hr)|
\end{aligned}$$

(k) $$\to_p 0 \quad \text{as } n \to \infty$$

by our hypothesis that (2.8) holds for $\mathbf{F} \equiv \{hr\colon h \in \mathbf{H}\}$. Hence (g) holds and this completes the proof of (2.30).

The proof of (2.40) is similar and we omit it.

Now we establish the covariance formula (2.19). First suppose that $G(h_1) = G(h_2) = 0$. Now from the covariance of $\mathbb{X}^*$ given in (2.10) and the definition of V in (2.17), it follows that

(l) $$\mathrm{Cov}(\mathbb{Z}(h_1), \mathbb{Z}(h_2)) = G(r^{-1}V(h_1)V(h_2)) - G(\underline{\tilde{w}}^T V(h_i))\underline{\lambda} G(\underline{\tilde{w}} V(h_2)).$$

But

$$\begin{aligned}
G(\underline{\tilde{w}} V(h_i)) &= \langle \underline{\tilde{w}}, rh_i \rangle_G + \langle \underline{\tilde{w}}, r\underline{\tilde{w}}^T \rangle_G M^- G(h_i r \underline{\tilde{w}}) \\
\text{(m)} \qquad &= (\mathrm{I} + \langle \underline{\tilde{w}}, r\underline{\tilde{w}}^T \rangle_G M^-) G(h_i r \underline{\tilde{w}}).
\end{aligned}$$

Now if M^- is the Moore–Penrose generalized inverse M^+ of M, it follows from (5.18) of Lemma 5.3 that $MM^+ = \mathrm{I} - \underline{\theta}\underline{\theta}^T$, where $\underline{\theta}$ is the (only) normalized eigenvector of M with eigenvalue 0. Since $M\underline{\lambda} = \underline{0}$, it follows that

$$\text{(n)} \qquad MM^{+} = \left(\underline{\underline{\lambda}}^{-1} - \langle \underline{\tilde{w}}, r\underline{\tilde{w}}^{T}\rangle_G\right)M^{+} = \mathrm{I} - \underline{\lambda}\underline{\lambda}^{T}/\underline{\lambda}^{T}\underline{\lambda}$$

and hence, with M^- taken to be M^+ in the definition (2.17) of V,

$$G(\underline{\tilde{w}}V(h_i)) = \left(\underline{\underline{\lambda}}^{-1}M^{+} + \underline{\lambda}\underline{\lambda}^{T}/\underline{\lambda}^{T}\underline{\lambda}\right)G(h_i r\underline{\tilde{w}})$$

$$\text{(o)} \qquad = \underline{\underline{\lambda}}^{-1}M^{+}G(h_i r\underline{\tilde{w}})$$

since $\underline{\lambda}^{T}G(h_i r\underline{\tilde{w}}) = G(h_i r r^{-1}) = G(h_i) = 0$. Also,

$$G\left(r^{-1}V(h_1)V(h_2)\right) = G\left(\left(h_1 + G\left(h_1 r\underline{\tilde{w}}^{T}\right)M^{+}\underline{\tilde{w}}^{T}\right)V(h_2)\right)$$

$$\text{(p)} \qquad = G(h_1V(h_2)) + G\left(h_1 r\underline{\tilde{w}}^{T}\right)M^{+}G(\underline{\tilde{w}}V(h_2)).$$

Thus

$$\begin{aligned}\mathrm{Cov}(\mathbb{Z}(h_1), \mathbb{Z}(h_2)) &= G(h_1V(h_2)) + G\left(r\underline{\tilde{w}}^{T}h_1\right)M^{+}G(\underline{\tilde{w}}V(h_2)) \\ &\quad - G\left(h_1 r\underline{\tilde{w}}^{T}\right)M^{+}\underline{\underline{\lambda}}^{-1}\underline{\underline{\lambda}}G(\underline{\tilde{w}}V(h_2)) \\ \text{(q)} \qquad &= G(h_1V(h_2)).\end{aligned}$$

Now for arbitrary h, since $\mathbb{Z}(1) = 0$,

$$\mathbb{Z}(h) = \mathbb{Z}(h - G(h)) = \mathbb{X}^{*}(V(h - G(h))),$$

so from (q) it follows that

$$\mathrm{Cov}(\mathbb{Z}(h_1), \mathbb{Z}(h_2)) = \langle h_1 - G(h_1), V(h_2 - G(h_2))\rangle_G,$$

and hence (2.19) holds with M^- taken to be the Moore–Penrose inverse M^+. But now note that

$$a_i \equiv G\left((h_i - G(h_i))r\underline{\tilde{w}}\right) \in \mathrm{Range}(M) = \left\{\underline{x}\colon \underline{\lambda}^{T}\underline{x} = 0\right\}$$

for $i = 1, 2$ and hence by Lemma 5.2(iv), the second term in (2.19) is the same for any $\{1, 2\}$-generalized inverse M^-. Thus (2.19) holds. □

Proof of (2.22) of Proposition 2.2 (Identification of $\mathbb{Z}_\alpha$). From the self-consistency equation (1.28) we have $\underline{\mathbb{W}}_n = \underline{\mathbb{G}}_n(\underline{w})$, and since $\underline{W} = G(\underline{w})$ by definition, it follows that

$$\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W}) = \mathbb{Z}_n(\underline{w}),$$

where $\max_{1 \le i \le s}\|w_i r\|_\infty \le \max_{1 \le i \le s}(W_i/\lambda_i) < \infty$. Thus, on the one hand by

Theorem 2.2

$$\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W}) \to_p \mathbb{Z}(\underline{w})$$

$$\text{(a)} \qquad = \mathbb{X}^*\big(r(\underline{w} - \underline{W})\big) + G\big((\underline{w} - \underline{W})r\underline{\tilde{w}}^T\big)M^-\mathbb{X}^*(r\underline{\tilde{w}}),$$

which proves (2.22). On the other hand, by (2.21),

$$\text{(b)} \qquad \sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W}) \to_p K^-\mathbb{X}^*(r\underline{\tilde{w}}) + \mathbb{Z}_\alpha \underline{W}.$$

Thus the expressions on the right sides in (a) and (b) must be equal, and upon multiplying across by $\underline{\lambda}^T \underline{\underline{W}}^{-1}$ and using $\underline{\lambda}^T \underline{\underline{W}}^{-1}\underline{W} = 1$, this yields

$$\mathbb{X}^*\big(r(r^{-1} - 1)\big) + G\big(r(r^{-1} - 1)\underline{\tilde{w}}^T\big)M^-\mathbb{X}^*(r\underline{\tilde{w}})$$
$$= \underline{1}^T M^-\mathbb{X}^*(r\underline{\tilde{w}}) + \mathbb{Z}_\alpha$$

or, since $\mathbb{X}^*(1) = 0$,

$$-\mathbb{X}^*(r) + \big(\underline{1}^T - G(r\underline{\tilde{w}}^T)\big)M^-\mathbb{X}^*(r\underline{\tilde{w}}) = \underline{1}^T M^-\mathbb{X}^*(r\underline{\tilde{w}}) + \mathbb{Z}_\alpha,$$

so that

$$\mathbb{Z}_\alpha = \big(-\mathbb{X}^*(r) + \{\underline{1}^T - G(r\underline{\tilde{w}}^T)\}M^-\mathbb{X}^*(r\underline{\tilde{w}})\big) - \underline{1}^T M^-\mathbb{X}^*(r\underline{\tilde{w}})$$

$$\text{(c)} \qquad = \big(-\mathbb{X}^*(r) - G(r\underline{\tilde{w}}^T)M^-\mathbb{X}^*(r\underline{\tilde{w}})\big).$$

Substitution of (c) into (b) yields yet another limiting form for $\sqrt{n}\,(\underline{\mathbb{W}}_n - \underline{W})$. □

Proof of Theorem 2.3. From the definitions (2.31) and (2.32) we can write, for fixed $h \in \mathbf{H}$,

$$
\begin{aligned}
\text{(a)} \quad \mathbb{Y}_{ni}(h) &= \sqrt{n}\big(\hat{\mathbb{F}}_{ni}(h) - F_i(h)\big) \\
&= \sqrt{n}\,(\mathbb{G}_n(w_i h) - G(w_i h))/\mathbb{W}_{ni} + G(w_i h)(1/\mathbb{W}_{ni} - 1/W_i) \\
&= \sqrt{n}\,(\mathbb{G}_n(\tilde{w}_i h) - G(\tilde{w}_i h)) \\
&\quad + G(\tilde{w}_i h)W_i^{-1}\sqrt{n}\,(\mathbb{W}_{ni} - W_i) + o_p(1) \\
&= \mathbb{Z}_n(\tilde{w}_i h) - G(\tilde{w}_i h)W_i^{-1}\sqrt{n}\,(\mathbb{W}_{ni} - W_i) + o_p(1) \\
&= \mathbb{X}^*\big(r(\tilde{w}_i h - G(\tilde{w}_i h))\big) \\
&\quad + G\big(r\underline{\tilde{w}}^T(\tilde{w}_i h - G(\tilde{w}_i h))\big)M^-\mathbb{X}^*(r\underline{\tilde{w}}) \\
\text{(b)} \quad &\quad - G(\tilde{w}_i h)W_i^{-1}\{\mathbb{X}^*(r(w_i - W_i)) \\
&\qquad + G\big(r(w_i - W_i)\underline{\tilde{w}}^T\big)M^-\mathbb{X}^*(r\underline{\tilde{w}})\} + o_p(1)
\end{aligned}
$$

by Theorems 2.2 and Proposition 2.2 since $|h| \le h_e$ and $\overline{F}(h_e^2) < \infty$ implies

$\bar{F}(h_e^2 \tilde{w}_i^2 r^2) \le \lambda_i^{-2} \bar{F}(h_e^2) < \infty$. Therefore

$$
\begin{aligned}
\text{(c)}\quad \underline{\mathbb{Y}}_n(h) &= \mathbb{X}^*(r(\underline{\tilde{w}}h - G(\underline{\tilde{w}}h))) \\
&\quad + G\big(r(\underline{\tilde{w}}h - G(\underline{\tilde{w}}h))\underline{\tilde{w}}^T\big)M^- \mathbb{X}^*(r\underline{\tilde{w}}) \\
&\quad - G\big(h\underline{\tilde{w}}\big)\underline{\underline{W}}^{-1}\{\mathbb{X}^*(r(\underline{w} - \underline{W})) \\
&\qquad + G\big(r(\underline{w} - \underline{W})\underline{\tilde{w}}^T\big)M^- \mathbb{X}^*(r\underline{\tilde{w}})\} + o_p(1) \\
&= \mathbb{X}^*(hr\underline{\tilde{w}}) + \Big[\big\{G\big(hr\underline{\tilde{w}}\underline{\tilde{w}}^T\big) - G\big(h\underline{\tilde{w}}\big)G\big(r\underline{\tilde{w}}\underline{\tilde{w}}^T\big)\big\}M^- - G\big(h\underline{\underline{\tilde{w}}}\big)\Big] \\
&\qquad \times \mathbb{X}^*(r\underline{\tilde{w}}) + o_p(1) \\
&= \mathbb{X}^*(hr\underline{\tilde{w}}) + \Big[G\big(hr\underline{\tilde{w}}\underline{\tilde{w}}^T\big)M^- - G\big(h\underline{\underline{\tilde{w}}}\big)\{I + AM^-\}\Big]\mathbb{X}^*(r\underline{\tilde{w}}) \\
&\quad + o_p(1) \\
&= \mathbb{X}^*(hr\underline{\tilde{w}}) + \left[G\big(hr\underline{\tilde{w}}\underline{\tilde{w}}^T\big)M^- - G\big(h\underline{\underline{\tilde{w}}}\big)\left\{\underline{\underline{\lambda}}^{-1}M^- + \frac{\underline{\lambda}\underline{\lambda}^T}{|\underline{\lambda}|^2}\right\}\right]\mathbb{X}^*(r\underline{\tilde{w}}) \\
&\quad + o_p(1) \\
&\qquad [\text{as in (k)–(l) of the proof of the covariance formula (2.19)}] \\
&= \mathbb{X}^*(hr\underline{\tilde{w}}) - \big\{G\big(h\underline{\underline{\tilde{w}}}\big)\underline{\underline{\lambda}}^{-1} - G\big(hr\underline{\tilde{w}}\underline{\tilde{w}}^T\big)\big\}M^- \mathbb{X}^*(r\underline{\tilde{w}}) + o_p(1) \\
\text{(d)}\quad &= \mathbb{Y}(h) + o_p(1) \quad [\text{as defined in (2.33)}].
\end{aligned}
$$

Hence (2.34) holds for a fixed $h \in \mathbf{H}$. The proof of uniformity of convergence for $h \in \mathbf{H}$ is similar to the proof of uniformity in Theorem 2.2 □

6. Further problems. Here we give a brief discussion of some further problems related to biased sampling.

PROBLEM 6.1 (Design questions). As mentioned briefly in Sections 3 and 4, if we have some control over the choice of the number of samples s, the sampling fractions λ_i or the biasing functions w_i, we may want to choose them to optimize some criteria such as the variance of an estimate.

Here are two simple examples of problems of this type which we have already solved.

First suppose that $s = 1$, $\mathbf{X} = R^1$ and that we want to estimate the mean $\mu = \int_{-\infty}^{\infty} x\, dG(x)$ of G. What is the optimal biasing function w? It follows easily from the covariance K calculated in Example 4.2 with $h(x) = x$ [so $G(h) = \mu$], that

$$\sqrt{n}\,(\hat{\mu}_n - \mu) = \sqrt{n}\,[\mathbb{G}_n(h) - G(h)] \to_d N(0, \sigma_w^2),$$

where

$$\sigma_w^2 = \int_{-\infty}^{\infty} w(x)\, dG(x) \int_{-\infty}^{\infty} \frac{1}{w(x)}(x-\mu)^2\, dG(x). \tag{6.1}$$

By the Cauchy–Schwarz inequality, the *optimal biasing function* w_0 is $w_0(x) = |x - \mu|$, and

$$\sigma_w^2 \geq \sigma_{w_0}^2 = \left\{ \int_{-\infty}^{\infty} |x - \mu|\, dG(x) \right\}^2. \tag{6.2}$$

(This is related to "importance sampling"; see, e.g., Rubenstein [(1981) page 122].) Note that w_0 depends on G only through μ and this raises the further interesting possibility of a two-step procedure using a preliminary estimate $\hat{\mu}_0$ of μ and then biased sampling with the (estimated optimal) biasing function $\hat{w}_0(x) = |x - \hat{\mu}_0|$. Of course the preceding argument generalizes immediately to estimation of $G(h) = \int h\, dG$: The biasing function w which minimizes the asymptotic variance of the nonparametric maximum likelihood estimate is

$$w_{0h}(x) = |h(x) - G(h)| = |h(x) - E_G(h)|.$$

Note that the comparison between ordinary sampling and length-biased sampling for estimation of the mean μ of G can go either way: From Example 4.2 it follows that the asymptotic relative efficiency of length-biased sampling with respect to random sampling for estimation of μ is, with $E \equiv E_G$,

$$e_\mu(\text{length-biased, ordinary}) = \frac{E(X^2) - \mu^2}{\mu^2[\mu E(1/X) - 1]} = \frac{E(X/\mu)^2 - 1}{E(\mu/X) - 1}, \tag{6.3}$$

which varies from 0 to ∞ as G varies [0 for exponential G; ∞ for G with density $g(x) = cx^{-3} 1_{[1/2, \infty)}(x)$]. Recall from Example 4.1 that combined ordinary and length-biased sampling avoids the troubles of each [$E(X^2) = \infty$ and $E(1/X) = \infty$, respectively] and yields an estimator of μ under only the assumption $\mu = \int x\, dG(x) < \infty$.

Here is another slightly more complicated example. First, consider estimation of the mean μ of G on $\mathbf{X} = R^1$ based on stratified sampling as in Example 1.3a. For this to be possible, it is necessary to assume that the stratum probabilities $W_i = G(D_i) = p_i$, $i = 1, \ldots, s$, are *known*. Under this assumption $\mathbb{G}_n^0$ of (1.11) and (1.24) is an estimator of G. Then it follows easily from (2.39) that the asymptotic variance of the estimator $\hat{\mu}_n^0 = \int y\, d\mathbb{G}_n^0(y)$ of $\mu = \int y\, dG(y)$ is

$$V_0^2 \equiv V_0^2(\underline{\lambda}) = \sum_{i=1}^{s} \frac{p_i^2}{\lambda_i} \operatorname{Var}(Y|D_i) \equiv \sum_{i=1}^{s} \frac{p_i^2}{\lambda_i} \sigma_i^2, \tag{6.4}$$

where $Y \sim G$. This is minimized as a function of the sampling fractions $\underline{\lambda} = (\lambda_1, \ldots, \lambda_s)$ by

$$\lambda_i^{\text{opt}} = \frac{p_i \sigma_i}{\sum_{j=1}^{s} p_j \sigma_j}. \tag{6.5}$$

This is the well-known "Neyman allocation"; see, e.g., Cochran [(1963), page 97].

The resulting minimum value of V_0^2 is

$$V_0^2(\underline{\lambda}^{\text{opt}}) = \left(\sum_{i=1}^{s} p_i \sigma_i\right)^2. \tag{6.6}$$

Now consider the same problem, optimal choice of the sampling fractions λ_i, $i = 1, \ldots, s$, for estimation of the mean of G, in the context of enriched stratified sampling as in Examples 1.3b and 4.5 where now the $W_i = G(D_i) = p_i$ are *unknown*. It follows easily from the covariance formula (2.19) that the asymptotic variance of the estimator $\hat{\mu}_n = \int y \, d\mathbb{G}_n(y)$ of $\mu = \int y \, dG(y)$ is

$$\begin{aligned} V^2 &\equiv V^2(\underline{\lambda}) \\ &= \sum_{i=1}^{s} \left\{ \frac{p_i}{1 + \lambda_i/(\lambda_{s+1} p_i)} \operatorname{Var}(Y|D_i) + \frac{p_i}{\lambda_{s+1}} [E(Y - \mu|D_i)]^2 \right\} \\ &= \sum_{i=1}^{s} \left\{ \frac{p_i}{1 + \lambda_i/(\lambda_{s+1} p_i)} \sigma_i^2 + \frac{p_i}{\lambda_{s+1}} B_i^2 \right\} \end{aligned} \tag{6.7}$$

with $\sigma_i^2 \equiv \operatorname{Var}(Y|D_i)$ and $B_i \equiv E(X - \mu|D_i)$. It is easily shown that $V^2(\underline{\lambda})$ is minimized by

$$\lambda_i^{\text{opt}} = \frac{p_i}{\sum_{j=1}^{s} p_j \sigma_j} \left(\sigma_i - \min_{1 \le j \le s} \sigma_j \right), \qquad i = 1, \ldots, s, \tag{6.8}$$

and

$$\lambda_{s+1}^{\text{opt}} = \frac{1}{\sum_{i=1}^{s} p_i \sigma_i} \min_{1 \le i \le s} \sigma_i. \tag{6.9}$$

Then

$$V^2(\underline{\lambda}^{\text{opt}}) = \left(\sum_{i=1}^{s} p_i \sigma_i\right)^2 \left\{ 1 + \frac{\sum_{i=1}^{s} p_i B_i^2}{\min_{1 \le i \le s} \sigma_i} \right\}. \tag{6.10}$$

This is of course closely related to the Neyman allocation in the case of known strata probabilities, and the cost of not knowing the strata probabilities $W_i = p_i$ is the factor in brackets in (6.10).

Clearly there are a great variety of related optimization problems which are of interest and potential importance. Again note that Theorem 3.1 asserts that the nonparametric maximum likelihood estimator $\mathbb{G}_n$ is optimal once the *design has been fixed*.

Problem 6.2 ($s = s_n \to \infty$). Regression problems with biased (stratified) sampling on the dependent variable as in Example 4.6 and Jewell (1985) can be reformulated in terms of our present model with $s = s_n$ depending on n and $s_n \to \infty$ as $n \to \infty$. A treatment of biased sampling regression models along these lines is given by Bickel and Ritov (1987).

Problem 6.3 (Biasing functions dependent on unknown parameters). As seen in Example 4.4, Problem 6.1 and Vardi [(1985a), page 198], situations in which the biasing functions w_i depend on an unknown parameter arise frequently. A thorough treatment of these situations would be of interest.

PROBLEM 6.4 (Biasing and censoring). Vardi [(1985a), page 195] suggests how to use an EM algorithm to estimate G from s samples involving both biasing and censoring. The large-sample behavior of the resulting estimator is still unknown.

Acknowledgments. We owe thanks to Peter Bickel and Ya'acov Ritov for conversations which lead to improvements and simplifications of our proofs.

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BERAN, R. (1977). Estimating a distribution function. *Ann. Statist.* **5** 400–404.

BERMAN, A. and PLEMMONS, R. J. (1979). *Nonnegative Matrices in the Mathematical Sciences.* Academic, New York.

BHATTACHARYA, P. K., CHERNOFF, H. and YANG, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11** 505–514.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1989). *Efficient and Adaptive Inference Semiparametric Models.* Johns Hopkins Univ. Press, Baltimore, Md. To appear.

BICKEL, P. J. and RITOV, Y. (1987). Large sample theory of estimation in biased sampling regression models I. Technical Report, Univ. California, Berkeley.

BRESLOW, N. E. and DAY, N. E. (1980). *The Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

COCHRAN, W. G. (1963). *Sampling Techniques*, 2nd ed. Wiley, New York.

COSSLETT, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49** 1289–1316.

COX, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, Jr., eds.) 506–527. Wiley, New York.

DUDLEY, R. M. (1984). A course on empirical processes. *Ecole d'Été de Probabilités de Saint Flour XII-1982. Lecture Notes in Math.* **1097** 1–142. Springer, New York.

DUDLEY, R. M. and PHILIPP, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. verw. Gebiete* **62** 509–552.

DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

GAENSSLER, P. (1983). *Empirical Processes.* IMS, Hayward, Calif.

GILL, R. D. (1988). Non- and semiparametric maximum likelihood estimators and the von Mises method, Part I. *Scand. J. Statist.* To appear.

GILL, R. D. (1989). Non- and semiparametric maximum likelihood estimators and the von Mises method, Part II.

GILL, R. D. and WELLNER, J. A. (1985). Large sample theory of empirical distributions in biased sampling models. Technical Report 75, Dept. Statistics, Univ. Washington; Technical Report MS-R8603, Dept. Mathematical Statistics, Centrum voor Wiskunde en Informatica, Amsterdam.

JEWELL, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* **72** 11–21.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887–906.

KIEFER, J. and WOLFOWITZ, J. (1959). Asymptotic minimax character of the sample distribution function for vector chance variables. *Ann. Math. Statist.* **30** 463–489.

LEVIT, B. YA. (1978). Infinite dimensional informational lower bounds. *Theory Probab. Appl.* **23** 371–377.

MANSKI, C. F. and LERMAN, S. R. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* **45** 1977–1988.

MANSKI, C. F. and MCFADDEN, D., EDS. (1981). *Structural Analysis of Discrete Data with Applications.* MIT Press, Cambridge, Mass.

MILLAR, P. W. (1979). Asymptotic minimax theorems for the sample distribution function. *Z. Wahrsch. verw. Gebiete* **48** 233–252.

MILLAR, P. W. (1985). Nonparametric applications of an infinite dimensional convolution theorem. *Z. Wahrsch. verw. Gebiete* **68** 545–556.

MORGENTHALER, S. and VARDI, Y. (1986). Choice based samples: A non-parametric approach. *J. Econometrics* **32** 109–125.

ORTEGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic, New York.

OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919.

POLLARD, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A* **33** 235–248.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd ed. Wiley, New York.

RAO, C. R. and MITRA, S. K. (1971). *Generalized Inverse of Matrices and Its Applications.* Wiley, New York.

ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton Univ. Press, Princeton, N.J.

RUBINSTEIN, R. Y. (1981). *Simulation and the Monte Carlo Method.* Wiley, New York.

SEBER, G. A. F. (1977). *Linear Regression Analysis.* Wiley, New York.

VAN DER VAART, A. (1988). *Statistical Estimation in Large Parameter Spaces.* CWI Tract **44**. Centrum voor Wiskunde en Informatica, Amsterdam.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10** 616–620.

VARDI, Y. (1985a). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–203.

VARDI, Y. (1985b). Asymptotics for empirical distributions in selection bias models. Report, AT & T Bell Laboratories, Murray Hill, N.J.

RICHARD D. GILL
CENTRUM VOOR WISKUNDE
EN INFORMATICA
KRUISLAAN 413
1098 SJ AMSTERDAM
THE NETHERLANDS

YEHUDA VARDI
DEPARTMENT OF STATISTICS
BUSCH CAMPUS
RUTGERS UNIVERSITY
NEW BRUNSWICK, NEW JERSEY 08903-0270

JON A. WELLNER
DEPARTMENT OF STATISTICS GN-22
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195