האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

STOCHASTIC COMPARISONS OF STRATIFED SAMPLING TECHNIQUES FOR SOME MONTE CARLO ESTIMATORS

By

LARRY GOLDSTEIN, YOSEF RINOTT and MARCO SCARSINI

Discussion Paper # 556 July 2010

מרכז לחקר הרציונליות

CENTER FOR THE STUDY
OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Stochastic comparisons of stratified sampling techniques for some Monte Carlo estimators

Larry Goldstein
Department of Mathematics
University of Southern California
Kaprielian Hall, Room 108
3620 Vermont Avenue
Los Angeles, CA 90089-2532, USA
larry@math.usc.edu

Yosef Rinott[*]
Department of Statistics
and Center for the Study of Rationality
Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905, Israel
and LUISS, Roma
rinott@mscc.huji.ac.il

Marco Scarsini[†]
Dipartimento di Scienze Economiche e Aziendali
LUISS
Viale Romania 12
I–00197 Roma, Italy
and HEC, Paris
marco.scarsini@luiss.it

May 27, 2010

[*] Partially supported by the Israel Science Foundation grant No. 473/04
[†] Partially supported by MIUR-COFIN

Abstract

We compare estimators of the (essential) supremum and the integral of a function f defined on a measurable space when f may be observed at a sample of points in its domain, possibly with error. The estimators compared vary in their levels of stratification of the domain, with the result that more refined stratification is better with respect to different criteria. The emphasis is on criteria related to stochastic orders. For example, rather than compare estimators of the integral of f by their variances (for unbiased estimators), or mean square error, we attempt the stronger comparison of convex order when possible. For the supremum the criterion is based on the stochastic order of estimators.

For some of the results no regularity assumptions for f are needed, while for others we assume that f is monotone on an appropriate domain.

AMS 2000 Subject Classification: Primary 65C05. Secondary 60E15, 62D05.

Keywords: convex loss, convex order, stochastic order, majorization, stratified sampling.

1 Introduction

In many situations the cost of computing the value of a function f is very high, either because the analytic expression of the function is extremely complex, or because the value is the result of a costly experiment. For example, f could be the level of toxicity as a reaction to different doses of certain drugs, or it could be the output of a chemical experiment, or it could be the survival time of a patient undergoing a certain treatment. Therefore the function can be computed only at a limited number of points. One standard way to choose these points is via some Monte Carlo randomization. Different possibilities arise: points could be sampled totally at random, or some stratification could be used. When properly carried out, stratification is known to improve the performance of estimators. The purpose of this paper is to qualify the above statement in some relevant cases, and to compare different sampling stratifications according to some suitable criteria.

Often the object of interest is some functional of f such as its supremum or integral. Monte Carlo estimation of such functionals is the subject of a very large number of papers. In most cases some regularity of the function f is assumed, see, for example, Novak (1988) or Zhigljavsky and Chekmasov (1996). Under some regularity conditions it is often reasonable to estimate the entire function and then use a plug-in method to estimate the functional. When no regularity is assumed for f, then it may be more reasonable to estimate the functional directly.

Given a measurable space $(\mathfrak{U}, \mathscr{U})$, let $f : \mathfrak{U} \to \mathbb{R}$ be a measurable function f. In order to estimate $\theta := \sup_{x \in \mathfrak{U}} f(x)$ we can draw a sample $X_1, \dots, X_n$ of n points in $\mathfrak{U}$ and use the estimator $T := \max(f(X_1), \dots, f(X_n))$. Alternatively we can sample the X's by resorting to some stratification. Ermakov, Zhiglyavskiĭ, and Kondratovich (1988), Kondratovich and Zhigljavsky (1998), and Zhigljavsky and Žilinskas (2008) prove that, if we consider two partitions of $\mathfrak{U}$, one of which is a refinement of the other, and we sample in proportion to the measure of each element of the partition, then the more refined partition produces a stochastically larger estimator of the supremum. Since these estimators are almost surely smaller than θ (hence biased), and consistent, the stochastically larger one performs better. Thus the more we stratify the better the estimator we obtain.

In our paper we extend this result and show that the stochastic comparison for estimators of the supremum holds also when observations are censored, that is, when for a sample of pairs of random variables (U_i, Z_i) we only know whether $Z_i \leq f(U_i)$ or not. In applications, there may be situations where exact evaluation of $f(u)$ at a given point is difficult or expensive, whereas a comparison of $f(u)$ to a given constant t is (at least for most values of t) much easier. For example, if $f(u)$ represents a lifetime, it may easier to see if it has exceeded a certain value, rather than wait to obtain the exact value $f(u)$ itself. This amounts to censoring.

When we want to estimate the integral $I(f)$ of the function f, then it is easy to construct an unbiased estimators of $I(f)$ by using different stratified samples.

Unbiasedness of these estimators implies that the comparison criterion cannot be the stochastic order, as used for the maximum.

In much of the literature estimators are compared in terms of a given loss function, which may be arbitrary. Typically the loss function is quadratic, so the criterion is the mean square error, i.e., the variance, when the estimator is unbiased. More generally, it may be possible to find comparison criteria that are valid for large classes of loss functions, for instance all losses of the type $|W - I(f)|^p$, where W is an estimator of $I(f)$ and $p \geq 1$, or even the class of all convex loss functions. The use of the entire class of convex loss functions in inference goes back at least to Laycock and Silvey (1968) and Laycock (1972). Similar ideas have later been used, e.g., by Berger (1976), Kozek (1977), Lin and Mousa (1982), Eberl (1984), Bai and Durairajan (1997), and Petropoulos and Kourouklis (2001). A comparison of the performance of different estimators, with respect to all convex loss functions, can be achieved by considering the convex order. Comparison of experiments in term of the convex order traces back to Blackwell (1951, 1953).

It is well known that stratification reduces the variance of estimators of $I(f)$, but, as will be shown below, stratification does not necessarily reduce $\mathbb{E}[|W - I(f)|^p]$, for $p \neq 2$, which implies that, even if stratification is useful in L_2, it may be counterproductive in L_1, for instance. We will show that in some circumstances stratified sampling is better not just in L_2, but in terms of the convex order, which in turn implies that it is better in L_p for every $p \geq 1$. This is the case when observations are censored, or the function f is univariate and monotone, or when it is multivariate and monotone and the sampling is independent across coordinates. Papageorgiou (1993) shows the computational advantage of using randomized methods to compute the integral of monotone d-variate functions, and shows how this depends on d.

Our results also hold when the function f can only be observed with noise, for instance, when f is observed as the outcome of some experiment. Moreover our regularity assumptions on the function f are rather nonrestrictive: measurability when estimating the maximum, boundedness when observations are censored, and sometimes monotonicity when estimating the integral.

We emphasize that in our framework evaluations of f by experiment is the costly part, and any precalculations, such as those required for computing strata and sampling from the conditional distributions in strata, even if computer-time consuming, are considered to have a relatively negligible cost.

The paper is organized as follows. Section 2 fixes notation and reviews various properties of stochastic orders and certain dependence structures. Section 3 compares estimators of the supremum of a function, considering also the case of censored observations. Section 4 compares estimators of integrals: first a variance comparison is shown to hold in general, even when observations are affected by errors, then a counter example is provided for a non-quadratic loss function. Then censored observations are considered and a comparison in terms of the convex order is proved in this case. Finally monotone functions are examined. In the univariate case a convex order comparison holds. In the multivariate case this is true under some additional conditions on the stratification and on the dependence of the underlying random vector. Finally Section 5 contains some numerical examples.

2 Notation and preliminaries

In the whole paper a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is assumed in the background. The *stochastic order* $\leq_{\text{st}}$, the *convex order* $\leq_{\text{cx}}$, the *increasing convex order* $\leq_{\text{icx}}$, and the *majorization order* $\prec$ are defined as follows (see, e.g., Marshall and Olkin (1979), Müller and Stoyan (2002), Shaked and Shanthikumar (2007)). Given two random vectors $\boldsymbol{X}, \boldsymbol{Y}$ we say that $\boldsymbol{Y} \leq_{\text{st}} \boldsymbol{X}$ if

$$\mathbb{E}[\phi(\boldsymbol{Y})] \leq \mathbb{E}[\phi(\boldsymbol{X})] \tag{2.1}$$

for all nondecreasing functions ϕ; we say that $\boldsymbol{Y} \leq_{\text{cx}} \boldsymbol{X}$ if (2.1) holds for all convex functions ϕ, and we say that $\boldsymbol{Y} \leq_{\text{icx}} \boldsymbol{X}$ if (2.1) holds for all nondecreasing convex functions ϕ. It is well known that $\boldsymbol{Y} \leq_{\text{st}} \boldsymbol{X}$ iff $\mathbb{P}(\boldsymbol{Y} \in A) \leq \mathbb{P}(\boldsymbol{X} \in A)$ for all increasing sets A, where we call a set *increasing* if its indicator function is nondecreasing. In the case of univariate random variables X, Y, the above inequality becomes $\mathbb{P}(Y \leq t) \geq \mathbb{P}(X \leq t)$ for all $t \in \mathbb{R}$. It is well known that $X \leq_{\text{cx}} Y$ implies $\mathbb{E}[X] = \mathbb{E}[Y]$ and $\operatorname{Var}[X] \leq \operatorname{Var}[Y]$.

The statement $\boldsymbol{Y} \leq_{\text{st}} \boldsymbol{X}$ depends only on the marginal laws $\mathscr{L}(\boldsymbol{Y})$ and $\mathscr{L}(\boldsymbol{X})$, so sometimes we write $\mathscr{L}(\boldsymbol{Y}) \leq_{\text{st}} \mathscr{L}(\boldsymbol{X})$, and analogously for $\leq_{\text{cx}}$ and $\leq_{\text{icx}}$.

Given two vectors $\boldsymbol{x} = (x_1, \dots, x_n)$, $\boldsymbol{y} = (y_1, \dots, y_n)$, we write $\boldsymbol{y} \prec \boldsymbol{x}$ if

$$\sum_{i=1}^{k} y_i^{\downarrow} \leq \sum_{i=1}^{k} x_i^{\downarrow} \quad \text{for} \quad k = 1, \dots, n-1, \qquad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i,$$

where $y_1^{\downarrow} \geq \cdots \geq y_n^{\downarrow}$ is the decreasing rearrangement of $\boldsymbol{y}$, and analogously for $\boldsymbol{x}$. The relation $\boldsymbol{y} \prec \boldsymbol{x}$ holds if and only if there exists an $n \times n$ doubly stochastic matrix $\boldsymbol{D}$ such that $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$.

A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is called Schur convex, or Schur concave, if $\boldsymbol{y} \prec \boldsymbol{x}$ implies $\psi(\boldsymbol{y}) \leq \psi(\boldsymbol{x})$, or $\psi(\boldsymbol{y}) \geq \psi(\boldsymbol{x})$, respectively. If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex then $\psi(\boldsymbol{x}) = \sum_{i=1}^{n} \varphi(x_i)$ is Schur convex.

A random vector $\boldsymbol{X}$ is *associated* if for all nondecreasing functions ϕ, ψ we have $\operatorname{Cov}[\phi(\boldsymbol{X}), \psi(\boldsymbol{X})] \geq 0$.

Recall that a subset $A \subset \mathbb{R}^d$ is a *lattice* if it is closed under componentwise maximum $\vee$ and minimum $\wedge$. A random vector $\boldsymbol{X}$ is *multivariate totally positive of order* 2 (MTP_2) if its support is a lattice and its density $f_{\boldsymbol{X}}$ with respect to some product measure on $\mathbb{R}^d$ satisfies $f_{\boldsymbol{X}}(\boldsymbol{s})\, f_{\boldsymbol{X}}(\boldsymbol{t}) \leq f_{\boldsymbol{X}}(\boldsymbol{s} \vee \boldsymbol{t})\, f_{\boldsymbol{X}}(\boldsymbol{s} \wedge \boldsymbol{t})$ for all $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$. MTP_2 implies association. Also, any vector having independent components is MTP_2.

Let U be a random variable with values in some measurable space $(\mathfrak{U}, \mathscr{U})$ with nonatomic law P_U. A finite sequence $\mathscr{B} = (B_1, \dots, B_b)$ of subsets of $\mathfrak{U}$ is called an *ordered partition* of $\mathfrak{U}$ if $B_i \cap B_j = \varnothing$ for $i, j \in \{1, \dots, b\}$, $i \neq j$, and $\cup_{i=1}^{b} B_i = \mathfrak{U}$. For the sake of brevity in the sequel whenever we say partition we mean ordered partition.

Here we consider partitions $\mathscr{B} = (B_1, \dots, B_b)$ of $\mathfrak{U}$ where the sets B_i are measurable and such that for $i = 1, \dots, b$ we have $\mathbb{P}(U \in B_i) = k_i/n$, for some $k_i \in \{1, \dots, n\}$ satisfying $\sum_i k_i = n$. We say that such a partition $\mathscr{B}$ of $\mathfrak{U}$ and a partition $\mathscr{B}^* = (B_1^*, \dots, B_b^*)$ of $N := \{1, \dots, n\}$ are associated if the cardinalities $|B_i^*|$ of the sets B_i^* satisfy $|B_i^*| = k_i$ for $i = 1, \dots, b$. We then have

$$\mathbb{P}(U \in B_i) = \frac{|B_i^*|}{n}. \tag{2.2}$$

The notation $B \in \mathscr{B}$ means that B is one of the sets B_i which comprise $\mathscr{B}$, and, given $B \in \mathscr{B}$ we let B^* denote the corresponding set B_i^* in $\mathscr{B}^*$ such that (2.2) holds.

Given two partitions $\mathscr{B}^* = (B_1^*, \dots, B_b^*)$ and $\mathscr{C}^* = (C_1^*, \dots, C_c^*)$ of N we write $\mathscr{C}^* \leq_{\text{ref}} \mathscr{B}^*$, that is, that $\mathscr{B}^*$ is a refinement of $\mathscr{C}^*$, when every set in $\mathscr{C}^*$ is the union of sets in $\mathscr{B}^*$. We will use the same order $\leq_{\text{ref}}$ also for partitions of $\mathfrak{U}$. Clearly, if $\mathscr{C}$ and $\mathscr{B}$ are partitions of $\mathfrak{U}$, each of which can be associated to some partition of N, then $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$ implies that there exist partitions $\mathscr{C}^*$ and $\mathscr{B}^*$ associated to $\mathscr{C}$ and $\mathscr{B}$, respectively, satisfying $\mathscr{C}^* \leq_{\text{ref}} \mathscr{B}^*$.

Call $\mathscr{A}^* = (\{1\}, \dots, \{n\})$ the finest partition of N and $\mathscr{D}^* = (N)$ the coarsest partition of N. Then $\mathscr{D}^* \leq_{\text{ref}} \mathscr{B}^* \leq_{\text{ref}} \mathscr{A}^*$ for all $\mathscr{B}^*$, and for any partition $\mathscr{A}$ of $\mathfrak{U}$ associated to $\mathscr{A}^*$ we have $\mathbb{P}(U \in A_i) = 1/n$.

For a partition $\mathscr{B}$ and $B \in \mathscr{B}$, let $P_{U|B}$ denote the conditional law of U given $U \in B$. Let $\{V_j^B, j \in B^*\}$ be random variables with law $P_{U|B}$ with $\{V_j^B, j \in B^*, B \in \mathscr{B}\}$ independent.

3 The supremum

Let $f : \mathfrak{U} \to \mathbb{R}$ be measurable, and define

$$W_{\text{S}}^{\mathscr{B}} = \max_{B \in \mathscr{B}} \max_{j \in B^*} f(V_j^B), \tag{3.1}$$

where the subscript S indicates that $W_{\text{S}}^{\mathscr{B}}$ will be used to estimate the (essential) supremum of the function f.

Given a random variable U with values in $(\mathfrak{U}, \mathscr{U})$, let $f^* := \operatorname{ess\,sup} f(U)$. It is clear that for any choice of partition $\mathscr{B}$, $\mathbb{P}(W_{\text{S}}^{\mathscr{B}} \leq f^*) = 1$. The following result compares two estimators of type $W_{\text{S}}^{\mathscr{B}}$. Since both estimators underestimate f^*, the stochastically larger one is preferable. This theorem, which goes back to Ermakov et al. (1988) and Kondratovich and Zhigljavsky (1998), can be found also in Zhigljavsky and Žilinskas (2008, Theorem 3.4)

Theorem 3.1. *If $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$, then $W_{\text{S}}^{\mathscr{C}} \leq_{\text{st}} W_{\text{S}}^{\mathscr{B}}$.*

A short proof of Theorem 3.1, different from the one in Zhigljavsky and Žilinskas (2008), can be found in Appendix A.

As mentioned in the Introduction, in many practical situations data are not always observed exactly, but may be censored, for various reasons, including budget constraints. We extend now the comparison result of Theorem 3.1 to the case of censored observations. Let $f : \mathfrak{U} \to \mathbb{R}$ be bounded; without loss of generality we take $0 \le f(u) \le 1$ for all $u \in \mathfrak{U}$. In this section we assume that for a sample of points of the type $(u, t) \in \mathfrak{U} \times [0, 1]$ we are allowed to observe only the value of t and whether $t > f(u)$.

For any partition $\mathscr{B}$ with associated partition $\mathscr{B}^*$, let $\{V_j^B, j \in B^*\}$, $B \in \mathscr{B}$, and $\{T_j, j \in N\}$ be independent random variables with law $P_{U|B}$ and the uniform distribution on $[0, 1]$, respectively, and let

$$S^{\mathscr{B}} = \bigcup_{B \in \mathscr{B}} \{j \in B^* : T_j \le f(V_j^B)\}, \quad \text{and} \quad W_{\text{CS}}^{\mathscr{B}} = \max_{j \in S^{\mathscr{B}}} T_j.$$

When $S^{\mathscr{B}} = \varnothing$ we set $W_{\text{CS}}^{\mathscr{B}} = 0$. The letter C in the subscript CS indicates censored data. Again it is clear that $\mathbb{P}(W_{\text{CS}}^{\mathscr{B}} \le f^*) = 1$, so the estimator $W_{\text{CS}}^{\mathscr{B}}$ underestimates f^*.

Theorem 3.2. *If $\mathscr{C} \le_{\text{ref}} \mathscr{B}$, then $W_{\text{CS}}^{\mathscr{C}} \le_{\text{st}} W_{\text{CS}}^{\mathscr{B}}$.*

Proof. Below when we write V_j^B without specifying B, we mean that $B \in \mathscr{B}$ corresponds in the sense of (2.2) to the set $B^* \in \mathscr{B}^*$ which contains the index j. For any $t \in [0, 1]$ we may calculate the distribution function of $W_{\text{CS}}^{\mathscr{B}}$ at t by writing

$$\begin{aligned}
\{W_{\text{CS}}^{\mathscr{B}} \le t\} &= \bigcup_{R \subset N} \left\{\max_{j \in S^{\mathscr{B}}} T_j \le t, S^{\mathscr{B}} = R\right\} \\
&= \bigcup_{R \subset N} \left\{T_j \le t, T_j \le f(V_j^B) \text{ for all } j \in R, \text{ and } T_j > f(V_j^B) \text{ for all } j \notin R\right\} \\
&= \bigcup_{R \subset N} \left\{T_j \le t \wedge f(V_j^B) \text{ for all } j \in R, \text{ and } T_j > f(V_j^B) \text{ for all } j \notin R\right\}.
\end{aligned}$$

Hence, conditionally on $\{V_j^B,\ j \in B^*,\ B \in \mathscr{B}\}$, using the fact that the T_j's are uniform, we obtain:

$$\begin{aligned}
\mathbb{P}(W_{\text{CS}}^{\mathscr{B}} \le t \,|\, V_j^B, j \in B^*, B \in \mathscr{B}) &= \sum_{R \subset N} \prod_{j \in R} \mathbb{P}(T_j \le t \wedge f(V_j^B)) \prod_{j \notin R} \mathbb{P}(T_j > f(V_j^B)) \\
&= \sum_{R \subset N} \prod_{j \in R} (t \wedge f(V_j^B)) \prod_{j \notin R} (1 - f(V_j^B)) \qquad (3.2) \\
&= \sum_{h_1=1}^{|B_1^*|} \cdots \sum_{h_b=1}^{|B_b^*|} \sum_{\substack{R \subset N \\ \forall i, |R \cap B_i^*| = h_i}} \prod_{j \in R} (t \wedge f(V_j^B)) \prod_{j \notin R} (1 - f(V_j^B)).
\end{aligned}$$

Taking expectation we obtain the unconditional distribution,

$$\begin{aligned}\mathbb{P}(W_{\mathrm{CS}}^{\mathscr{B}} \le t) &= \sum_{h_1=1}^{|B_1^*|} \cdots \sum_{h_b=1}^{|B_b^*|} \prod_{i=1}^{b} \binom{|B_i^*|}{h_i} \left(\int_{B_i} (t \wedge f(u)) \,\mathrm{d}P_{U|B_i}(u) \right)^{h_i} \\ &\qquad \cdot \left(\int_{B_i} (1 - f(u)) \,\mathrm{d}P_{U|B_i}(u) \right)^{|B_i^*|-h_i} \\ &= \prod_{B \in \mathscr{B}} \left(\int_B (t \wedge f(u)) \,\mathrm{d}P_{U|B}(u) + \int_B (1 - f(u)) \,\mathrm{d}P_{U|B}(u) \right)^{|B^*|}.\end{aligned}$$

Let

$$q^B = \int_B (t \wedge f(v)) \,\mathrm{d}P_{U|B}(v) + \int_B (1 - f(v)) \,\mathrm{d}P_{U|B}(v) = \int_B [(t \wedge f(v)) + (1 - f(v))] \,\mathrm{d}P_{U|B}(v).$$

If C is a union of disjoint sets B_i then

$$q^C = \sum_i q^{B_i} \frac{\mathbb{P}(U \in B_i)}{\mathbb{P}(U \in C)} = \sum_i q^{B_i} \frac{|B_i^*|}{|C^*|}. \tag{3.3}$$

If $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$ then

$$(\underbrace{q^{C_1}, \ldots, q^{C_1}}_{|C_1^*|}, \ldots, \underbrace{q^{C_c}, \ldots, q^{C_c}}_{|C_c^*|}) \prec (\underbrace{q^{B_1}, \ldots, q^{B_1}}_{|B_1^*|}, \ldots, \underbrace{q^{B_b}, \ldots, q^{B_b}}_{|B_b^*|}).$$

To see this, observe that (3.3) implies that the vector on the left-hand side above is obtained from the one on the right by multiplying it by the $n \times n$ doubly stochastic matrix $\boldsymbol{D}$ which is block diagonal where the i-th block is the $|C_i^*| \times |C_i^*|$ matrix with all entries equal to $1/|C_i^*|$. Therefore, by the Schur concavity of the function $(\theta_1, \ldots, \theta_n) \mapsto \prod_{i=1}^n \theta_i$, we have

$$\mathbb{P}(W_{\mathrm{CS}}^{\mathscr{C}} \le t) = \prod_{C \in \mathscr{C}} \left(q^C\right)^{|C^*|} \ge \prod_{B \in \mathscr{B}} \left(q^B\right)^{|B^*|} = \mathbb{P}(W_{\mathrm{CS}}^{\mathscr{B}} \le t).$$

□

For every $n \in \mathbb{N}$ and for every partition $\mathscr{B}_n$ associated to a partition $\mathscr{B}_n^*$ of $\{1, \ldots, n\}$, we have $W_{\mathrm{CS}}^{\mathscr{B}_n} \le_{\mathrm{st}} W_{\mathrm{S}}^{\mathscr{B}_n}$. Therefore

$$W_{\mathrm{CS}}^{\mathscr{D}_n} \le_{\mathrm{st}} W_{\mathrm{CS}}^{\mathscr{B}_n} \le_{\mathrm{st}} W_{\mathrm{S}}^{\mathscr{B}_n} \le_{\mathrm{st}} f^*.$$

Since $W_{\mathrm{CS}}^{\mathscr{D}_n}$ is consistent for f^* as $n \to \infty$, we have that $W_{\mathrm{CS}}^{\mathscr{B}_n}$ and $W_{\mathrm{S}}^{\mathscr{B}_n}$ are consistent, too.

4 The integral

With the subscript I standing for integral, let

$$W_{\mathrm{I}}^{\mathscr{B}} = \frac{1}{n} \sum_{B\in\mathscr{B}} \sum_{j\in B^*} f(V_j^B) \tag{4.1}$$

$$W_{\mathrm{IE}}^{\mathscr{B}} = \frac{1}{n} \sum_{B\in\mathscr{B}} \sum_{j\in B^*} \left(f(V_j^B) + \varepsilon_j\right), \tag{4.2}$$

where the variables ε_j are independent copies of a random variable ε having mean 0 and finite variance, independent of the variables V_j^B. Clearly $W_{\mathrm{I}}^{\mathscr{B}}$ and $W_{\mathrm{IE}}^{\mathscr{B}}$ are both unbiased estimators of $\overline{f} := \mathbb{E}[f(U)] = \int f(U)\ \mathrm{d}\mathbb{P}$ when $\int |f(U)|\ \mathrm{d}\mathbb{P}$ is finite, and $W_{\mathrm{I}}^{\mathscr{B}}$ is the special case of $W_{\mathrm{IE}}^{\mathscr{B}}$ when the error has zero variance, that is, there is no measurement error.

The following result is well-known when the error has zero variance (see, e.g., Glasserman, 2004, Section 4.3). We extend it to a more general case, relevant when the evaluation of f is the result of an experiment.

Theorem 4.1. *If* $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$*, then* $\mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{B}}] \leq \mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{C}}]$.

The proof of Theorem 4.1 can be found in Appendix A.

It follows immediately from Theorem 4.1 that $\mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{A}}] \leq \mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{D}}]$, hence, in particular, $\mathrm{Var}[W_{\mathrm{I}}^{\mathscr{A}}] \leq \mathrm{Var}[W_{\mathrm{I}}^{\mathscr{D}}]$. The following counterexample shows nevertheless that, even when the function is observed without error, $W_{\mathrm{I}}^{\mathscr{A}} \not\leq_{\mathrm{cx}} W_{\mathrm{I}}^{\mathscr{D}}$, that is, domination in the convex order does not hold. In the counterexample we consider the absolute error, that is, (L_1), rather than mean square error, (L_2).

Example 4.2. Let $\mathfrak{U} = [0,1]$ and U have a uniform distribution on $[0,1]$. Furthermore let $n = 2$, and $A_1 = [0,1/2], A_2 = (1/2,1]$. Define

$$f(u) = 4I_{[0,1/2]}(u) + 2I_{(1/2,3/4]}(u) + 6I_{(3/4,1]}(u).$$

Then $W_{\mathrm{I}}^{\mathscr{D}}$ takes the values $2,3,4,5,6$ with probabilities $(1,4,6,4,1)/16$, respectively. The variable $W_{\mathrm{I}}^{\mathscr{A}}$, based on one random observation from each of the above intervals A_i, takes the values 3 and 5 each with probability $1/2$. Therefore $\mathbb{E}[W_{\mathrm{I}}^{\mathscr{A}}] = 4 = \mathbb{E}[W_{\mathrm{I}}^{\mathscr{D}}]$.

We have $\mathrm{Var}[W_{\mathrm{I}}^{\mathscr{D}}] = \mathrm{Var}[W_{\mathrm{I}}^{\mathscr{A}}] = 1$, but for the convex function $\psi(u) = |u-4|$ we have

$$\mathbb{E}[\psi(W_{\mathrm{I}}^{\mathscr{D}})] = \mathbb{E}|W_{\mathrm{I}}^{\mathscr{D}} - 4| = 2\frac{2}{16} + 2\frac{4}{16} = \frac{12}{16} < 1 = \mathbb{E}|W_{\mathrm{I}}^{\mathscr{A}} - 4| = \mathbb{E}[\psi(W_{\mathrm{I}}^{\mathscr{A}})].$$

A more general example can be constructed as follows. Consider a partition $\mathscr{A}$ associated to the finest partition $\mathscr{A}^*$ of N. Split A_1 into two measurable subsets

A_{1a}, A_{1b} such that $\mathbb{P}(U \in A_{1a}) = \mathbb{P}(U \in A_{1b}) = 1/(2n)$. Consider now a function f defined as follows:

$$f(u) = \begin{cases} 1 & \text{if } u \in A_{1a}, \\ -1 & \text{if } u \in A_{1b}, \\ 0 & \text{elsewhere.} \end{cases} \tag{4.3}$$

For all $i \in N$ we have $\mathbb{E}[f(U)|U \in A_i] = 0$ and

$$\operatorname{Var}[f(U)|U \in A_i] = \begin{cases} 1 & \text{for } i = 1, \\ 0 & \text{for } i \neq 1. \end{cases}$$

Hence

$$\operatorname{Var}[W_{\mathrm{I}}^{\mathscr{A}}] = \mathbb{E}[(W_{\mathrm{I}}^{\mathscr{A}})^2] = \frac{1}{n^2}.$$

Moreover, if $V_1, \dots, V_n$ are i.i.d. copies of U,

$$\begin{aligned} \operatorname{Var}[W_{\mathrm{I}}^{\mathscr{D}}] &= \operatorname{Var}\left[\frac{1}{n}\sum_{j=1}^{n} f(V_j)\right] \\ &= \frac{1}{n^2}\sum_{j=1}^{n} \operatorname{Var}[f(V_j)] \\ &= \frac{1}{n^2} \\ &= \operatorname{Var}[W_{\mathrm{I}}^{\mathscr{A}}]. \end{aligned}$$

Analogously

$$\mathbb{E}[|f(U)||U \in A_i] = \begin{cases} 1 & \text{for } i = 1, \\ 0 & \text{for } i \neq 1. \end{cases}$$

Therefore

$$\mathbb{E}|W_{\mathrm{I}}^{\mathscr{A}}| = \sqrt{\mathbb{E}[(W_{\mathrm{I}}^{\mathscr{A}})^2]} = \frac{1}{n}.$$

For any square integrable random variable Y we have $\mathbb{E}|Y| \leq \sqrt{\mathbb{E}[Y^2]}$ and the inequality is strict if Y is not almost surely constant. Hence

$$\mathbb{E}|W_{\mathrm{I}}^{\mathscr{D}}| < \sqrt{\mathbb{E}[(W_{\mathrm{I}}^{\mathscr{D}})^2]} = \sqrt{\mathbb{E}[(W_{\mathrm{I}}^{\mathscr{A}})^2]} = \mathbb{E}|W_{\mathrm{I}}^{\mathscr{A}}| = \frac{1}{n}.$$

Example 4.2 proves that the convex order does not hold in general between estimators $W_{\mathrm{I}}^{\mathscr{B}}$ and $W_{\mathrm{I}}^{\mathscr{C}}$ when $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$. Nevertheless, in the following subsections we show that, under some natural conditions, comparisons in the convex order are possible.

4.1 Censored observations

Keeping the notation and spirit of Section 3, consider a function f such that $0 \le f(u) \le 1$ for all $u \in \mathfrak{U}$. Assume that for a sample of points of the type $(u,t) \in \mathfrak{U} \times [0,1]$ we are allowed to observe only the value of t and whether $t \le f(u)$, and let

$$W_{\mathrm{CI}}^{\mathscr{B}} = \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} I_{\{T_j \le f(V_j^B)\}}.$$

Note that $W_{\mathrm{CI}}^{\mathscr{B}}$ is an unbiased estimator of $\overline{f} = \mathbb{E}[f(U)]$, as

$$\begin{aligned}\mathbb{E}[W_{\mathrm{CI}}^{\mathscr{B}}] &= \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} \mathbb{P}(T_j \le f(V_j^B)) = \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} \int_{\mathfrak{U}} \int_0^1 I_{\{t \le f(u)\}} \, \mathrm{d}t \, \mathrm{d}P_{U|B}(u) \\ &= \sum_{B \in \mathscr{B}} \frac{|B^*|}{n} \int_{\mathfrak{U}} f(u) \, \mathrm{d}P_{U|B}(u) = \sum_{B \in \mathscr{B}} \mathbb{P}(B) \mathbb{E}[f(U)|U \in B] \\ &= \mathbb{E}[f(U)].\end{aligned}$$

Theorem 4.3. *If $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$, then $W_{\mathrm{CI}}^{\mathscr{B}} \le_{\mathrm{cx}} W_{\mathrm{CI}}^{\mathscr{C}}$.*

Proof. By a result in Karlin and Novikoff (1963) (see also Marshall and Olkin, 1979, Sections 12.F and 15.E), if

$$X_{\boldsymbol{p}} = \frac{1}{n} \sum_{i=1}^{n} \xi_i,$$

where $\xi_1, \ldots, \xi_n$ are independent Bernoulli variables with parameters $p_1, \ldots, p_n$, and $\boldsymbol{p} = (p_1, \ldots, p_n)$, then

$$\boldsymbol{p} \prec \boldsymbol{q} \quad \text{implies} \quad X_{\boldsymbol{q}} \le_{\mathrm{cx}} X_{\boldsymbol{p}}. \tag{4.4}$$

Define

$$p^C = \mathbb{P}(T_j \le f(V_j^C)), \quad p^B = \mathbb{P}(T_j \le f(V_j^B)),$$

and

$$\boldsymbol{p} = (\underbrace{p^{C_1}, \ldots, p^{C_1}}_{|C_1^*|}, \ldots, \underbrace{p^{C_c}, \ldots, p^{C_c}}_{|C_c^*|}), \quad \boldsymbol{q} = (\underbrace{p^{B_1}, \ldots, p^{B_1}}_{|B_1^*|}, \ldots, \underbrace{p^{B_b}, \ldots, p^{B_b}}_{|B_b^*|}).$$

If $C = \bigcup_i B_i$ then

$$p^C = \sum_i p^{B_i} \frac{|B_i|}{|C|},$$

so $\boldsymbol{p} \prec \boldsymbol{q}$ and invoking (4.4) completes the proof. □

Notice that in the case of censored observations the comparison holds in the convex order, whereas in the case of perfect observation a variance comparison holds, but Example 4.2 shows that comparisons in the convex order do not.

4.2 Univariate monotone functions

In the rest of this subsection the space $\mathfrak{U}$ is totally ordered, and without loss of generality we choose $\mathfrak{U} = [0,1]$. For subsets G and H of the real line, we write $G \leq H$ if $g \leq h$ for every $g \in G$ and $h \in H$. We call a partition $\mathscr{B} = (B_1, \dots, B_b)$ of $\mathfrak{U}$ monotone if $B_1 \leq \cdots \leq B_b$.

Theorem 4.4. *Let $\mathscr{B}$ and $\mathscr{C}$ be monotone partitions of $\mathfrak{U}$ and let $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$. If f is nondecreasing, then*

$$W_{\mathrm{IE}}^{\mathscr{B}} \leq_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}}. \tag{4.5}$$

To prove Theorem 4.4 we will apply the following lemma.

Lemma 4.5. *Let ξ and η be random variables such that $\xi \leq_{\mathrm{st}} \eta$, and let ξ_i and η_j be independent copies of ξ and η respectively. Let K be an integer valued random variable, independent of all ξ_j and η_j, satisfying $K \leq m$ for some integer m, and having an integer valued expectation, $\mathbb{E}[K] = k$. Then*

$$\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j \leq_{\mathrm{cx}} \sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j. \tag{4.6}$$

Proof. Since $\xi \leq_{\mathrm{st}} \eta$ we may construct i.i.d. pairs (ξ_i, η_i) with $\mathbb{P}(\xi_i \leq \eta_i) = 1$ for all $i = 1, \dots, m$. We adopt the usual convention that if $k = 0$ then $\sum_{j=1}^{k} \xi_j = 0$. First note that, by Wald's Lemma,

$$\mathbb{E}\left[\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j\right] = \mathbb{E}\left[\sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j\right].$$

Therefore (see, e.g., Müller and Stoyan, 2002, Theorem 1.5.3) it suffices to show that

$$\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j \leq_{\mathrm{icx}} \sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j.$$

Let ϕ be an increasing convex function and set

$$g(k) := \mathbb{E}\left[\phi\left(\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j\right)\right].$$

Note that

$$g(k) = \mathbb{E}\left[\phi\left(\sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j\right) \,\middle|\, K = k\right] \quad \text{and} \quad \mathbb{E}[g(K)] = \mathbb{E}\left[\phi\left(\sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j\right)\right].$$

Thus we have to show that $g(k) \leq \mathbb{E}[g(K)]$. Since $\mathbb{E}[K] = k$, this follows readily by Jensen's inequality, once we prove that $g(k)$ is a convex function.

The following part of the proof follows ideas of Ross and Schechner (1984). Setting

$$S_k = \sum_{j=1}^{k} \xi_j + \sum_{j=k+2}^{m} \eta_j,$$

We have

$$g(k+1) - g(k) = \mathbb{E}[\phi(\xi_{k+1} + S_k)] - \mathbb{E}[\phi(\eta_{k+1} + S_k)].$$

Since ϕ is convex, and $\xi_{k+1} \leq \eta_{k+1}$, the function

$$h(s) := \mathbb{E}[\phi(\xi_{k+1} + S_k) \mid S_k = s] - \mathbb{E}[\phi(\eta_{k+1} + S_k) \mid S_k = s]$$

is decreasing in s. Now note that

$$S_{k+1} = \sum_{i=1}^{k+1} \xi_i + \sum_{i=k+3}^{m} \eta_i \leq_{\text{st}} S_k = \sum_{i=1}^{k} \xi_i + \sum_{i=k+2}^{m} \eta_i$$

because $\xi_{k+1} \leq_{\text{st}} \eta_{k+2}$. Hence $g(k+1) - g(k) = \mathbb{E}[h(S_k)]$ is increasing in k, thus proving that g is convex, as required. □

Proof of Theorem 4.4. Since $\mathscr{B} = (B_1, \ldots, B_b)$ and $\mathscr{C} = (C_1, \ldots, C_c)$ are monotone partitions satisfying $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$ there exist $1 = i_1 < i_2 < \cdots < i_c < i_{c+1} = b+1$ such that

$$C_q = \bigcup_{j=i_q}^{i_{q+1}-1} B_j, \quad \text{for} \quad q = 1, \ldots, c.$$

As the union above may be formed by taking the union of two consecutive sets at a time, it suffices to prove (4.5) for the case where $c = b-1$, $C_m = B_m \cup B_{m+1}$, $C_k = B_k$ for $k \in \{1, \ldots, m-1\}$, and $C_k = B_{k+1}$ for $k \in \{m+1, \ldots, c\}$.

In this case we have

$$W_{\text{IE}}^{\mathscr{B}} = \frac{1}{n} \left[\sum_{C \neq C_m} \sum_{j \in C^*} f(V_j^C) + \sum_{j \in B_m^*} f(V_j^{B_m}) + \sum_{j \in B_{m+1}^*} f(V_j^{B_{m+1}}) + \sum_{j \in N} \varepsilon_j \right],$$

$$W_{\text{IE}}^{\mathscr{C}} = \frac{1}{n} \left[\sum_{C \neq C_m} \sum_{j \in C^*} f(V_j^C) + \sum_{j \in C_m^*} f(V_j^{C_m}) + \sum_{j \in N} \varepsilon_j \right].$$

Note that

$$\mathscr{L}\left(\sum_{j \in C_m^*} f\left(V_j^{C_m}\right) \right) = \mathscr{L}\left(\sum_{j=1}^{K} f\left(V_j^{B_m}\right) + \sum_{j=K+1}^{|C_m^*|} f\left(V_j^{B_{m+1}}\right) \right),$$

where K is binomially distributed with parameters

$$\left(|C_m^*|, \frac{|B_m^*|}{|C_m^*|}\right).$$

It is easy to see that if two variables are ordered by the convex order (see (2.1)) and we add the same independent variable to each one, to wit, $\sum_{j\in N} \varepsilon_j$, then the convex order is preserved. This fact and Lemma 4.5 now yield (4.5). □

4.3 Multivariate monotone functions

In this section we extend the results in Section 4.2 to the multivariate case. When we consider multivariate monotone functions, stratifying can still yield improvement in the convex order, but some restrictions are needed, both on the distribution of the random vector used for sampling and on the stratifying partitions. More specifically, we consider estimation of an integral with respect to a random vector whose components are independent, and under a stratification that preserves independence on each set of the partition. The result we prove below actually only requires that the random vector have an MTP_2 distribution (independence being a particular case of it), and that the stratification preserves MTP_2.

Let $f : [0,1]^d \to [0,1]$ be nondecreasing in each variable, and let $\boldsymbol{U}$ be a random vector taking values in $[0,1]^d$ with a nonatomic distribution. Our goal is to show that the estimate of $\mathbb{E}[f(\boldsymbol{U})]$ improves by refining stratifications as follows: recalling the definitions in Section 2, start with a partition $\mathscr{C} = (C_1, \dots, C_b)$ of $[0,1]^d$ such that for some i the distribution $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_i)$ is associated. Then split C_i into $C_i \cap G$ and $C_i \cap G^c$, where G is an increasing set. Lemma 4.8 below shows that the new partition obtained by this splitting achieves a better estimator of the integral in terms of the convex order, and Theorem 4.6 provides some conditions for its application.

Theorem 4.6. *Consider a partition $\mathscr{C} = (C_1, \dots, C_c)$ of $[0,1]^d$ where each C_i is a lattice. Let $\mathscr{B}$ be a partition obtained by a sequence of refinements $\mathscr{C} = \mathscr{C}_1 \le_{\mathrm{ref}} \cdots \le_{\mathrm{ref}} \mathscr{C}_m = \mathscr{B}$, such that for $k = 1, \dots, m-1$ the partition $\mathscr{C}_{k+1}$ is obtained from $\mathscr{C}_k$ by splitting one set of $\mathscr{C}_k$, say $C_{i_k,k}$, into $C_{i_k,k} \cap G_k$ and $C_{i_k,k} \cap G_k^c$, where $G_k = \{\boldsymbol{x} = (x_1, \dots, x_d) \in [0,1]^d : a_k \le x_j\}$ for some $a_k \in [0,1]$ and some $j \in \{1, \dots, d\}$.*

If $\boldsymbol{U}$ is MTP_2 on $[0,1]^d$ and $f : [0,1]^d \to [0,1]$ is nondecreasing, then $W_{\mathrm{IE}}^{\mathscr{B}} \le_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}}$.

As mentioned earlier, independence is a particular (and in our framework the most important) case of MTP_2. Independence makes simulation of a multivariate random vector easy, even when conditioned on an interval, since the strata can be constructed by knowing only the quantiles of the marginal distributions. If the cost of simulation is negligible relative to the cost of evaluating f, then even rejective sampling can be used, once the strata are defined.

The proof of Theorem 4.6 is preceded by the following lemmas.

Lemma 4.7. *If $\boldsymbol{U}$ is an associated random vector, and G is an increasing set, then*

$$\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in G^c) \leq_{\mathrm{st}} \mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in G). \tag{4.7}$$

Conversely, if (4.7) *holds for every increasing set G, then $\boldsymbol{U}$ is associated.*

Proof. First note that (4.7) is equivalent to

$$\mathbb{P}(\boldsymbol{U} \in A \,|\, \boldsymbol{U} \in G) \geq \mathbb{P}(\boldsymbol{U} \in A \,|\, \boldsymbol{U} \in G^c)$$

holding for all increasing sets A. The latter inequality is easily seen to be equivalent to

$$\mathbb{P}(\boldsymbol{U} \in A \cap G)[1 - \mathbb{P}(\boldsymbol{U} \in G)] \geq [\mathbb{P}(\boldsymbol{U} \in A) - \mathbb{P}(\boldsymbol{U} \in A \cap G)]\mathbb{P}(\boldsymbol{U} \in G).$$

By simple cancelation this inequality is equivalent to

$$\mathbb{P}(\boldsymbol{U} \in A \cap G) \geq \mathbb{P}(\boldsymbol{U} \in A)\mathbb{P}(\boldsymbol{U} \in G),$$

which is equivalent to association of the random vector $\boldsymbol{U}$ by e.g., Shaked (1982). □

Lemma 4.8. *Consider a partition $\mathscr{C} = (C_1, \ldots, C_c)$ of $[0,1]^d$ such that for some C_i the distribution $\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C_i)$ is associated. Let G be an increasing set and let $\mathscr{B} = (C_1, \ldots, C_{i-1}, C_i \cap G, C_i \cap G^c, C_{i+1}, \ldots, C_c)$. If $f : [0,1]^d \to [0,1]$ is nondecreasing, then $W_{\mathrm{IE}}^{\mathscr{B}} \leq_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}}$.*

Proof. With $\mathscr{L}(\boldsymbol{V}_1) = \mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C_i \cap G^c)$ and $\mathscr{L}(\boldsymbol{V}_2) = \mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C_i \cap G)$, Lemma 4.7 yields $\boldsymbol{V}_1 \leq_{\mathrm{st}} \boldsymbol{V}_2$. The monotonicity of f implies $f(\boldsymbol{V}_1) \leq_{\mathrm{st}} f(\boldsymbol{V}_2)$, and Lemma 4.5 now proves the claim, applying arguments as in the proof of Theorem 4.4. □

The following result can be found in Karlin and Rinott (1980).

Lemma 4.9. *If an MTP$_2$ vector $\boldsymbol{U}$ takes values in a lattice of which C is a sublattice, then $\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C)$ is MTP$_2$ and hence associated.*

The following corollary is obvious, and only requires the fact that the intersection of sublattices is a lattice.

Corollary 4.10. *If an MTP$_2$ vector $\boldsymbol{U}$ takes values in some lattice, and C, G and G^c, are all sublattices, then both $\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C \cap G)$ and $\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C \cap G^c)$ are MTP$_2$, and hence also associated.*

Proof of Theorem 4.6. We first prove by induction that $\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C_{i,k})$ are MTP$_2$ for all $C_{i,k} \in \mathscr{C}_k$ and $k = 1, \ldots, m$. For $k = 1$ this follows from Lemma 4.9 and the assumptions that $\boldsymbol{U}$ is MTP$_2$ and that $C_i = C_{i,1}$ are sublattices of $[0,1]^d$. Assuming the statement true for $1 \leq k < m$, to verify that it is true for $k+1$ we need only show that $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i_k,k} \cap G_k)$ and $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i_k,k} \cap G_k^c)$ are MTP$_2$, which follows from Lemma 4.9, thus completing the induction.

Hence, again using Lemma 4.9, $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i_k,k})$ is associated. Since G_k is increasing, Lemma 4.8 now yields $W_{\mathrm{IE}}^{\mathscr{C}_{k+1}} \leq_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}_k}$ for all $k = 1, \ldots, m-1$, and, therefore, the theorem. □

A sequence of partitions as in Theorem 4.6 can be generated as follows: start with the whole space $[0,1]^d$, then split it into boxes by repeatedly subdividing one element of the partition by an intersection with some G and G^c. In $[0,1]^2$ the resulting partition forms a tiling of the square by rectangles. Note that from the first step, a sequence of partitions created using G as above has at least one line which crosses the whole square from side to side. Therefore the tiling of Figure 1 is not attainable by such a sequence.

FIGURE 1 ABOUT HERE

Lastly, recall that the hypothesis of MTP$_2$ includes as a particular case the uniform distribution on $[0,1]^d$, so Theorem 4.6 applies to the estimation of the integral $\int f(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$ on $[0,1]^d$, or any lattice.

5 Numerical examples

In this section we estimate the integral of different functions with different levels of stratification. In all our examples the functions are defined on the unit square $[0,1]^2$ and the strata are obtained by recursively splitting the squares into squares of the same size. We consider six possible stratifications where the number of strata is 2^{2i} with $i \in \{0, \ldots, 5\}$ and each evaluation of the integral involves 1024 observations. Therefore when $i = 0$ no stratification is performed, and when $i = 5$ each stratum contains only one observation. This sequence of stratifications is performed according to the procedure used in Theorem 4.6. Observations are drawn uniformly from each stratum in $[0,1]^2$. Since a uniform distribution on an interval of $[0,1]^d$ is MTP$_2$, the hypotheses of Theorem 4.6 are satisfied. The simulation generates 10,000 estimates of the integral for each level of stratification. The numbers on the horizontal axis indicate the number of strata in all the figures that follow. All losses in the plots below are multiplied by 1024.

Figure 2 compares the L_1 loss in estimating the integral of two functions observed without error, under different stratifications. The first function f is of the type defined in (4.3) and is nonmonotone; the second function g is a slight alteration that makes it monotone. Specifically

$$f(u) = \begin{cases} 1 & \text{if } u \in A_a, \\ -1 & \text{if } u \in A_b, \\ 0 & \text{elsewhere,} \end{cases} \qquad g(u) = \begin{cases} 1 & \text{if } u \in A_a, \\ 1/2 & \text{if } u \in A_b, \\ 0 & \text{elsewhere,} \end{cases} \tag{5.1}$$

where $A_a = [63/64, 1] \times [31/32, 1]$ and $A_b = [62/64, 63/64) \times [31/32, 1]$.

FIGURE 2 ABOUT HERE

The maximal sample standard deviation of the simulations for line a is less than .002. The maximal sample standard deviation for line b is less than .004. Note that the loss increases in stratification for the nonmonotone function f. The increase in the loss going from 256 to 1024 strata is statistically significant. That the loss decreases in stratification for the monotone function g is a numerical validation of Theorem 4.6.

In Figures 3 to 11 all functions are observed with an error having standard deviation either zero (i.e., no error), or .1, or .25.

Figures 3, 4, and 5 show the L_1, L_2, and L_7 loss for the monotone continuous function $f(x, y) = ((1 + x)^y(1 + y) - 1)/3$ on $[0, 1]^2$. Going from no stratification to some stratification the estimation improves dramatically, especially when the function is observed without error. Further refinements of the stratification produce smaller benefits.

FIGURES 3, 4, AND 5 ABOUT HERE

Figure 6, 7, and 8 show the L_1, L_2, and L_7 loss for the monotone discontinuous function

$$f(x, y) = \frac{1}{5} I_{[1/2,1]}(x) I_{[7/10,1]}(y) + \frac{1}{2} I_{[3/5,1]}(x) I_{[3/10,1]}(y) + \frac{3}{2} I_{[7/10,1]}(x) I_{[9/10,1]}(y)$$

on $[0, 1]^2$. Going from no stratification to some stratification the estimation improves dramatically, especially when the function is observed without error. Unlike the continuous example above, even further refinements of the stratification produce sizable benefits. The improvement is smaller when the error has a higher variance.

FIGURES 6, 7, AND 8 ABOUT HERE

Figure 9, 10, and 11 show the L_1, L_2, and L_7 loss for the monotone discontinuous function

$$f(x, y) = \frac{1}{5} I_{[1/2,1]}(x) I_{[1/2,1]}(y) + \frac{1}{2} I_{[1/4,1]}(x) I_{[3/4,1]}(y) + \frac{3}{2} I_{[3/4,1]}(x) I_{[7/8,1]}(y)$$

on $[0, 1]^2$. Unlike the previous example, here, when we refine the stratification, at some point this function becomes constant on each stratum. Therefore any further refinement becomes useless, as the figures clearly show.

FIGURES 9, 10, AND 11 ABOUT HERE

Figure 12, 13, and 14 show the L_1, L_2, and L_7 loss for the nonmonotone discontinuous function

$$f(x,y) = \frac{1}{10} I_{[0,0.5]}(x) I_{[0,0.7]}(y) + \frac{1}{4} I_{[0.6,1]}(x) I_{[0.3,1]}(y) + \frac{3}{4} I_{[0.7,1]}(x) I_{[0.9,1]}(y)$$

on $[0,1]^2$, when the function is observed with censoring as in Theorem 4.3. As expected from this theorem, stratification is beneficial, even if the function is not monotone.

FIGURES 12, 13, AND 14 ABOUT HERE

A Appendix

Lemma A.1. *Given a partition $\mathscr{B}^*$ of N, consider a collection of independent random variables $\{\xi_j^{B^*}\}$, $B^* \in \mathscr{B}^*$, $j \in B^*$, with those indexed by the same element B^* of the partition being identically distributed.*

For $\mathscr{C}^* \leq_{\mathrm{ref}} \mathscr{B}^*$ let $\{\xi_j^{C^*}\}$ with $C^* \in \mathscr{C}^*$ and $j \in C^*$ be a collection of independent random variables with the mixture distribution

$$\mathscr{L}(\xi_j^{C^*}) = \sum_{B^* \subset C^*} \frac{|B^*|}{|C^*|} \mathscr{L}(\xi_j^{B^*}). \tag{A.1}$$

Then

$$\max_{C^* \in \mathscr{C}^*} \max_{j \in C^*} \xi_j^{C^*} \leq_{\mathrm{st}} \max_{B^* \in \mathscr{B}^*} \max_{j \in B^*} \xi_j^{B^*}. \tag{A.2}$$

Proof. Let $p^{B^*} = \mathbb{P}(\xi_1^{B^*} \leq t)$ for $B^* \in \mathscr{B}^*$, and $p^{C^*} = \mathbb{P}(\xi_1^{C^*} \leq t)$ for $C^* \in \mathscr{C}^*$.

We claim that

$$(\underbrace{p^{C_1^*}, \ldots, p^{C_1^*}}_{|C_1^*|}, \ldots, \underbrace{p^{C_c^*}, \ldots, p^{C_c^*}}_{|C_c^*|}) \prec (\underbrace{p^{B_1^*}, \ldots, p^{B_1^*}}_{|B_1^*|}, \ldots, \underbrace{p^{B_b^*}, \ldots, p^{B_b^*}}_{|B_b^*|}).$$

To see this, observe that (A.1) implies that the vector on the left-hand side above is obtained from the one on the right by multiplying it by the $n \times n$ doubly stochastic matrix $\boldsymbol{D}$ which is block diagonal where the i-th block is the $|C_i^*| \times |C_i^*|$ matrix with all entries equal to $1/|C_i^*|$.

Hence, by the Schur concavity of the function $(\theta_1, \ldots, \theta_n) \mapsto \prod_{i=1}^n \theta_i$, we have

$$\mathbb{P}\left(\max_{C^* \in \mathscr{C}^*} \max_{j \in C^*} \xi_j^{C^*} \leq t \right) = \prod_{C^* \in \mathscr{C}^*} (p^{C^*})^{|C^*|} \geq \prod_{B^* \in \mathscr{B}^*} (p^{B^*})^{|B^*|} = \mathbb{P}\left(\max_{B^* \in \mathscr{B}^*} \max_{j \in B^*} \xi_j^{B^*} \leq t \right),$$

which is equivalent to (A.2). □

Proof of Theorem 3.1. Let $\mathscr{B}^*$ and $\mathscr{C}^*$ be partitions associated with $\mathscr{B}$ and $\mathscr{C}$, respectively, satisfying $\mathscr{C}^* \leq_{\text{ref}} \mathscr{B}^*$, and let $\{\xi_j^{B^*}, B^* \in \mathscr{B}^*, j \in B^*\}$ and $\{\xi_j^{C^*}, C^* \in \mathscr{C}^*, j \in C^*\}$ be collections of independent random variables with distributions

$$\begin{aligned}\mathbb{P}(\xi_j^{B^*} \leq t) &= \mathbb{P}(f(U) \leq t \,|\, U \in B)\\ \mathbb{P}(\xi_j^{C^*} \leq t) &= \mathbb{P}(f(U) \leq t \,|\, U \in C).\end{aligned}$$

Then (A.1) holds (law of total probability), and the result follows by Lemma A.1. □

Proof of Theorem 4.1. In what follows we consider conditional expectation with respect to a partition. Though the notion is standard, specifically, by $\mathbb{E}[f(U)+\varepsilon \,|\, \mathscr{B}]$ we mean the random variable that takes values $\overline{f}_B := \mathbb{E}[f(U) \,|\, U \in B]$ with probability $|B^*|/n$. Then

$$\begin{aligned}\mathrm{Var}[f(U)+\varepsilon \,|\, \mathscr{B}] &= \mathbb{E}\left[\{f(U)+\varepsilon - \mathbb{E}[f(U)+\varepsilon \,|\, \mathscr{B}]\,\}^2 \,|\, \mathscr{B}\right]\\ &= \mathbb{E}\left[\{f(U)+\varepsilon - \mathbb{E}[f(U) \,|\, \mathscr{B}]\}^2 \,|\, \mathscr{B}\right]\end{aligned}$$

is a random variable taking values $\mathbb{E}\left[\left(f(U)+\varepsilon-\overline{f}_B\right)^2 \,|\, U \in B\right]$ with probability $|B^*|/n$, and

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[f(U)+\varepsilon \,|\, \mathscr{B}]] &= \sum_{B\in\mathscr{B}} \frac{|B^*|}{n}\mathbb{E}\left[(f(U)+\varepsilon-\overline{f}_B)^2 \,|\, U \in B\right]\\ &= \frac{1}{n}\sum_{B\in\mathscr{B}} |B^*|\,\mathbb{E}\left[(f(V_1^B)+\varepsilon-\overline{f}_B)^2\right]\\ &= \frac{1}{n}\,\mathrm{Var}\left[\sum_{B\in\mathscr{B}}\sum_{j\in B_i^*} f(V_j^B)+\varepsilon_j^B\right]\\ &= n\,\mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{B}}].\end{aligned}$$

If $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$, then for any random variable Y, say, $\mathrm{Var}[\mathbb{E}[Y \,|\, \mathscr{B}]] \geq \mathrm{Var}[\mathbb{E}[Y \,|\, \mathscr{C}]]$ by Jensen's inequality, and now the usual variance decomposition of Y (see, e.g., Rosenthal, 2006, Theorem 13.3.1) implies $\mathbb{E}[\mathrm{Var}[Y \,|\, \mathscr{B}]] \leq \mathbb{E}[\mathrm{Var}[Y \,|\, \mathscr{C}]]$. Therefore

$$\mathbb{E}[\mathrm{Var}[f(U)+\varepsilon \,|\, \mathscr{B}]] \leq \mathbb{E}[\mathrm{Var}[f(U)+\varepsilon \,|\, \mathscr{C}]],$$

and hence

$$\mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{B}}] = \frac{1}{n}\mathbb{E}[\mathrm{Var}[f(U)+\varepsilon \,|\, \mathscr{B}]] \leq \frac{1}{n}\mathbb{E}[\mathrm{Var}[f(U)+\varepsilon \,|\, \mathscr{C}]] = \mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{C}}].$$

□

Acknowledgments

We thank Abram Kagan for sparking our curiosity in the topic with a simple version of Theorem 3.1, Erich Novak for an important bibliographical reference, and Pierpaolo Brutti for his help with R. We are indebted to the Editor, an Associate Editor, and three referees for their accurate reading of the paper and their helpful comments.

References

Bai, S. K. and Durairajan, T. M. (1997) Optimal equivariant estimator with respect to convex loss function. *J. Statist. Plann. Inference* **64**, 283–295.

Berger, J. O. (1976) Admissibility results for generalized Bayes estimators of coordinates of a location vector. *Ann. Statist.* **4**, 334–356.

Blackwell, D. (1951) Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, 93–102. University of California Press, Berkeley and Los Angeles.

Blackwell, D. (1953) Equivalent comparisons of experiments. *Ann. Math. Statistics* **24**, 265–272.

Eberl, Jr., W. (1984) On unbiased estimation with convex loss functions. *Statist. Decisions* 177–192.

Ermakov, S. M., Zhiglyavskiĭ, A. A., and Kondratovich, M. V. (1988) Reduction of a problem of random estimation of an extremum of a function. *Dokl. Akad. Nauk SSSR* **302**, 796–798.

Glasserman, P. (2004) *Monte Carlo Methods in Financial Engineering.* Springer-Verlag, New York.

Karlin, S. and Novikoff, A. (1963) Generalized convex inequalities. *Pacific J. Math.* **13**, 1251–1279.

Karlin, S. and Rinott, Y. (1980) Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10**, 467–498.

Kondratovich, M. and Zhigljavsky, A. (1998) Comparison of independent and stratified sampling schemes in problems of global optimization. In *Monte Carlo and Quasi-Monte Carlo Methods 1996 (Salzburg)*, 292–299. Springer, New York.

Kozek, A. (1977) Efficiency and Cramér-Rao type inequalities for convex loss functions. *J. Multivariate Anal.* **7**, 89–106.

Laycock, P. J. (1972) Convex loss applied to design in regression problems. *J. Roy. Statist. Soc. Ser. B* **34**, 148–170, 170–186.

Laycock, P. J. and Silvey, S. D. (1968) Optimal designs in regression problems with a general convex loss function. *Biometrika* **55**, 53–66.

Lin, P. E. and Mousa, A. (1982) Proper Bayes minimax estimators for a multivariate normal mean with unknown common variance under a convex loss function. *Ann. Inst. Statist. Math.* **34**, 441–456.

Marshall, A. W. and Olkin, I. (1979) *Inequalities: Theory of Majorization and Its Applications.* Academic Press Inc., New York.

Müller, A. and Stoyan, D. (2002) *Comparison Methods for Stochastic Models and Risks.* John Wiley & Sons Ltd., Chichester.

Novak, E. (1988) *Deterministic and Stochastic Error Bounds in Numerical Analysis.* Springer-Verlag, Berlin.

Papageorgiou, A. (1993) Integration of monotone functions of several variables. *J. Complexity* **9**, 252–268.

Petropoulos, C. and Kourouklis, S. (2001) Estimation of an exponential quantile under a general loss and an alternative estimator under quadratic loss. *Ann. Inst. Statist. Math.* **53**, 746–759.

Rosenthal, J. S. (2006) *A First Look at Rigorous Probability Theory.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition.

Ross, S. M. and Schechner, Z. (1984) Some reliability applications of the variability ordering. *Oper. Res.* **32**, 679–687.

Shaked, M. (1982) A general theory of some positive dependence notions. *J. Multivariate Anal.* **12**, 199–218.

Shaked, M. and Shanthikumar, J. G. (2007) *Stochastic Orders.* Springer, New York.

Zhigljavsky, A. and Žilinskas, A. (2008) *Stochastic Global Optimization.* Springer, New York.

Zhigljavsky, A. A. and Chekmasov, M. V. (1996) Comparison of independent, stratified and random covering sample schemes in optimization problems. *Math. Comput. Modelling* **23**, 97–110.

May 27, 2010

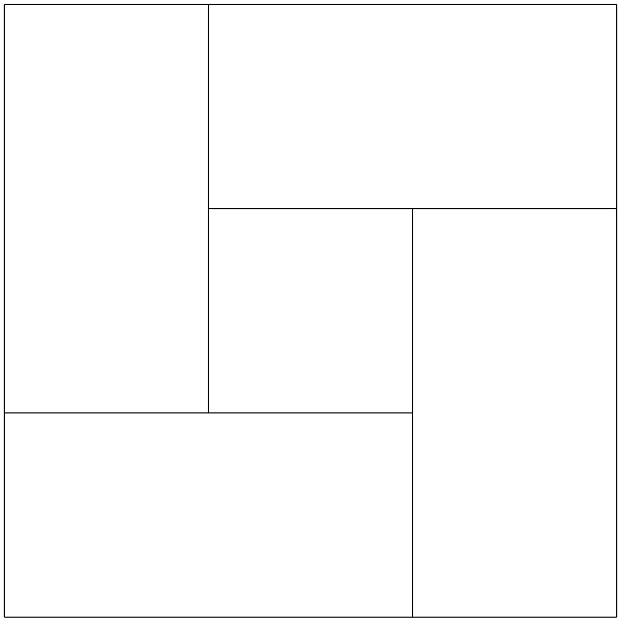

Figure 1: Non-attainable tiling.

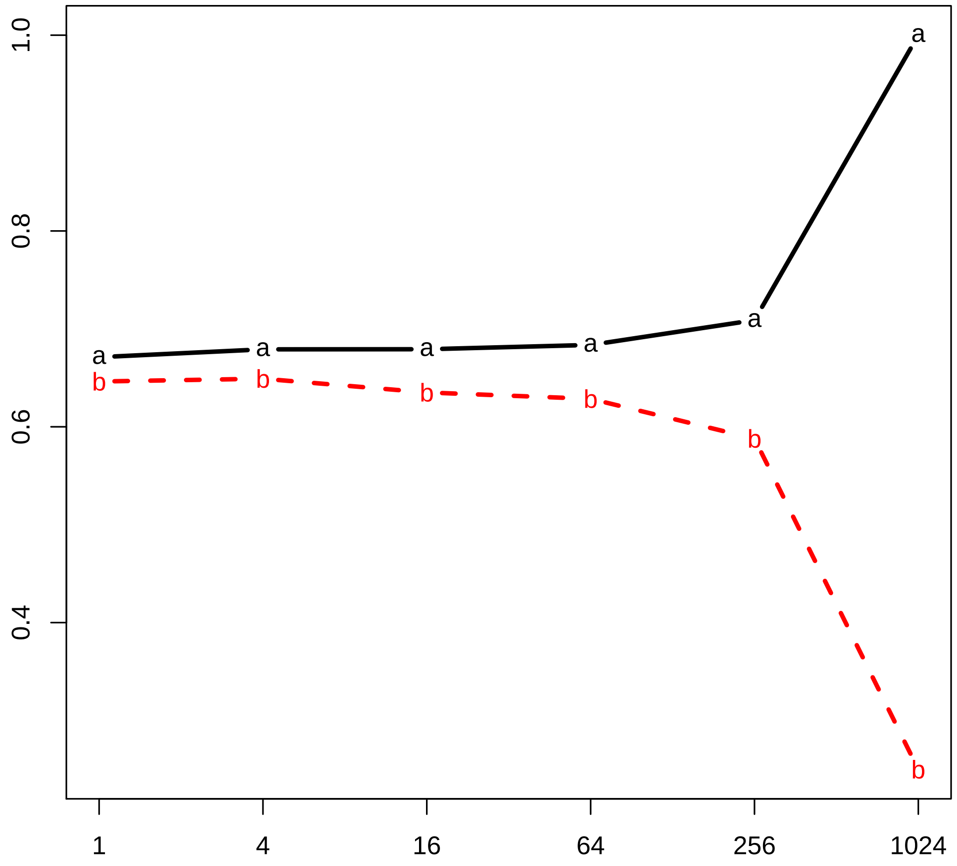

Figure 2: L_1 loss of function f (line a) and of function g (line b), with f and g defined as in (5.1), for different numbers of strata.

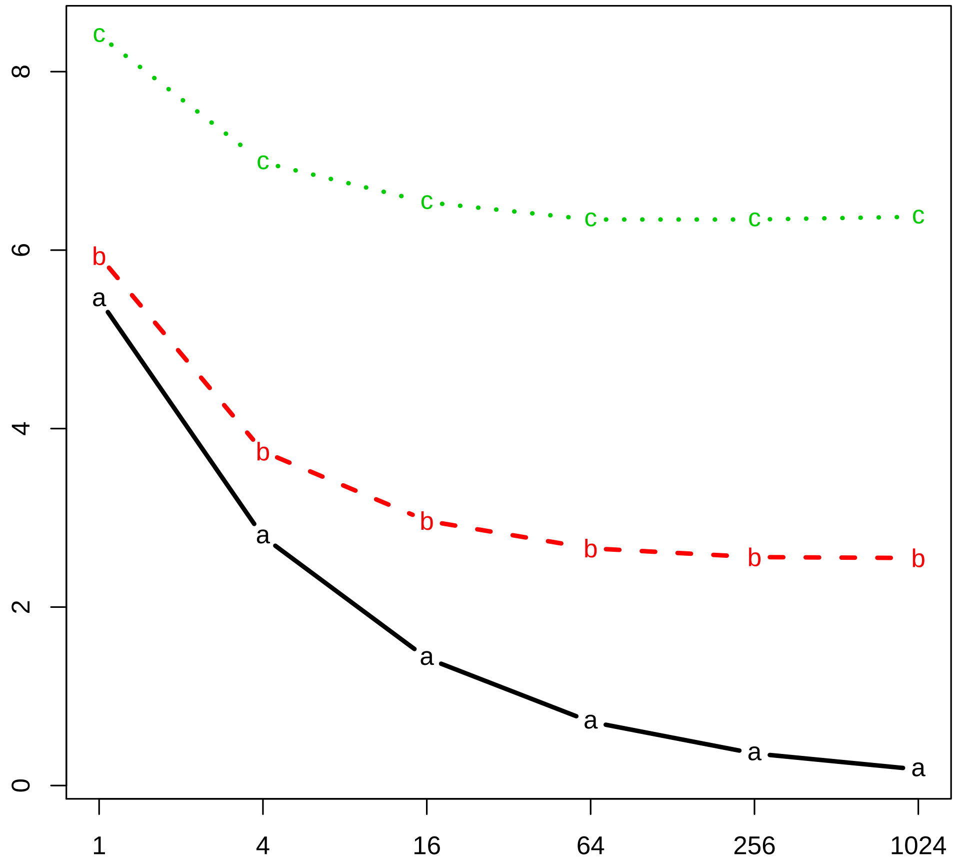

Figure 3: L_1 loss for $f(x, y) = ((1 + x)^y(1 + y) - 1)/3$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

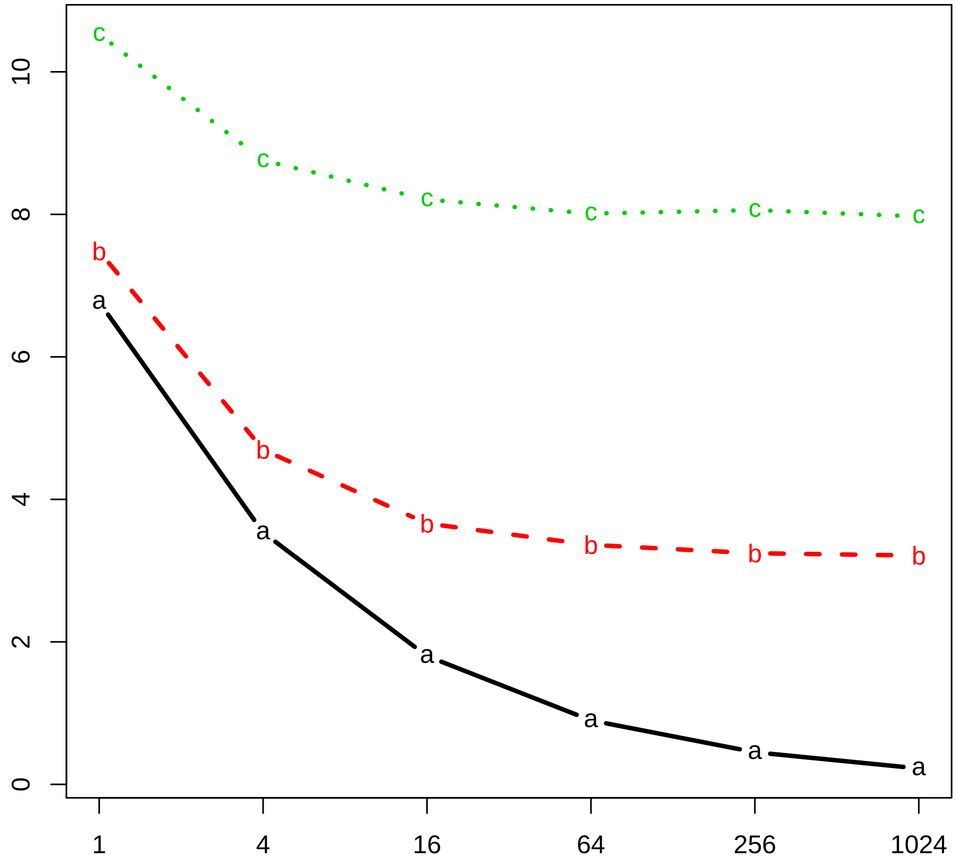

Figure 4: L_2 loss for $f(x,y) = ((1+x)^y(1+y)-1)/3$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

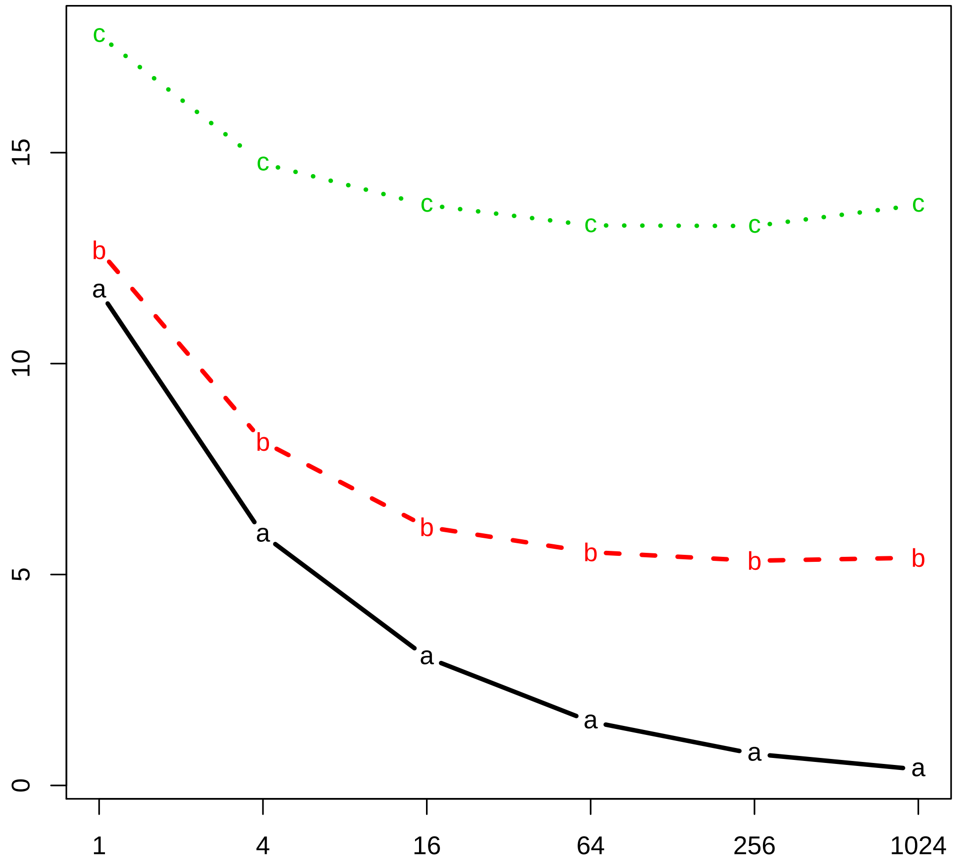

Figure 5: L_7 loss for $f(x, y) = ((1 + x)^y(1 + y) - 1)/3$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

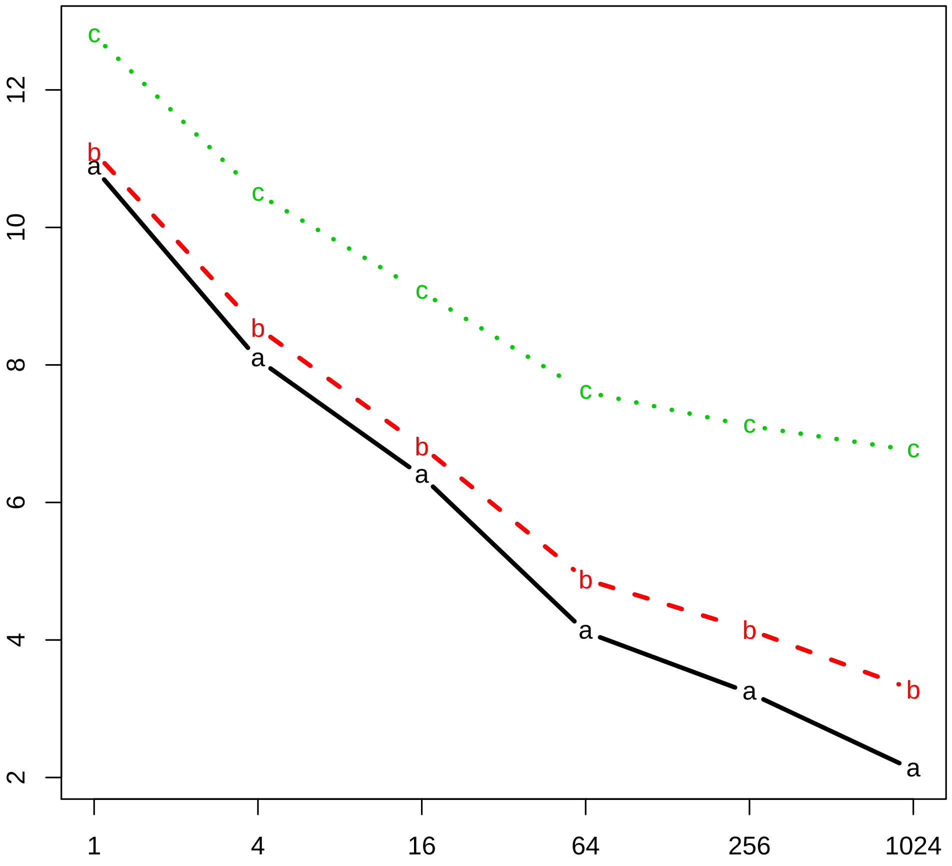

Figure 6: L_1 loss for $f(x,y) = \frac{2}{10}I_{[1/2,1]}(x)I_{[7/10,1]}(y) + \frac{1}{2}I_{[3/5,1]}(x)I_{[3/10,1]}(y) + \frac{3}{2}I_{[7/10,1]}(x)I_{[9/10,1]}(y)$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

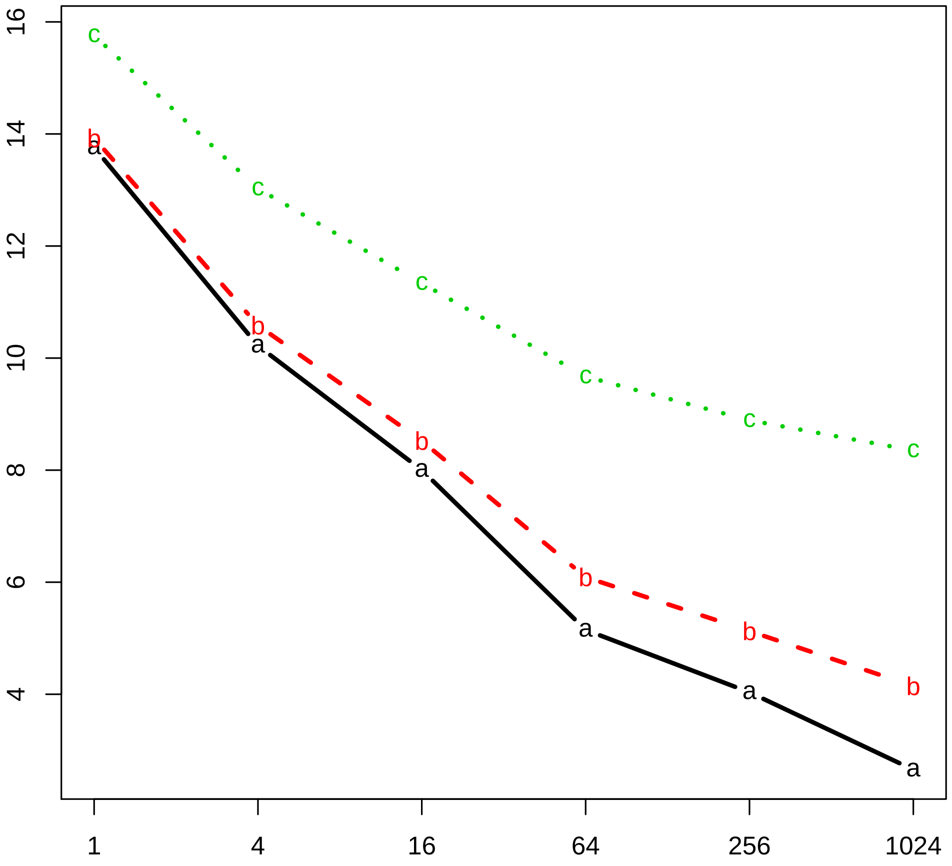

Figure 7: L_2 loss for $f(x,y) = \frac{1}{5}I_{[1/2,1]}(x)I_{[7/10,1]}(y) + \frac{1}{2}I_{[3/5,1]}(x)I_{[3/10,1]}(y) + \frac{3}{2}I_{[7/10,1]}(x)I_{[9/10,1]}(y)$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

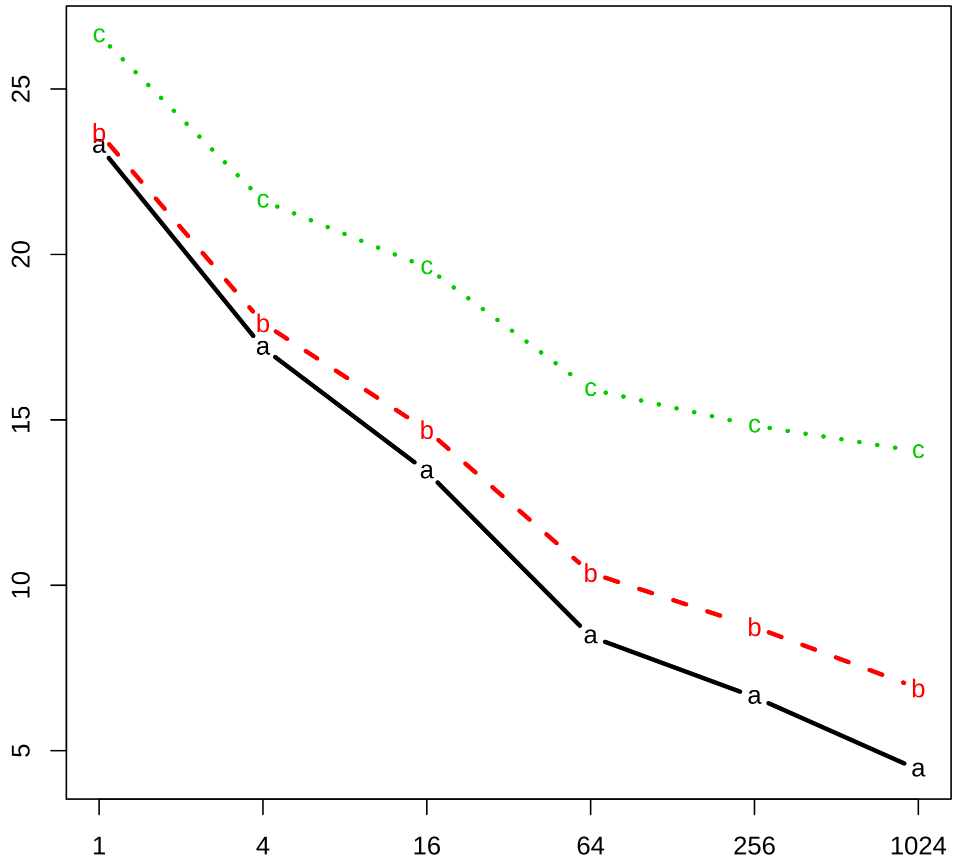

Figure 8: L_7 loss for $f(x,y) = \frac{1}{5}I_{[1/2,1]}(x)I_{[7/10,1]}(y) + \frac{1}{2}I_{[3/5,1]}(x)I_{[3/10,1]}(y) + \frac{3}{2}I_{[7/10,1]}(x)I_{[9/10,1]}(y)$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

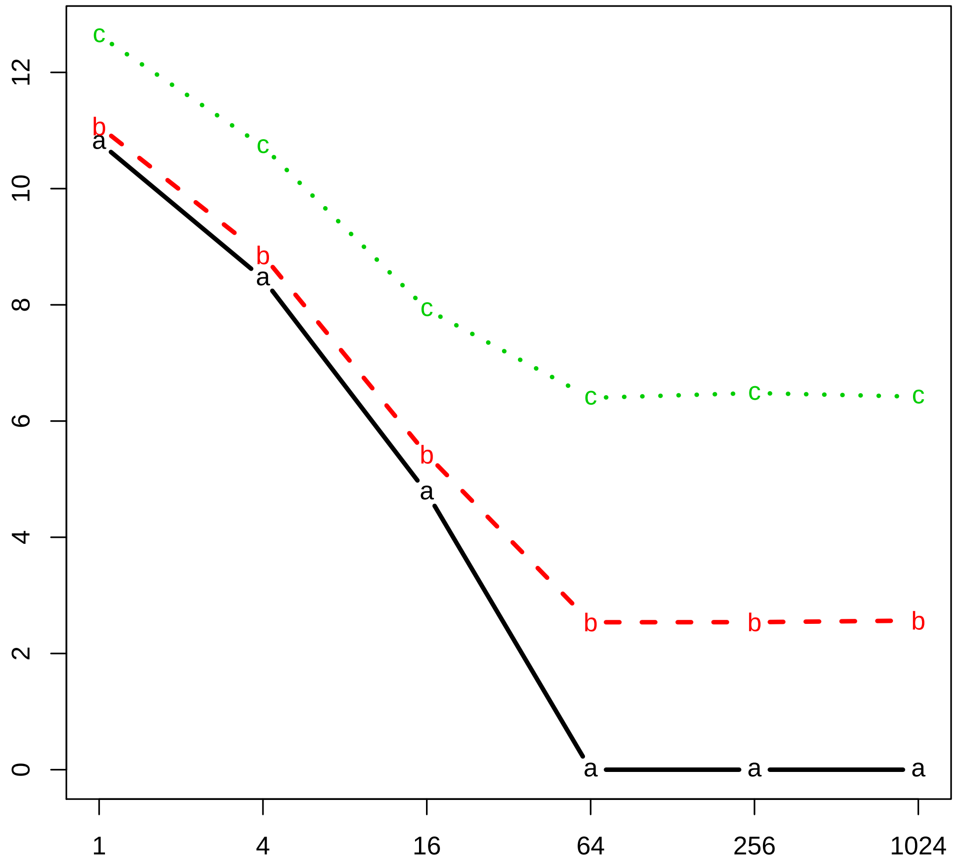

Figure 9: L_1 loss for $f(x, y) = \frac{1}{5}I_{[1/2,1]}(x)I_{[1/2,1]}(y) + \frac{1}{2}I_{[1/4,1]}(x)I_{[3/4,1]}(y) + \frac{3}{2}I_{[3/4,1]}(x)I_{[7/8,1]}(y)$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

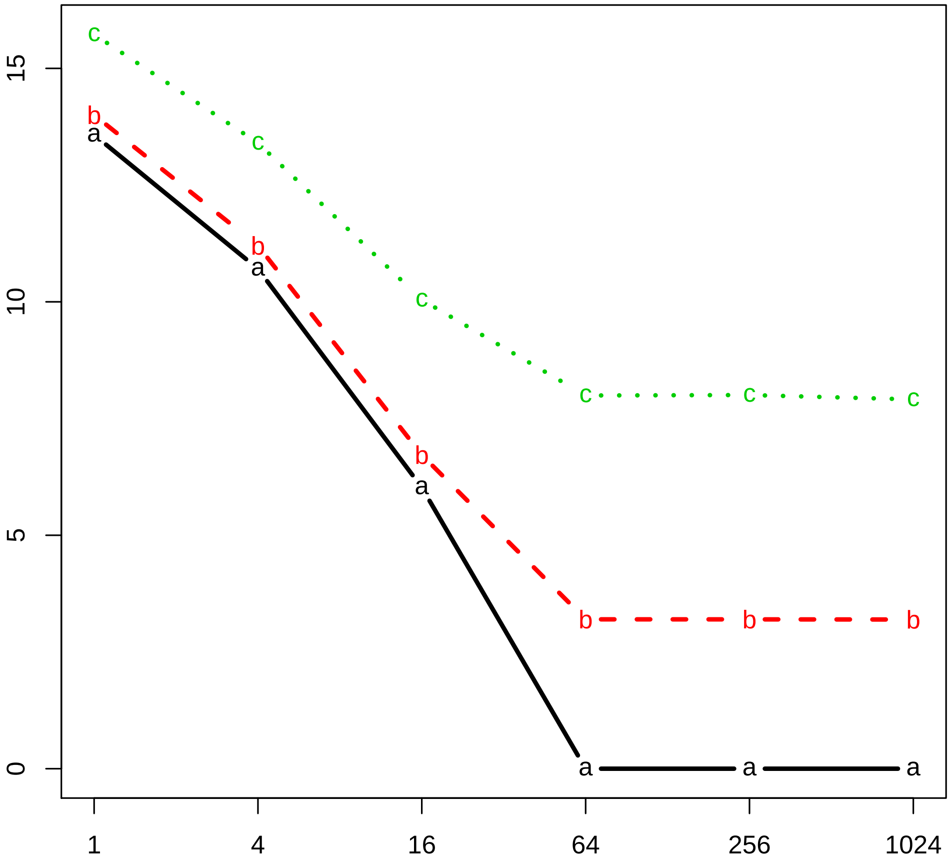

Figure 10: L_2 loss for $f(x,y) = \frac{1}{5}I_{[1/2,1]}(x)I_{[1/2,1]}(y) + \frac{1}{2}I_{[1/4,1]}(x)I_{[3/4,1]}(y) + \frac{3}{2}I_{[3/4,1]}(x)I_{[7/8,1]}(y)$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

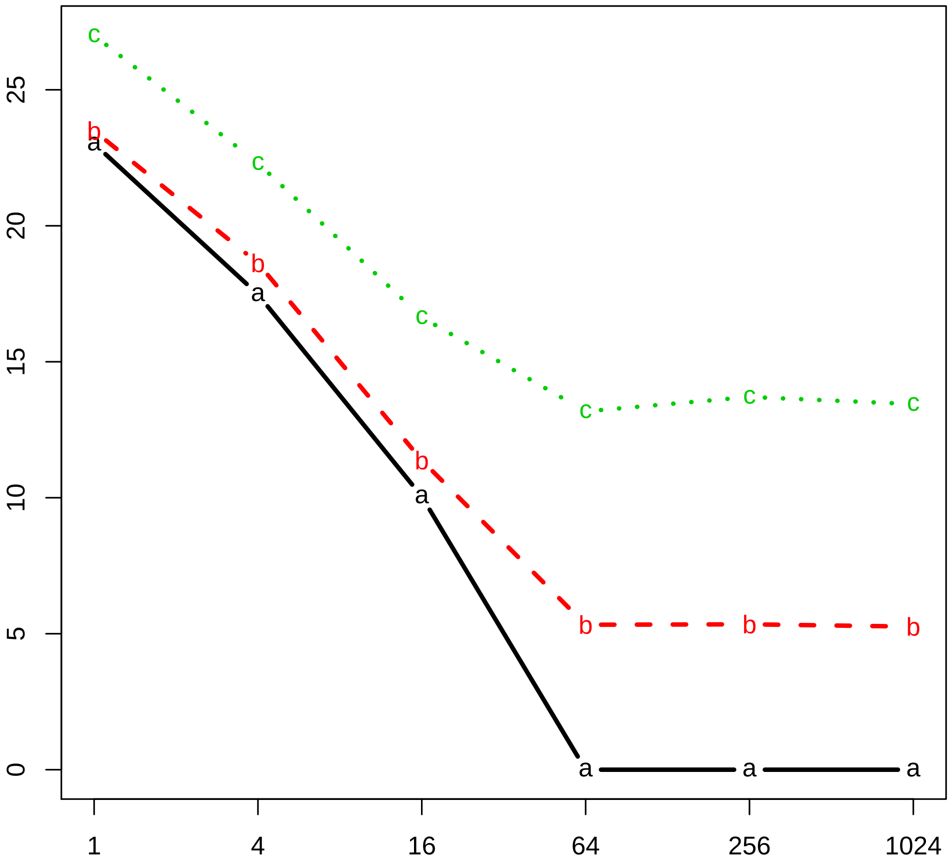

Figure 11: L_7 loss for $f(x,y) = \frac{1}{5}I_{[1/2,1]}(x)I_{[1/2,1]}(y) + \frac{1}{2}I_{[1/4,1]}(x)I_{[3/4,1]}(y) + \frac{3}{2}I_{[3/4,1]}(x)I_{[7/8,1]}(y)$ when the standard deviation of the error is 0 (line a), 0.1 (line b), or 0.25 (line c), for different numbers of strata.

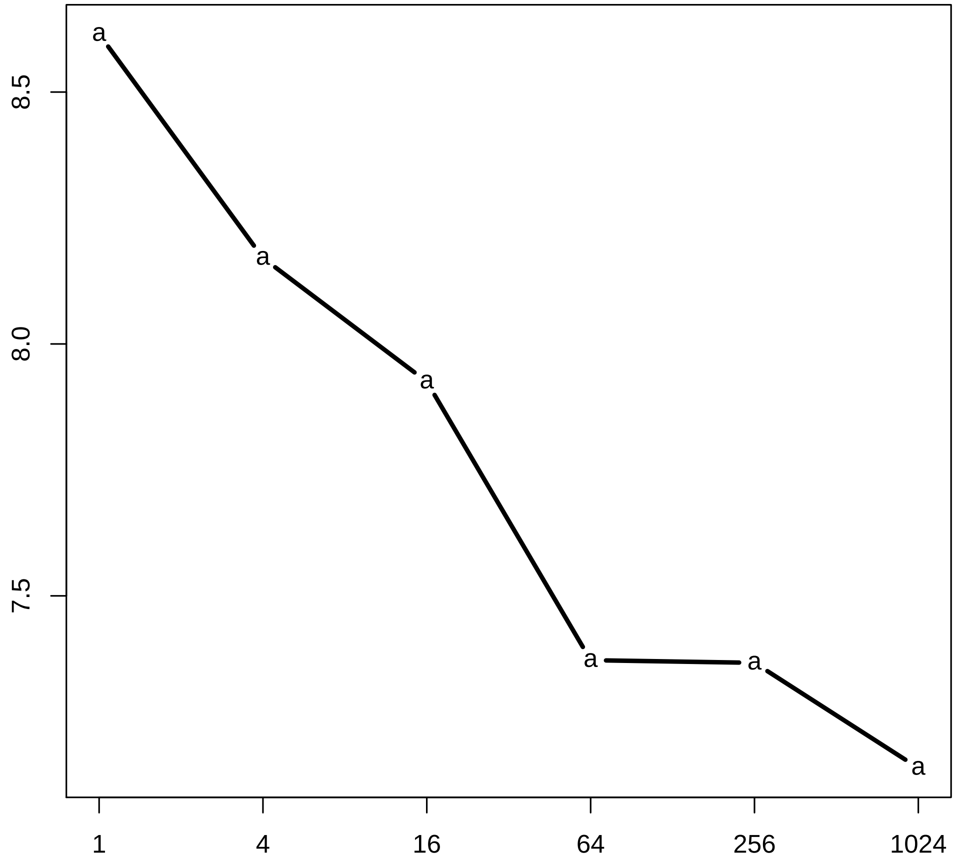

Figure 12: L_1 loss for $f(x,y) = \frac{1}{10}I_{[0,0.5]}(x)I_{[0,0.7]}(y) + \frac{1}{4}I_{[0.6,1]}(x)I_{[0.3,1]}(y) + \frac{3}{4}I_{[0.7,1]}(x)I_{[0.9,1]}(y)$, with censoring, for different numbers of strata.

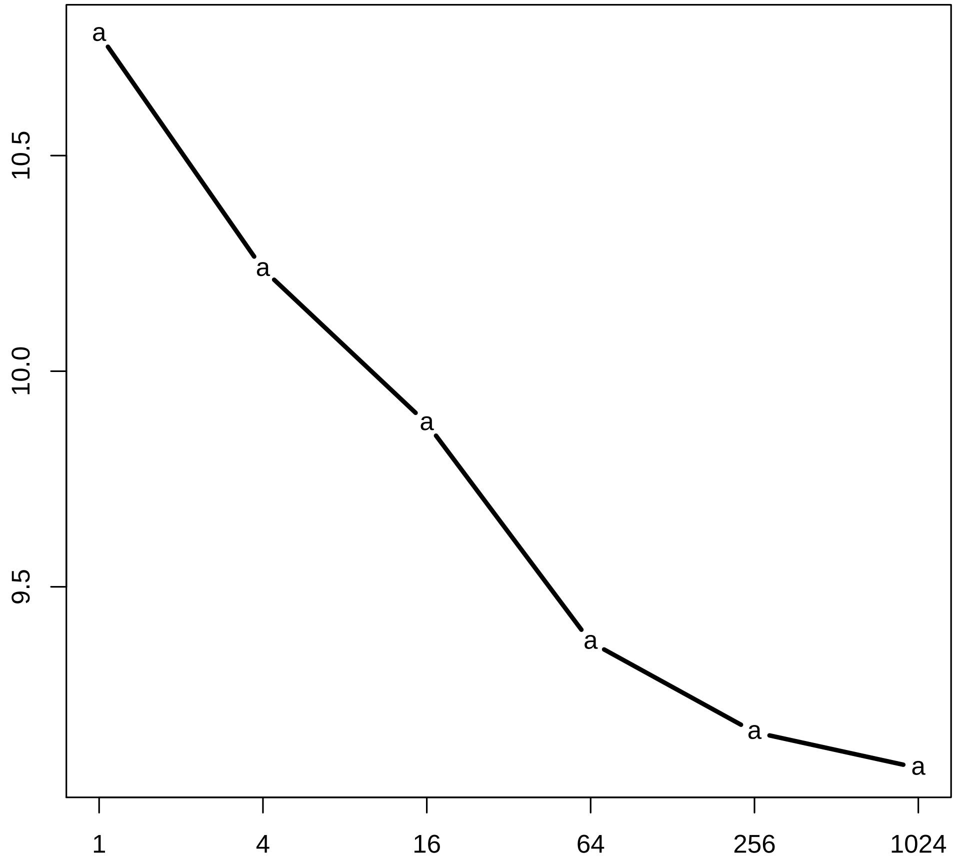

Figure 13: L_2 loss for $f(x,y) = \frac{1}{10}I_{[0,0.5]}(x)I_{[0,0.7]}(y) + \frac{1}{4}I_{[0.6,1]}(x)I_{[0.3,1]}(y) + \frac{3}{4}I_{[0.7,1]}(x)I_{[0.9,1]}(y)$, with censoring, for different numbers of strata.

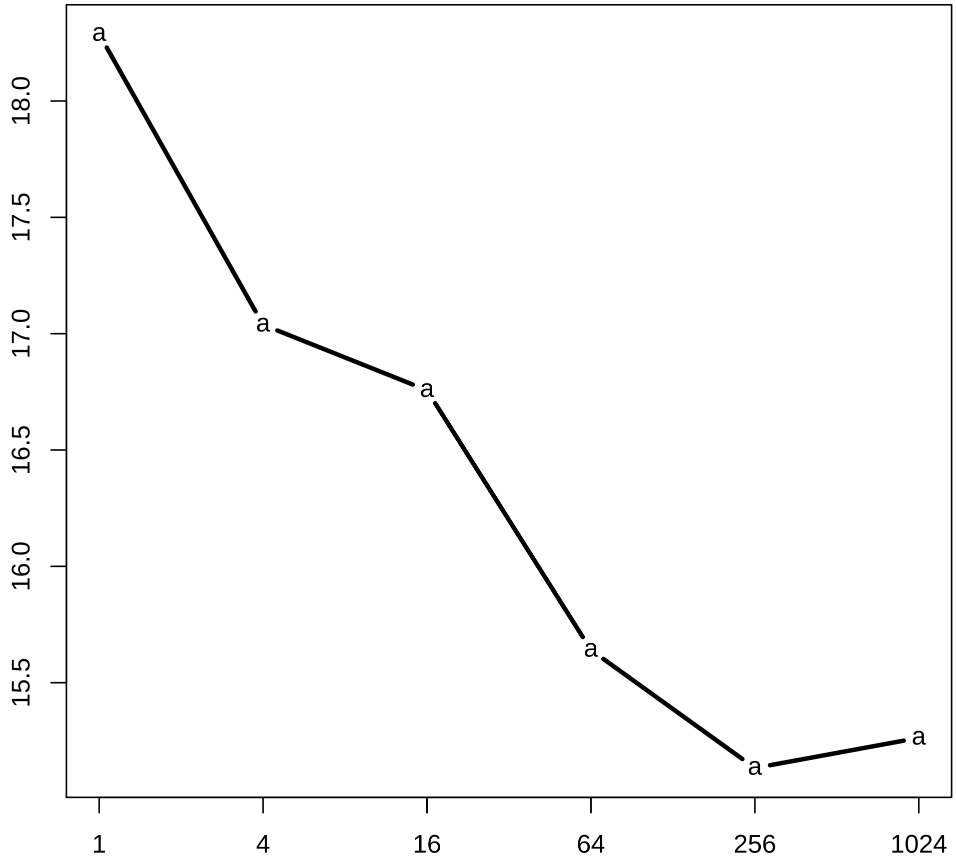

Figure 14: L_7 loss for $f(x,y) = \frac{1}{10}I_{[0,0.5]}(x)I_{[0,0.7]}(y) + \frac{1}{4}I_{[0.6,1]}(x)I_{[0.3,1]}(y) + \frac{3}{4}I_{[0.7,1]}(x)I_{[0.9,1]}(y)$, with censoring, for different numbers of strata.