Variance Competitiveness for Monotone Estimation: Tightening the Bounds

Edith Cohen*

Abstract

Random samples are extensively used to summarize massive data sets and facilitate scalable analytics. Coordinated sampling, where samples of different data sets "share" the randomization, is a powerful method which facilitates more accurate estimation of many aggregates and similarity measures.

We recently formulated a model of *Monotone Estimation Problems* (MEP), which can be applied to coordinated sampling, projected on a single item. MEP estimators can then be used to estimate sum aggregates, such as distances, over coordinated samples. For MEP, we are interested in estimators that are unbiased and nonnegative. We proposed *variance competitiveness* as a quality measure of estimators: For each data vector, we consider the minimum variance attainable on it by an unbiased and nonnegative estimator. We then define the competitiveness of an estimator as the maximum ratio, over data, of the expectation of the square to the minimum possible. We also presented a general construction of the L^* estimator, which is defined for any MEP for which a nonnegative unbiased estimator exists, and is at most 4-competitive.

Our aim here is to obtain tighter bounds on the *universal ratio*, which we define to be the smallest competitive ratio that can be obtained for any MEP. We obtain an upper bound of 3.375, improving over the bound of 4 of the L* estimator. We also establish a lower bound of 1.44. The lower bound is obtained by constructing the *optimally competitive* estimator for particular MEPs. The construction is of independent interest, as it facilitates estimation with instance-optimal competitiveness.

1 Introduction

We consider sampling schemes where the randomization is captured by a single parameter, u, which we refer to as the *seed*. The sampling scheme is specified by a data domain $\mathbf{V}$, the *seed* $u \in [0,1]$, and a function $S(u,\boldsymbol{v})$, which maps a data $\boldsymbol{v}$ and a value of u to the *sample* (or the outcome). For each outcome S (and seed u), we can consider the set

$$S^* = \{\boldsymbol{z} \in \mathbf{V} \mid S(u,\boldsymbol{z}) = S(u,\boldsymbol{v})\}$$

of all data vectors $\boldsymbol{v} \in \mathbf{V}$, which are consistent with S. The set $S^*(u,\boldsymbol{v})$ must clearly include $\boldsymbol{v}$, but can include many, a possibly infinite number of data vectors.

We recently introduced a framework of monotone sampling and estimation [7], which we briefly review and motivate here.

We say that the sampling scheme of the form above is *monotone* if for all $\boldsymbol{v}$, the function $S^*(u,\boldsymbol{v})$ is non-decreasing with u. The sample S can be interpreted as a lossy measurement of the data $\boldsymbol{v}$, which

*Microsoft Research, Mountain View, CA, USA editco@microsoft.com

provides some partial information, where the seed u is the granularity of our measurement instrument, the lower u is, the more we know on the data.

A *monotone estimation problem* (MEP) is defined by a monotone sampling scheme together with a function $f : \mathbf{V} \geq 0$, which we would like to estimate. The estimator $\hat{f}(S)$ is applied to the sample S (or equivalently, depends on the set S^*) and we require that for all $\boldsymbol{v}$ and u, $\hat{f}(S(u, \boldsymbol{v})) \geq 0$ (estimator is nonnegative), and for all $\boldsymbol{v}$, $\mathsf{E}_{u \sim U[0,1]}[\hat{f}(S(u, \boldsymbol{v}))] = f(\boldsymbol{v})$ (estimator is unbiased).

The main motivation for MEP comes from coordinated sampling, which dates back to Brewer et al [2], and was extensively applied in Statistics and Computer Science [15, 5, 13, 14, 4, 3, 8, 12, 10]. With coordinated sampling, we can think of our data set as a nonnegative matrix $\{v_{ij}\}$. Each row corresponds to a different time or location and each column to a particular key or feature. The data set is sampled, for example, each row can be Reservoir sampled, or Probability Proportional to Size (PPS) sampled, so only a small number of nonnegative entries are retained. With coordinated sampling, the samples of different rows utilize the same randomization (this can be achieved by applying a hash function to each key ID). As a result, we obtain the property that the samples of different rows are more similar when the data is. This property is also known as the Locality Sensitive Hashing (LSH) property. One of the benefits of coordinating the samples of rows is that it facilitates tighter estimates of many important functions of multiple rows, such as their similarity.

We are interested in estimating *sum aggregates* from the sampled data. A sum aggregate is a sum over selected keys (columns) of some function f of the values the key assumes in one or more instances.

An example of an aggregate that is of particular importance in data analysis is L_p^p of two rows, which is the sum over keys of the exponentiated range function, $\textsc{rg}_p$, which is the absolute difference between their values in the two instances, raised to the power of $p > 0$. The L_p distance, an extensively used distance measure, is the pth root of L_p^p. We studied estimation of $\textsc{rg}_p$ (over coordinated and independent samples) in [6].

We can estimate such sums by considering each key (column) separately. When the samples of different rows are coordinated, we obtain a simpler estimation problem for each key: If $\boldsymbol{v} = (v_{1h}, v_{2h}, \ldots, v_{rh})$ are the values the key h assumes in different rows, we would like to estimate $f(\boldsymbol{v})$ from the sample. For example, to estimate L_p^p, we estimate $f(\boldsymbol{v}) = \textsc{rg}_p(\boldsymbol{v})$ for each selected key, and then sum the estimates. Note that in this framework, the estimate is 0 on keys for which we have no information, we therefore, similar to estimation from a single set (row), we only need to actively compute the estimator for keys that are sampled in at least one row. This allows the computation of the estimate of the sum aggregate to be scalable.

When the sample of different rows are *coordinated*, the problem for each key is a MEP. We are often interested in nonnegative f, which means that we favor nonnegative estimators $\hat{f}$. Since our main application is estimating sums, and the variance will be high for a typical key (since we typically have no or little information on the values), unbiased estimators are desirable.

Classic point estimation theory studies estimating the parameter(s) θ of a distribution the data was drawn from, and estimators and quality measures of estimators are extensively studied since the time of Gauss [11]. A *risk* function which assigns cost for the deviation of the estimator from the true value is used. A popular risk function is the expected squared error (which for unbiased estimators is the variance). The risk, however, depends on the parameter. Ideally, we would want a Uniform Minimum Variance Unbiased Estimator (UMVUE), which minimizes variance for all parameter values. In reality, for most estimation problems, including monotone estimation in general, a UMVUE does not exist. Instead, we desire to somehow balance the risk across possible parameter values. In estimation theory, the main approaches either assume a distribution over θ and minimizing the average risk, or a Minimax estimator which minimizes the maximum risk. Either way, we are interested in an admissible (Pareto optimal) estimator, which means that it can not be strictly improved, attaining strictly lower risk for some parameters without increasing the risk on some others.

In our MEP formulation, which is suited for data analysis from samples, there are no distribution assumptions. Instead, we estimate an arbitrary function of the data $\boldsymbol{v}$. As we explained earlier, we only consider unbiased nonnegative estimators, since we are interested in sum estimators and nonnegative functions. The risk function we work with is the squared error, which since we only consider unbiased estimators, is the same as the variance.

For MEP, a UMVUE generally does not exist and we similarly aim to "balance" performance over the domain. We recently proposed a relative "risk" measure, inspired by *competitive analysis* in theoretical computer science [1]. For each vector $\boldsymbol{v}$, we consider the minimum variance attainable by a nonnegative unbiased estimator. We then define the variance ratio of an estimator to be the maximum over $\boldsymbol{v} \in \mathbf{V}$, of the expectation of the square of the estimator to the minimum possible by an unbiased nonnegative estimator. That is, instead of considering the absolute deviation, we compare performance,point wise, to the best possible. The advantage of competitiveness over the more classic measures is that it does not require distribution assumptions on the data and also that it captures performance in a way that is loosens up the dependence on the magnitude of f. A property that is not critical in the parameter estimation setting but is important in our setting.

Surprisingly perhaps, we learned that we can characterize this point-wise variance optimum for any MEP and data $\boldsymbol{v} \in \mathbf{V}$ [9, 7]. Moreover, for any MEP for which an unbiased nonnegative estimator with finite variances exists, there exists an estimator with a constant ratio [9]. In [7] we presented a particularly natural estimator, the L* estimator, that is guaranteed to be 4 competitive. We also showed that the ratio of 4 is tight for the L* estimator: For any $\epsilon > 0$, there is an MEP for which the L* estimator has ratio $\geq 4 - \epsilon$.

Our previous work [9, 7], left open two natural questions on variance competitive estimator constructions for MEPs. The first is mostly of theoretical interest, the second is of both theoretical and practical interest.

- What is the smallest possible competitive ratio which can be guaranteed for all queries and data domains (for which an unbiased nonnegative estimator with bounded variances exist) ? We refer to this ratio as the *universal ratio* for monotone estimation.
- For a specific MEP, can we construct an estimator with minimum ratio for this MEP ? We refer to such an estimator as *optimally competitive*.

We partially address the first question in Section 3 by presenting a parameterized construction of estimators, which are valid for any MEP for which an unbiased nonnegative estimator with finite variances exist. This family of estimators, which we name the $\alpha L*$ estimators, has a parameter $\alpha \geq 1$. When $\alpha = 1$, we obtain the L* estimator of [7], which is 4-competitive. The L* estimators follows the lower bound of the optimal range, having a ratio closer to 1 for data $\boldsymbol{v}$ where $f(\boldsymbol{v})$ is small. For $\alpha > 1$, the estimator lies in the middle of the optimal range of estimators. We show that for $\alpha = 1.5$ we obtain an upper bound of $27/8 \approx 3.38$ on the ratio for any MEP. There fore we obtain a tighter upper bound which strictly improves over the previous bound of 4 obtained by our L* estimator.

In Section 4, we obtain a lower bound on the universal ratio by first devising a method to construct an optimally competitive estimator for MEP over a finite domain. We then conduct a computer search over certain function families on finite domains. For these instances we computed the instance-optimal competitive ratio. The highest ratio we encountered in our search was 1.44, which gives a lower bound of 1.44 on the universal ratio.

2 Preliminaries

We review some material that is necessary for our presentation. For a set $Z \subset \mathbf{V}$, we define $\underline{f}(Z) = \inf\{f(v) \mid v \in Z\}$ to be the infimum of f on Z. For an outcome $S(u, \boldsymbol{v})$, we use the notation $\underline{f}(S) \equiv \underline{f}^{(\boldsymbol{v})}(u) \equiv \underline{f}(S^*)$ for the infimum of f on all data vectors S^*consistent with the outcome.

From monotonicity of the sampling scheme, it follows that $\forall \boldsymbol{v},\ \underline{f}^{(\boldsymbol{v})}(u)$ is monotone non increasing in u. It is also not hard to see that any unbiased and nonnegative estimator $\hat{f}$ must satisfy

$$\forall \boldsymbol{v}, \forall \rho, \int_{\rho}^{1} \hat{f}(u, \boldsymbol{v}) du \leq \underline{f}^{(\boldsymbol{v})}(\rho) \ . \tag{1}$$

The lower bound function $\underline{f}^{(\boldsymbol{v})}$, and its lower hull $H_f^{(\boldsymbol{v})}$, can be used to determine the existence of estimators with certain properties [9]:

- $\exists$ unbiased nonnegative f estimator $\iff$ (2)
  $\forall \boldsymbol{v} \in \mathbf{V},\ \lim_{u \rightarrow 0^+} \underline{f}^{(\boldsymbol{v})}(u) = f(\boldsymbol{v})$. (3)
- If f satisfies (3), $\exists$ unbiased nonnegative estimator with finite variance for $\boldsymbol{v}$
  $$\iff \int_0^1 \left(\frac{dH_f^{(\boldsymbol{v})}(u)}{du}\right)^2 du < \infty \ . \tag{4}$$

We work with *partial specifications* $\hat{f}$ of (nonnegative and unbiased) estimators. The specification is for a set of outcomes that is closed to increased u: For all $\boldsymbol{v}$, there is ρ_v, so that $S(u, \boldsymbol{v})$ is specified if and only if $u > \rho_v$. If $\rho_v = 0$, the estimator is *fully specified* for $\boldsymbol{v}$.

The partial specification is nonnegative, and we also require that for any $\boldsymbol{v}$, the estimate values on the specified part never exceed $f(\boldsymbol{v})$, which from (1), is clearly a necessary condition for extending the specification to a nonnegative unbiased estimator. We established in [9] that if a MEP satisfies (3) (has a nonnegative unbiased estimator), then any partially specified estimator can be extended to an unbiased nonnegative estimator.

Our derivations of estimators in [9, 7] utilize partial specifications: We express the estimate value on an outcome as a function of the estimate values of all "less-informative" outcomes (those with larger u). The specification is such that the estimate on an outcome is selected to be "optimal" in some respect. In particular, we can precisely consider an optimal choice of $\hat{f}(S)$ with respect to a particular consistent vector $\boldsymbol{v} \in S^*$.

Given a partially specified estimator $\hat{f}$ so that $\rho_v > 0$ and $M = \int_{\rho_v}^{1} \hat{f}(u, \boldsymbol{v}) du$, a ***v**-optimal extension* is an extension which is fully specified for $\boldsymbol{v}$ and, among all such extensions, minimizes variance for $\boldsymbol{v}$. The $\boldsymbol{v}$-optimal extension is defined on outcomes $S(u, \boldsymbol{v})$ for $u \in (0, \rho_v]$ and minimizes $\int_0^{\rho_v} \hat{f}(u, \boldsymbol{v})^2 du$ subject to $\int_0^{\rho_v} \hat{f}(u, \boldsymbol{v}) du = f(\boldsymbol{v}) - M$ (unbiasedness), $\forall u, \hat{f}(u, \boldsymbol{v}) \geq 0$ (nonnegativity), and $\forall u, \int_u^{\rho_v} \hat{f}(x, \boldsymbol{v}) dx \leq \underline{f}^{(\boldsymbol{v})}(u) - M$ (necessary nonnegativity for other data). At the point ρ_v, the ***v**-optimal estimate* is

$$\lambda(\rho, \boldsymbol{v}, M) = \inf_{0 \leq \eta < \rho} \frac{\underline{f}(\eta, \boldsymbol{v}) - M}{\rho - \eta} \ . \tag{5}$$

For the outcome $S(\rho, v)$, we can also consider the range of optimal estimates (with respect to M). The infimum and supremum of this range are

$$\lambda_U(S,M) = \sup_{\boldsymbol{z}\in S^*(\rho,\boldsymbol{v})} \lambda(\rho,\boldsymbol{z},M) \tag{6}$$

$$\lambda_L(S,M) = \inf_{\boldsymbol{z}\in S^*(\rho,\boldsymbol{v})} \lambda(\rho,\boldsymbol{z},M) = \frac{\underline{f}(\rho,\boldsymbol{v})-M}{\rho} \tag{7}$$

Estimators that are outside the range with finite probability can not be (unbiased and nonnegative) admissible, that is, they can be strictly improved.

For $\rho_v \in (0,1]$ and $M \in [0, \underline{f}^{(\boldsymbol{v})}(\rho_v)]$, we define the function $\hat{f}^{(\boldsymbol{v},\rho_v,M)} : (0,\rho_v] \to R_+$ as the solution of

$$\hat{f}^{(\boldsymbol{v},\rho_v,M)}(u) = \inf_{0\leq\eta<u} \frac{\underline{f}^{(\boldsymbol{v})}(\eta) - M - \int_u^{\rho_v} \hat{f}^{(\boldsymbol{v},\rho_v,M)}(u)du}{\rho-\eta} . \tag{8}$$

Geometrically, the function $\hat{f}^{(\boldsymbol{v},\rho_v,M)}$ is the negated derivative of the lower hull of the lower bound function $\underline{f}^{(\boldsymbol{v})}$ on $(0,\rho_v)$ and the point (ρ_v, M).

Theorem 2.1 *[9] Given a partially specified estimator $\hat{f}$ so that $\rho_v > 0$ and $M = \int_{\rho_v}^1 \hat{f}(u,\boldsymbol{v})du$, then $\hat{f}^{(\boldsymbol{v},\rho_v,M)}$ is the unique (up to equivalence) $\boldsymbol{v}$-optimal extension of $\hat{f}$.*

The $\boldsymbol{v}$*-optimal* estimates are the minimum variance extension of the empty specification. We use $\rho_v = 1$ and $M = 0$ and obtain $\hat{f}^{(\boldsymbol{v})} \equiv \hat{f}^{(\boldsymbol{v},1,0)}$. $\hat{f}^{(\boldsymbol{v})}$ is the solution of

$$\hat{f}^{(\boldsymbol{v})}(u) = \inf_{0\leq\eta<u} \frac{\underline{f}^{(\boldsymbol{v})}(\eta) - \int_u^1 \hat{f}^{(\boldsymbol{v})}(u)du}{\rho-\eta} , \tag{9}$$

which is the negated slope of the lower hull of the lower bound function $\underline{f}^{(\boldsymbol{v})}$.

Variance competitiveness [9] of an estimator is defined with respect to the expectation of the square. An estimator $\hat{f}$ is c*-competitive* if

$$\forall\boldsymbol{v},\ \int_0^1 \left(\hat{f}(u,\boldsymbol{v})\right)^2 du \leq c \inf_{\hat{f}'} \int_0^1 \left(\hat{f}'(u,\boldsymbol{v})\right)^2 du,$$

where the infimum is over all unbiased nonnegative estimators of f. An estimator that minimizes the expectation of the square also minimizes the expected squared error. When unbiased, it minimizes the variance.

3 Upper bound on the universal ratio

We define the family of αL* estimators, with respect to a parameter $\alpha \geq 1$. This family extends the definition of the L* estimator we presented in [7], which is the special case of $\alpha = 1$. The L* estimator is defined by searching for an estimate that is the minimum possible in the "optimal range" of admissible estimators. As a result, the estimator is "optimized" that is, has variance that is close to the minimum possible for data vectors with a smaller $f(\boldsymbol{v})$. The L* estimator is also the unique *monotone* estimator, meaning that for any $\boldsymbol{v}$ the estimate is never lower on a more informative outcomes.

For larger α, the αL* estimator gives more weight to the less informative outcomes. More precisely, the $\alpha L*$ estimator, $\hat{f}^{(\alpha L)}(x,\boldsymbol{v})$, for random seed value x and on outcomes consistent with some fixed data $\boldsymbol{v}$, is the solution of the integral equation, $\forall\boldsymbol{v}$, $\forall x \in (0,1]$,

$$\hat{f}^{(\alpha L)}(x,\boldsymbol{v}) = \frac{\alpha}{x}\left(\underline{f}^{(\boldsymbol{v})}(x) - \int_x^1 \hat{f}^{(\alpha L)}(u,\boldsymbol{v})du\right) . \tag{10}$$

We assume here that the lower bound function satisfies $\underline{f}^{(\boldsymbol{v})}(1) = 0$: Otherwise, if we are interested in estimating functions f where this is not the case, we can shift the lower bound function by subtracting $\underline{f}^{(\boldsymbol{v})}(1)$, compute the estimator with respect to the shifted function, and then add back the constant $\underline{f}^{(\boldsymbol{v})}(1)$ to the estimate. The expectation-of-square ratio computed for the shifted function can only be lower than the ratio obtained when $\underline{f}^{(\boldsymbol{v})}(1) = 0$. From (10), we get $\hat{f}^{(\alpha L)}(1, \boldsymbol{v}) = \alpha \underline{f}^{(\boldsymbol{v})}(1) = 0$.

Similarly to the special case of the L* estimator we treated in [7], the $\alpha L*$ estimate value depends only on information available from the outcome, which is the values of the lower bound function and the estimate value on less informative outcomes. Therefore, the estimates are consistently defined across the data domain. We note that for $\alpha < 1$, these estimators lie outside the optimal range on *every* outcome. Therefore, the $\alpha L*$ estimator in this case is dominated by the L* estimator and thus is not interesting.

For $\alpha > 1$, the $\alpha L*$ estimators, which solve $\hat{f}(\rho, S) = \alpha \lambda_L(S)$ are not necessarily in-range.

To force the estimator to be in-range (which results in strict improvement) we can instead define it the solution of $\hat{f}(S) = \min\{\lambda_U, \alpha\lambda_L\}$. Unbiasedness and nonnegativity of $\alpha L*$ follow immediately then from being in-range [7], but also hold without the truncation to λ_U. The upper bound establish next on the competitiveness of the αL* estimators also applies to the definition without this truncation.

We establish the following:

Theorem 3.1 *The $\alpha L*$ estimator is $\frac{4\alpha^3}{(2\alpha-1)^2}$-competitive. The supremum of the ratio over instances is at least $\frac{4\alpha^2}{(2\alpha-1)^2}$.*

Fixing the data $\boldsymbol{v}$, the lower bound function $\underline{f}^{(\boldsymbol{v})}(x)$ is bounded (upper bounded by $f(\boldsymbol{v})$ and lower bounded by 0) and monotone non-increasing and hence differentiable almost everywhere. We multiply (10) by x and take a derivative with respect to x and obtain the first-order differential equation

$$x\frac{\partial \hat{f}(x, \boldsymbol{v})}{\partial x} - (\alpha - 1)\hat{f}(x, \boldsymbol{v}) = \alpha \frac{\partial \underline{f}(x, \boldsymbol{v})}{\partial x} \; . \tag{11}$$

The solution is uniquely determined when we incorporate the initial condition $\hat{f}(1, \boldsymbol{v}) = 0$:

$$\hat{f}^{(\alpha L)}(x, \boldsymbol{v}) = -\alpha x^{\alpha-1} \int_x^1 y^{-\alpha} \frac{\partial \underline{f}(y, \boldsymbol{v})}{\partial y} dy \; . \tag{12}$$

To study competitiveness, we can consider the estimate values and the lower bound function with respect to a fixed data $\boldsymbol{v}$. We therefore omit the reference to $\boldsymbol{v}$ in the notation. For convenience, we define $g(x) = -\frac{\partial \underline{f}(x, \boldsymbol{v})}{\partial x} \geq 0$ and obtain the equation for $\hat{f}$ (with initial condition) and solution $\hat{f}_{\alpha,g}$:

$$x\hat{f}'(x) - (\alpha - 1)\hat{f}(x) \;\; = \;\; -\alpha g(x) \; , \hat{f}(1) = 0 \tag{13}$$

$$\hat{f}_{\alpha,g}(x) \;\; = \;\; \alpha x^{\alpha-1} \int_x^1 y^{-\alpha} g(y) dy \tag{14}$$

We now bound the ratio of $\int_0^1 \hat{f}_{\alpha,g}(x)^2 dx$ to $\int_0^1 g(x)^2 dx$. This corresponds to the ratio of the expectation of the square of the $\alpha L*$ estimator to $\int_0^1 g(x)^2 dx = \int_0^1 \left(\frac{\partial \underline{f}^{(\boldsymbol{v})}(x)}{\partial x}\right)^2 dx$. When the lower bound function is convex ($g(x)$ is monotone non-increasing), from Theorem 2.1, $g(x)$ are the $\boldsymbol{v}$-optimal estimates, and $\int_0^1 g(x)^2 dx$ is the minimum expectation of the square for $\boldsymbol{v}$, over all unbiased nonnegative estimators.

Theorem 3.2 *Let $g(x) \geq 0$ on $(0,1]$ be such that $\int_0^1 g(x)^2 dx < \infty$. For $\alpha \geq 1$, let $\hat{f}(x) \equiv \hat{f}_{\alpha,g}$ be the solution of* (13). *Then*

$$\int_0^1 \hat{f}(x)^2 dx \leq \left(\frac{2\alpha}{2\alpha - 1}\right)^2 \int_0^1 g(x)^2 dx \; . \tag{15}$$

Proof Rearranging (13), we obtain

$$x\hat{f}'(x) = (\alpha - 1)\hat{f}(x) - \alpha g(x) \; . \tag{16}$$

$$\begin{aligned}
(\hat{f}(x)^2)' &= 2\hat{f}(x)\hat{f}'(x) \implies && (17)\\
\hat{f}(x)^2 &= -2\int_x^1 \hat{f}(y)\hat{f}'(y)dy \implies && (18)\\
\int_0^1 \hat{f}(x)^2 dx &= -2\int_0^1 \int_x^1 \hat{f}(y)\hat{f}'(y)dydx = -2\int_0^1 x\hat{f}'(x)\hat{f}(x)dx && (19)\\
&= -2\int_0^1 \hat{f}(x)\Big((\alpha-1)\hat{f}(x) - \alpha g(x)\Big)dx && (20)\\
&= -2(\alpha-1)\int_0^1 \hat{f}(x)^2 dx + 2\alpha\int_0^1 \hat{f}(x)g(x)dx \; . && (21)
\end{aligned}$$

We applied integration by parts to obtain (18), and then changed order of double integration (19), using the initial condition $\hat{f}(1) = 0$, and reduced to a single integral. To obtain (20), we substituted (16). Rearranging (21), and using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\int_0^1 \hat{f}(x)^2 dx &= \frac{2\alpha}{2\alpha-1}\int_0^1 \hat{f}(x)g(x)dx && (22)\\
&\leq \frac{2\alpha}{2\alpha-1}\sqrt{\int_0^1 \hat{f}(x)^2 dx}\sqrt{\int_0^1 g(x)^2 dx} \; . && (23)
\end{aligned}$$

Finally, the claim of the theorem follows by dividing both sides by $\sqrt{\int_0^1 \hat{f}(x)^2 dx}$ and squaring. ∎

We now show that the expectation of the square of the $\alpha L*$ estimates with respect to a lower bound function $\underline{f}(x)$ with lower hull $H(x)$, is bounded by α times the expectation of the square of the estimator computed with respect to the convex lower bound function $H(x)$. The statement of the theorem is in terms of the negated derivatives, $h(x)$ and $g(x)$, of $H(x)$ and $\underline{f}(x)$:

Lemma 3.1 *Let $h(x) \geq 0$ be monotone non-increasing on $(0,1]$ such that $\int_0^1 h(x)^2 dx < \infty$. Define $H(x) \equiv \int_x^1 h(u)du$. Let $g(x)$ be such that the lower hull of $G(x) \equiv \int_x^1 g(u)du$ is equal to $H(x)$. Then for $\alpha \in (1,2]$,*

$$\int_0^1 \hat{f}_{\alpha,g}(x)^2 dx \leq \alpha \int_0^1 \hat{f}_{\alpha,h}(x)^2 dx \; .$$

Proof Let $\hat{f} \equiv \hat{f}_{\alpha,g}$ be the solution (12) of (13). From the proof of Theorem 3.2, $\hat{f}$ satisfies (22). Substituting (12) in (22), we obtain

$$\begin{aligned}
\int_0^1 \hat{f}(x)^2 dx &= \frac{2\alpha}{2\alpha-1}\int_0^1 \hat{f}(x)g(x)dx\\
&= \frac{2\alpha^2}{2\alpha-1}\int_0^1 g(x)x^{\alpha-1}\int_x^1 y^{-\alpha}g(y)dydx \; . && (24)
\end{aligned}$$

We have $\int_0^1 g(x)dx = \int_0^1 h(x)dx$ and for all $x \in (0,1]$, $\int_0^x g(u)du \le \int_0^x h(u)du$.

Consider the *defining points* of the hull H. These are the points so that for all g defining the same hull, we must have $\int_x^1 g(x)dx = \int_x^1 h(x)dx$. It suffices to show that $\int_a^b \hat{f}_{\alpha,g}(x)^2 dx \le \alpha \int_a^b \hat{f}_{\alpha,h}(x)^2 dx$ between any two such points. Moreover, it suffices to consider only intervals between such points (the discontinuities). For such an interval $[a,b]$, the function h must be fixed (a linear part of the hull). We have

$$\int_a^b \hat{f}_{\alpha,g}(x)^2 dx = \tag{25}$$

$$= \frac{2\alpha^2}{2\alpha-1}\int_a^b g(x)x^{\alpha-1}\int_x^1 y^{-\alpha}g(y)dydx \tag{26}$$

$$= \frac{2\alpha^2}{2\alpha-1}\int_a^b g(x)x^{\alpha-1}\left(\int_x^b y^{-\alpha}g(y)dy + \int_b^1 y^{-\alpha}g(y)dy\right)dx\ . \tag{27}$$

Between any two defining points, $\int_a^b g(x)dx = \int_a^b h(x)dx$ and also $\int_a^x g(u)du \le \int_a^x h(u)du$. We now fix $g(x)$ in the interval $[b,1]$ and the integral $B_g = \int_b^1 y^{-\alpha}g(y)dy$. Since both b and 1 are defining points of the hull, the properties above, and monotonicity of $y^{-\alpha}$, imply that $B_g \le B_h$.

It suffices to show that

$$\frac{\int_a^b \hat{f}_{\alpha,g}(x)^2 dx}{\int_a^b \hat{f}_{\alpha,h}(x)^2 dx} \le \alpha\ . \tag{28}$$

The function $h(x)$ is constant on (a,b). Let $h(x) = A$ on (a,b). To bound the ratio (28), we separately consider and bound the ratio of g to h for each of two summands: $\int_a^b g(x)x^{\alpha-1}B_g dx$ and $\int_a^b g(x)x^{\alpha-1}\int_x^b y^{-\alpha}g(y)dydx$.

For the first summand, we have

$$\begin{aligned}\int_a^b g(x)x^{\alpha-1}B_g dx \le B_g b^{\alpha-1}(b-a)A &\le B_h A(b^\alpha - ab^{\alpha-1}) \\ &\le B_h A(b^\alpha - a^\alpha)\ .\end{aligned}$$

We have $\int_a^b h(x)x^{\alpha-1}B_h dx = B_h A(b^\alpha - a^\alpha)/\alpha$. We get that the ratio is at most α.

We now consider the ratio of the second summand

$$\int_a^b g(x)x^{\alpha-1}\int_x^b y^{-\alpha}g(y)dydx\ ,$$

for g and h.

At the denominator, we have the expression for $h(x)$, which is

$$\int_a^b h(x)x^{\alpha-1}\int_x^b y^{-\alpha}h(y)dydx =$$

$$= A^2\int_a^b x^{\alpha-1}\int_x^b y^{-\alpha}dydx$$

$$= \frac{A^2}{\alpha-1}\int_a^b x^{\alpha-1}(x^{1-\alpha} - b^{1-\alpha})dx \tag{29}$$

$$= \frac{A^2}{\alpha-1}\left((b-a) - \frac{b^{1-\alpha}}{\alpha}(b^\alpha - a^\alpha)\right)$$

$$= \frac{A^2}{\alpha-1}\left((b-a) - \frac{b}{\alpha}(1 - \left(\frac{a}{b}\right)^\alpha)\right)\ . \tag{30}$$

We now consider $\int_a^b g(x)x^{\alpha-1}\int_x^b y^{-\alpha}g(y)dy$

We approximate g by a piecewise constant function, on n pieces, each containing $1/n$ of the mass. The breakpoints are $a \equiv t_0 < t_1 \cdots < t_n \equiv b$ satisfy $\int_a^{t_i} g(x)dx = i(b-a)A/n$. The breakpoints must satisfy $t_i \geq a + i(b-a)/n$. The fixed value in (t_i, t_{i+1}) is $W_i = \frac{(b-a)A}{n(t_{i+1}-t_i)}$ We have for $j > i$,

$$\begin{aligned}
T_{ij} &\equiv \int_{t_i}^{t_{i+1}} g(x)x^{\alpha-1}\int_{t_j}^{t_{j+1}} g(y)y^{-\alpha}dydx \\
&= \int_{t_i}^{t_{i+1}} W_i x^{\alpha-1}\int_{t_j}^{t_{j+1}} W_j y^{-\alpha}dydx \\
&= W_i W_j \frac{t_i^{\alpha-1} - t_{i+1}^{\alpha-1}}{\alpha}\frac{t_j^{-\alpha+1} - t_{j+1}^{-\alpha+1}}{\alpha-1} \\
&= \frac{(b-a)^2A^2}{n^2\alpha(\alpha-1)}\frac{(t_i^{\alpha-1} - t_{i+1}^{\alpha-1})(t_j^{-\alpha+1} - t_{j+1}^{-\alpha+1})}{(t_{i+1}-t_i)(t_{j+1}-t_j)}
\end{aligned}$$

For i,

$$\begin{aligned}
T_{ii} &\equiv \int_{t_i}^{t_{i+1}} g(x)x^{\alpha-1}\int_{x}^{t_{i+1}} g(y)y^{-\alpha}dydx \\
&= W_i^2 \int_{t_i}^{t_{i+1}} x^{\alpha-1}\frac{x^{-\alpha+1} - t_{i+1}^{-\alpha+1}}{\alpha-1} \\
&= W_i^2 \frac{\left((t_{i+1}-t_i) - \frac{t_{i+1}^{-\alpha+1}}{\alpha}(t_{i+1}^{\alpha} - t_i^{\alpha})\right)}{\alpha-1} \\
&= W_i^2 \frac{\left((t_{i+1}-t_i) - \frac{t_{i+1}}{\alpha}(1 - \frac{t_i^{\alpha}}{t_{i+1}^{\alpha}})\right)}{\alpha-1} \\
&= \frac{(b-a)^2A^2}{n^2(\alpha-1)}\frac{\left((t_{i+1}-t_i) - \frac{t_{i+1}}{\alpha}(1 - \frac{t_i^{\alpha}}{t_{i+1}^{\alpha}})\right)}{(t_{i+1}-t_i)^2}
\end{aligned}$$

The expression is $\sum_{i=0}^{n-1}\sum_{j=i}^{n-1} T_{ij}$. We need to show that for all n, the maximum over sequences t is bounded by α times (30).

For $\alpha = 2$, we obtain $T_{ii} = \frac{(b-a)^2A^2}{2n^2}\frac{1}{t_{i+1}}$ and $T_{ij} = \frac{(b-a)^2A^2}{2n^2}\frac{1}{t_j t_{j+1}}$. The sum is maximized when all t_i are at their minimum value of $t_i a + i(b-a)A/n$, which means all the W_i are equal to A.

More generally, for $\alpha \in (1, 2]$, the partial derivatives of T_{ij} with respect to t_i, t_j, t_{i+1}, t_{j+1}, and of T_{ii} with respect to t_i and t_{i+1}, are all negative. This means that the sum is maximized when t_i are as small as possible, and we can use the same argument.

■

Combining the results from Theorem 3.2 and Lemma 3.1, we obtain that the αL estimator is $4\alpha^3/(2\alpha-1)^2$ competitive. This expression is minimized for $\alpha = 1.5$, where we get a competitive ratio $27/8 = 3.375$.

To conclude the proof of Theorem 3.1 we need to show that for any $\epsilon > 0$ there are instances where $\alpha L*$ has ratio at least $\frac{4\alpha^2}{(2\alpha-1)^2} - \epsilon$:

Lemma 3.2 *The supremum of the ratio of the $\alpha L*$ estimator is $\geq \frac{4\alpha^2}{(2\alpha-1)^2}$.*

Proof Consider the function $f(v) = 1 - v^p$ ($p \in (0.5, 1]$), where $v \in [0, 1]$. For data $v = 0$, the lower bound function is $1 - v^p$ and is square integrable for $p \in (0, 5, 1]$. Since the lower bound function is convex, the 0-optimal estimates are $\hat{f}^{(0)}(x) = \underline{f}(x)' = p/x^{1-p}$. The optimal expectation of the square is $\frac{p^2}{2p-1}$.

The $\alpha L*$ estimator is $\hat{f}^{(\alpha L)}(x) = \frac{\alpha p}{\alpha - p}(x^{p-1} - x^{\alpha-1})$. The expectation of the square is

$$\int_0^1 \hat{f}^{(\alpha L)}(x)^2 dx = \frac{\alpha^2 p^2}{(\alpha - p)^2}\left(\frac{1}{2\alpha - 1} + \frac{1}{2p-1} - \frac{2}{\alpha + p - 1}\right) .$$

Simplifying, we obtain the ratio of $\int_0^1 \hat{f}^{(\alpha L)}(x)^2 dx$ to the optimum:

$$\frac{2\alpha^2}{(2\alpha - 1)(\alpha + p - 1)} .$$

Fixing α, we look at the supremum over $p \in (0.5, 1]$ of this ratio, which is obtained for $p \to 0.5^+$ and is equal to $\frac{4\alpha^2}{(2\alpha-1)^2}$. ■

We obtain ratio ≥ 4 for $\alpha = 1$ (the L* estimator) and $\geq 16/9$ for $\alpha = 2$.

4 Lower bound on the universal ratio

We start with a simple example of a MEP where any (nonnegative unbiased) estimator has ratio that is at least $10/9$. This gives a lower bound of $10/9$ on the universal ratio.

The data domain has 3 points: $\mathbf{V} = \{0, 0.5, 1\}$ and the function is $f(0) = 2$, $f(0.5) = 1$, and $f(1) = 0$. The sampling scheme is such that data $v \in \mathbf{V}$, is sampled $\iff u < v$. That is, if $u < v$ then $S^* = \{v\}$ and otherwise, $S^* = [0, u) \cap \mathbf{V}$. The lower bound function for $v = 1$ is $\underline{f}^{(1)}(u) \equiv 0$, for $v = 0.5$ is $\underline{f}^{(0.5)}(u) = 1$ for $u \in (0, 1)$ and for $v = 0$, we have is $\underline{f}^{(0)}(u) = 2$ for $u \in (0, 0.5]$ and $\underline{f}^{(0)}(u) = 1$ for $u \in (0.5, 1)$. The v-optimal estimates for each of $v \in \{0, 0.5, 1\}$ are fixed $\hat{f}^{(v)}(u) \equiv f(v)$ for $u \in (0, 1)$. The optimal expectation of the square is therefore $f(v)^2$.

Any variance optimal nonnegative unbiased estimator must be 0 when the data is 1. When the data is $\{0, 0.5\}$, the estimator must have the same fixed value $y \in [0, 2]$ for $x \in (0.5, 1)$ and a different fixed value (determined by y, v, and unbiasedness) when $v \in \{0.5, 1\}$. This value is equal to $2 - y$ when $v = 0.5$ and to $4 - y$ when $v = 1$. (since information is the same on all these outcomes, variance is minimized when the estimate is the same). The respective expectation of the square, as a function of y, is accordingly $y^2/2 + (2-y)^2/2 = y^2 - 2y + 2$ for $v = 0.5$ and is $y^2/2 + (4-y)^2/2 = y^2 + 8 - 4y$ for $v = 0$. The two ratios are respectively $y^2 - 2y + 2$ for $v = 0.5$ and $y^2/4 + 2 - y$ for $v = 1$. The competitive ratio is minimized by y which minimizes the maximum of $y^2 - 2y + 2$ and $y^2/4 + 2 - y$. The maximum is minimizes when $y = 4/3$. The corresponding ratio of this estimator is $10/9$.

4.1 Computer search for a tighter lower bound

Using a computer program we computed the optimal ratio on MEPs on discrete domains which included thousands of points. We obtained instances where any estimator must have ratio that is at least 1.44.

Providing more detail, we considered discrete one-dimensional domain $\mathbf{V} = \{i/n\}$ for $i = 0, \ldots, n$. The sampling scheme we use is PPS sampling of v: For $u \sim U[0, 1]$, we "sample" v if and only if $v \geq u$.

The respective monotone sampling scheme has $S^*(u,v) = \{v\}$ when $u \leq v$ and $S^*(u,v) = [0,u) \cap \mathbf{V}$ otherwise.

We are then interested in estimating $f(v) = 1 - v^p$ for $p \in (0,1]$ and estimating $f(v) = (1-v)^p$ for $p > 1$. It is easy to verify that on this finite domain, unbiased nonnegative estimators with finite variances exist for all p and n.

Any (nonnegative unbiased) admissible estimator must have a very particular structure. The value $\hat{f}(v,u)$ is the same for all $v < u$. For $u < v$, when we know v exactly, the estimate is determined by the values for $\hat{f}(v,u)$ for $u > v$ and unbiasedness. Moreover, an admissible estimator is also fixed in each interval $X_i = (i/n, (i+1)/n]$, since the information we have, in terms of S^*, within the interval is the same.

It therefore suffices to consider the values of the estimator on the n points i/n, and ensure for unbiasedness that for any v, the integral over $u > v$ does not exceed $f(v)$.

We first implemented a subroutine which (attempts to) construct a c-competitive estimator for a particular c. We also compute the v-optimal estimate for any $v \in \mathbf{V}$. The subroutine considers the intervals X_i in decreasing i order. At each step, we use the maximum estimate so that the ratio on affected data points (those consistent with $v \leq i/n$ remains below c. We can then compute the full estimate for $u \leq i/n$ for the point $v = i/n$ and test its competitiveness. If there is no c-competitive estimator, that is, the input choice of c was too low, our subroutine reveals that and stops. Otherwise, it finds a c competitive estimator.

We apply this subroutine in a binary search, looking for the minimum value c for which the subroutine succeeds in building an estimator. This allows us to approximate or tightly lower bound, the optimal ratio for this MEP. The highest ratio we found on the MEPs we examined was 1.44. This implies a lower bound of 1.44 on the universal ratio.

Finally, we note that this construction of the optimally-competitive estimator only applies with certain simple family of functions. It would be interesting to come up with a general construction. The particular function $1 - v^p$ is interesting, since the L* estimator has a ratio which approaches its worst-case ratio of 4 [7]. The particular function $(1-v)^p$ is also interesting. It is a special case of the exponentiated range, $\textsc{rg}_p(\boldsymbol{v}) = |v_1 - v_2|^p$ which is the basis of Manhattan and Euclidean distance estimation [6] with PPS sampling. In fact, our construction yields an optimally competitive ratio of ≈ 1.204 for $p = 1$ and of ≈ 1.35 for $p = 2$, whereas the ratio of the L* estimator is respectively 2 ($p = 1$) and 2.5 ($p = 2$).

References

[1] A. Borodin, N. Linial, and M. E. Saks. An optimal on-line algorithm for metrical task system. *J. ACM*, 39(4):745–763, 1992.

[2] K. R. W. Brewer, L. J. Early, and S. F. Joyce. Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3):231–239, 1972.

[3] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29. IEEE, 1997.

[4] A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proc.of the 11th Annual Symposium on Combinatorial Pattern Matching*, volume 1848 of *LNCS*, pages 1–10. Springer, 2000.

[5] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.

[6] E. Cohen. Distance queries from sampled data: Accurate and efficient. In *KDD*. ACM, 2014. full version: `http://arxiv.org/abs/1203.4903`.

[7] E. Cohen. Estimation for monotone sampling: Competitiveness and customization. In *PODC*. ACM, 2014. full version `http://arxiv.org/abs/1212.0243`.

[8] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *ACM SIGMETRICS*, 2009.

[9] E. Cohen and H. Kaplan. What you can do with coordinated samples. In *The 17th. International Workshop on Randomization and Computation (RANDOM)*, 2013. full version: `http://arxiv.org/abs/1206.5637`.

[10] E. Cohen, H. Kaplan, and S. Sen. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. *Proceedings of the VLDB Endowment*, 2(1–2), 2009. full version: `http://arxiv.org/abs/0906.4560`.

[11] E. L. Lehmann. *Theory of point estimation*. Wadsworth, California, 1983.

[12] P. Li, , K. W. Church, and T. Hastie. One sketch for all: Theory and application of conditional random sampling. In *NIPS*, 2008.

[13] E. Ohlsson. Coordination of pps samples over time. In *The 2nd International Conference on Establishment Surveys*, pages 255–264. American Statistical Association, 2000.

[14] B. Rosén. Asymptotic theory for order sampling. *J. Statistical Planning and Inference*, 62(2):135–158, 1997.

[15] P. J. Saavedra. Fixed sample size pps approximations with a permanent random number. In *Proc. of the Section on Survey Research Methods, Alexandria VA*, pages 697–700. American Statistical Association, 1995.