The decision in a data-driven decision-making problem is generally a high-dimensional function of data. When can the decision be reduced to a single-dimensional function of a statistic? This study addresses this question based on the operational statistics literature. The study introduces the notion of sufficient operational statistics and derives the factorization theorem for identifying such statistics. Further, the study proposes a solution procedure based on the statistics and derives the finite-sample performance bound of the proposed solution.

Key words: data-driven decision-making, operational statistics, data reduction, sufficiency

History: Received: April 2021; Accepted: January 2022 by Zuo-Jun (Max) Shen, after 2 revisions.

1 Introduction and Problem Formulation

In recent years, advances in internet technology and data science have sparked a heightened interest in data-driven decision-making. Numerous concepts, technologies, and models have emerged to address various business problems. Yet, despite the rich body of literature in this field, one important aspect seems to remain unclear, that is, data reduction in decision-making.

Generally speaking, the decision in a data-driven decision-making problem is a high-dimensional function of data. From a statistical point of view, the decision can also be viewed as a point estimator; it is not an estimator of an unknown distribution parameter as in classical point estimation, but an estimator of the true (unknown) optimal decision. We note that in classical point estimation, a major endeavor is data reduction, which answers the question: can a point estimator (a high-dimensional function of data) be reduced to a single-dimensional function of a sufficient statistic? This naturally promotes a related inquiry for data-driven decision-making: can the decision (a point estimator of the true optimal decision) be reduced to a function of a statistic? This study addresses this question.

This study builds upon the stream of literature on operational statistics (e.g., Liyanage and Shanthikumar 2005, Lim et al. 2006, Chu et al. 2008, Ramamurthy et al. 2012, Lu et al. 2015).
Specifically, consider a data-driven decision-making problem involving $m$ random variables, $\mathbf{X} = (X_1, ..., X_m)$, that are independently drawn from a distribution $F_{\mathbf{X}}$. The decision maker does not fully know the distribution $F_{\mathbf{X}}$, but instead observes a size-$n$ sample, $\hat{\mathbf{x}} = (\hat{x}_1, ..., \hat{x}_n)$, as a realization of the random sample, $\hat{\mathbf{X}} = (\hat{X}_1, ..., \hat{X}_n)$, from the same distribution.\(^1\) The decision maker makes a decision based on (as a function of) the sample, denoted by $r(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$, and gains the corresponding expected payoff:

$$
E \left[ \phi \left( r(\hat{\mathbf{X}}), \mathbf{X} \right) \right].
$$

(1)

The operational statistics literature considered a problem where the payoff function $\phi(\cdot)$ is homogeneous in the decision and random variables. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be degree-$k$ homogeneous if $f(\alpha \mathbf{x}) = \alpha^k f(\mathbf{x})$ for $\alpha \in \mathbb{R}^+$ and $\mathbf{x} \in \mathbb{R}^n$. Homogeneity essentially means that scaling the inputs to a model will scale the output accordingly. In economics, a degree-$k$ homogeneous utility function is said to exhibit decreasing, constant, or increasing returns to scale when $k < 1$, $k = 1$, or $k > 1$ respectively. Homogeneous functions are very common in economic models. “Profit functions and cost functions that are derived from production functions, and demand functions that are derived from utility functions are automatically homogeneous in the standard economic models” (Simon and Blume 1994, Chapter 20). Simon and Blume (1994) provide many examples of homogeneous functions. For example, the Cobb-Douglas function (Cobb and Douglas 1928), a utility function frequently used by economists, is homogeneous. The usual demand function of a price vector and an income level is homogeneous (Simon and Blume 1994, p. 486). A well-known model in operations management and applied economics, the newsvendor model, is homogeneous (see, e.g., Liyanage and Shanthikumar 2005, Chu et al. 2008, Ramamurthy et al. 2012). Another common economic model, the auction model, is also homogeneous in certain cases (see, e.g., Krishna 2010). More details about the newsvendor and auction models will be discussed in Section 2.

In addition, the operational statistics literature has focused on a scale parametric setting where the distribution $F_{\mathbf{X}}$ is known up to a scale. In other words, a random variable $X$ that follows $F_X$ can be represented as $X = \theta Z$ where $\theta \in \mathbb{R}^+$ is an unknown scalar and $Z$ has a known distribution (extensions with more than one unknown parameter have also been considered; see, e.g., Chu et al. 2008, Ramamurthy et al. 2012). The parametric setting serves our purpose of examining data

\(^1\)Our result holds as long as $\mathbf{X}$ is independent of $\hat{\mathbf{X}}$; the respective components in $\mathbf{X}$ or $\hat{\mathbf{X}}$ can be dependent and follow nonidentical marginal distributions without affecting the result.
reduction, because similar to classical estimation theory, identifying sufficient statistics essentially entails a factorization rule built upon parametric models.

Thus, following the literature, we consider data-driven decision-making problems with homogeneous payoff functions and scale parametric settings (we later also examine the location-scale parametric case). For such problems, Chu et al. (2008) developed a solution approach to find the decision rule that maximizes the expected payoff (1) within the class of decision rules that are degree-one homogeneous in data (i.e., the decision scales with data). This restriction to a class of decision rules is necessary: there is no universally optimal decision rule, so the rule must be sought in a reasonable class. This is again similar to classical point estimation, where a universally “best” estimator does not exist and the best estimator has to be sought in a class (e.g., the best unbiased estimator and the best linear unbiased estimator in regression). In this current problem, since the scale of the distribution is unknown, considering decision rules that scale with data is reasonable.

While the approach in Chu et al. (2008) can be used to solve the aforementioned problems, the solution is in general a high-dimensional function of data and may not be easy to obtain even numerically. Thus, a natural question is: Can the solution be reduced to a single-dimensional function of a simple statistic, as in classical estimation? Our analysis next answers this question.

We introduce the notation of sufficient operational statistics and derive the factorization theorem for identifying such statistics through inspection of the sample distribution. The statistics contain all the information needed for the decision-making and greatly facilitate the derivation of the optimal solution. We further propose a solution procedure based on the statistics and derive the finite-sample performance bound of the solution. It is worth noting that although our problem setting follows the operational statistics literature, our solution procedure is a stand-alone approach; the approach itself does not rely on the analysis in Chu et al. (2008). This approach applies to any problems with homogeneous payoff functions. Given the plethora of homogeneous economic models as discussed earlier, we naturally expect further applications of our approach to a variety of such problems.

## 2 Sufficient Operational Statistics

Given the aforementioned problem formulation, we denote by \( r^* (\cdot) \) the optimal decision rule that maximizes the expected payoff (1) within the class of degree-one homogeneous decision rules. We propose the following definition:
Definition 1 We say that a statistic \( t(\cdot) \) is a sufficient operational statistic if \( r^*(\hat{x}) = g(t(\hat{x})) \) for any sample \( \hat{x} \) and some function \( g(\cdot) \).

Analogous to the definition of the classical sufficient statistics, if the optimal decision rule can be expressed as a function of only a statistic, then that statistic is sufficient (in fact, there are many ways to define sufficiency; this definition is just one version). It is also worth noting that the operational statistics literature has referred to the optimal decision rule as an operational statistic. Therefore, the “sufficient operational statistics” notion here means “sufficient statistics for identifying operational statistics (optimal decision rules)” or “sufficient statistics for operational purposes.”

Similar to the classical sufficient statistics theory, identifying a sufficient operational statistic through its definition is usually unwieldy. Fortunately, the next factorization theorem allows us to characterize the statistic by simply inspecting the sample distribution function.

Theorem 1 (Factorization Theorem) For a random sample \( \hat{X} \), if its density or mass function \( f_{\hat{X}}(\hat{x}) = h(\hat{x})g(t(\hat{x}), \theta) \) for an unknown parameter \( \theta \) and some homogeneous functions \( h(\cdot) \) and \( t(\cdot) \), then \( t(\cdot) \) is a sufficient operational statistic.

Many sample distributions (uniform, exponential, beta, gamma, etc.) have this format. For example, the sample from a gamma distribution with an unknown parameter \( \beta \) has the density function: 
\[
(\prod_i \hat{x}_i^{\alpha-1}) e^{-\sum_i \hat{x}_i/\beta} / [\Gamma(\alpha)\beta^\alpha]^n.
\]
Let \( h(\hat{x}) = \prod_i \hat{x}_i^{\alpha-1} \) and \( t(\hat{x}) = \sum_i \hat{x}_i \), then the sample sum (or sample mean) is a sufficient operational statistic. The sample sum (or mean) is a common sufficient operational statistic for distributions in the exponential family; examples of other sufficient operational statistics include the sample product (e.g., Pareto) and the sample power sum (\( \sum_i \hat{x}_i^\gamma \), e.g., Weibull).

Clearly, sufficient operational statistics are not unique, as any one-to-one homogeneous function of a sufficient operational statistic characterized by Theorem 1 is also a sufficient operational statistic (analysis in Appendix A). For example, if the sample sum is a sufficient operational statistic, then the sample mean is for sure also a sufficient operational statistic. Since we can always normalize sufficient operational statistics to a degree-one homogeneous function, we consider only degree-one homogeneous statistics in the following analysis.

While sufficient operational statistics achieve data reduction analogously to classical sufficient statistics, they serve very different purposes. Sufficient operational statistics achieve data reduction
for the optimal decision rules, and hence, the statistics highly depend on the properties of the pay-off function in the decision-making problem. Since we consider problems with homogeneous payoff functions, homogeneity becomes a natural requirement for the sufficient operational statistics in the factorization theorem. In contrast, classical sufficient statistics are usually defined with respect to some unknown distribution parameters independent of the decision-making problem. Correspondingly, the classical Fisher-Neyman factorization theorem does not have any further requirement beyond factorization. Thus, the sufficient operational statistics and factorization theorem herein extend the classical data reduction theory to data-driven decision-making.

It is also worth noting that unlike the classical factorization theorem, the factorization rule in Theorem 1 is only a sufficient, but not necessary, condition for sufficient operational statistics. In other words, there exist sufficient operational statistics that do not satisfy the factorization rule. For example, in Appendix C, we discuss another approach to identify sufficient operational statistics based on ancillary and complete statistics. The statistics identified by this approach are much more restrictive than Theorem 1 and mostly covered by Theorem 1 (hence, we relegate the details to the appendix). However, since Theorem 1 does not fully cover the statistics identified by this approach, Theorem 1 is not a necessary condition.

In fact, the optimal decision rule is itself a sufficient operational statistic, because the optimal decision rule can be trivially expressed as a function of itself, satisfying Definition 1. For the decision-making problems considered here, Chu et al. (2008) has shown that the optimal decision rule exists. In this sense, a sufficient operational statistic always exists for the problems considered. However, a statement like “the optimal decision rule conveys sufficient information about itself” is certainly of no interest for data reduction. Thus, we focus on the sufficient operational statistics that can be characterized through simple inspection of the sample distribution function, which is achieved by the factorization rule in Theorem 1.

We next address two other questions related to the link between sufficient operational statistics and classical sufficient statistics: is a sufficient statistic always a sufficient operational statistic? And is a sufficient operational statistic always a sufficient statistic? The answer to the first question is straightforward: Since sufficient operational statistics require additional properties (such as homogeneity), a sufficient statistic may not be a sufficient operational statistic. For the second question, we note that any statistic characterized by Theorem 1 also satisfies the classical factorization theorem and, hence, is also a sufficient statistic. However, as discussed, not all sufficient operational statistics can be characterized by Theorem 1, meaning that a sufficient operational
statistic may not be a sufficient statistic (e.g., the optimal decision rule itself). In other words, although a classical sufficient statistic carries sufficient information about unknown distribution parameters, the optimal decision rule may depend on not only the distribution parameters, but also other information about the distribution.

The aforementioned relationships are demonstrated in the Ballentine Venn diagram in Figure 1. The sufficient operational statistics characterized by factorization (Theorem 1) belong to the intersection of sufficient operational statistics and sufficient statistics. The intersection may also contain some special cases beyond Theorem 1 (as characterized by the approach in Appendix C), but the statistics characterized by the factorization rule in Theorem 1 cover most commonly used distributions, especially distributions in the exponential family.

Next, we show that the sufficient operational statistics allow a simple procedure to find the optimal decision rule:

**Theorem 2 (Solution)** If \( t(\cdot) \) satisfies Theorem 1, then \( r^*(\hat{x}) = a^*t(\hat{x}) \), where

\[
a^* = \arg\max_a \mathbb{E} \left[ \phi \left( at(\hat{Z}), Z \right) \right].
\]

In (2), the distributions of the base random variables \( Z \) and the base sample \( \hat{Z} \) and the sufficient operational statistic \( t(\cdot) \) are all known, so (2) is a regular single-dimensional optimization problem and solving it is usually unchallenging. If (2) does not have a unique solution, then any solution serves the purpose. We next illustrate this solution procedure through a newsvendor and an auction problem to show that the procedure allows us to conveniently solve these problems.

**Example 1 (Newsvendor Example)** We first consider a newsvendor problem, where a decision maker faces an exponentially distributed demand \( X \) with an unknown mean \( \theta \) (or equivalently,
\[ X = \theta Z \text{ where } Z \text{ follows a standard exponential distribution}. \] The decision maker observes a sample \( \hat{X} \) from the distribution and decides an order quantity \( r(\hat{X}) \). The unit selling price is \( s \) and the unit procurement cost is \( c \). The decision maker solves:

\[
\max_{r(\cdot) \in \mathcal{R}} E \left[ s \min \left( r(\hat{X}), X \right) - cr(\hat{X}) \right],
\]

where \( \mathcal{R} \) contains all degree-one homogeneous decision rules. By Theorem 1, for exponential distribution, the sample sum, denoted by \( |\hat{x}| \), is a sufficient operational statistic. Thus, by Theorem 2,

\[
a^* = \arg \max_a \left( s \min \left( a|\hat{Z}|, c|\hat{Z}| \right) = \arg \max_a sE |\hat{Z}| \left( \min \left( a|\hat{Z}|, Z \right) \right) \right) - can = \arg \max_a s \left( 1 - (1 + a)^{-n} \right) - can = \left( \frac{s}{c} \right)^{1/(n+1)} - 1,
\]

where the fourth equality holds because \( |\hat{Z}| \) follows \( \text{gamma}(n, 1) \) and its moment generating function \( E \left( e^{t|\hat{Z}|} \right) = (1 - t)^{-n}. \) Thus, the optimal decision rule \( r^*(\hat{x}) = \left( \left( \frac{s}{c} \right)^{1/(n+1)} - 1 \right) |\hat{x}|. \) This solution was first derived by Liyanage and Shanthikumar (2005) through a different approach. Our approach, based on sufficient statistics, provides a simple way to derive the decision.

**Example 2 (Auction Example)** We next consider an auction problem where \( m \) bidders draw independent private values \( X = (X_1, ..., X_m) \) from a uniform distribution on \([0, \theta]\) with \( \theta \) unknown (or equivalently, each \( X_i = \theta Z_i \) where \( Z_i \) follows a standard uniform distribution). The seller’s own value for the auctioned asset is normalized to 0, which is usual in the literature. The auction format is a second-price auction. In such an auction, buyers bid their true values. The seller observes a sample \( \hat{X} \) from the value distribution (or a sample of historical bids since bids are true values) and decides on a reserve price \( r(\hat{X}) \). The auction succeeds when the highest bid exceeds the reserve price, in which case the highest bidder wins and pays the maximum of the second highest bid and the reserve price. Let \( X_{[i]} \) denote the \( i \)th highest order statistic in \( X \). The seller solves:

\[
\max_{r(\cdot) \in \mathcal{R}} E \left[ X_{[2]} \mathbb{I}(r(\hat{X}) < X_{[2]}) + r(\hat{X}) \mathbb{I}(X_{[2]} \leq r(\hat{X}) \leq X_{[1]}) \right],
\]

where \( \mathbb{I}(\cdot) \) is the indicator function and \( \mathcal{R} \) contains all degree-one homogeneous decision rules. This payoff function is clearly homogenous. For the uniform distribution, the sample distribution is \( f_{\hat{X}}(\hat{x}) = \theta^{-n} \mathbb{I}(\max\{\hat{x}_1, ..., \hat{x}_n\} \leq \theta). \) Thus, by Theorem 1, the sample maximum, \( \max\{\hat{x}\} \), is a sufficient operational statistic. We then apply Theorem 2 and obtain (details in appendix):

\[
r^*(\hat{x}) = m + n + 1 \frac{m + n}{2(m + n)} \max\{\hat{x}\}.
\]

This solution was first derived in Jia and Katok (2021) by using the approach in Chu et al. (2008). One prominent feature here is that, unlike the classical auction theory and the classical “estimate-and-then-optimize” approach, the decision rule here depends on the number of bidders.
3 Performance Analysis

In the previous section, we have developed a solution procedure based on sufficient operational statistics to find the best decision rule within the class of degree-one homogeneous decision rules. As previously discussed, the restriction to a class of decision rules is necessary, as no universally optimal decision rule exists. However, one question remains unclear: how does the best rule from this class perform relative to the true (unachievable) optimal performance assuming a known distribution? Fortunately, our solution offers a convenient way to establish a finite-sample performance guarantee.

By Theorem 2, the optimal decision rule is in the format of $a^*t(\hat{x})$. Thus, the performance analysis entails bounding the coefficient $a^*$ and the sufficient operational statistic, respectively. Here, $a^*$ is the solution to the payoff optimization problem (2), for which the first-order condition is often used to examine properties of the optimal solution. Many of such first-order analyses in the literature require some type of continuity of the payoff function, with one common requirement being Lipschitz continuity (see, e.g., Lemma 2.1 in Levi et al. 2007 and the proof of Proposition 2 in Ban and Rudin 2019; e.g., the property has been established for newsvendor and auction models under certain regularity conditions). A function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if $|f(x_1) - f(x_2)| \leq s|x_1 - x_2|$ for some $s > 0$ (called a Lipschitz constant) and any $x_1, x_2 \in \mathbb{R}$. We thus follow the literature and consider Lipschitz continuous payoff functions.

Now consider a sufficient operational statistic $t(\hat{Z})$ of the base sample $\hat{Z}$. Since infinitely many such statistics exist, it is important for us to focus on a statistic with desirable concentration properties to derive results. Without loss of generality, we can focus on the case that $t(\hat{Z})$ has a finite mean. As an example, the sample sum is a usual sufficient operational statistic for distributions in the exponential family, but its magnitude increases with the sample size; we can focus on the sample mean instead, which often has desired concentration properties. In general, if $\mathbb{E}[t(\hat{Z})] = g(n)$ for some increasing function $g(\cdot)$, then we can focus on the normalized statistic $t(\hat{Z})/g(n)$ instead, which has a finite mean. In fact, in this analysis, we consider only a normalized sufficient operational statistic such that $\mathbb{E}[t(\hat{Z})] = 1$. By Theorem 1, such scaling preserves sufficiency.

We say that a statistic $t(\hat{Z})$ of the base sample $\hat{Z}$ is bounded in mean (concentrated around its mean) by a function $f(n)$ if:

$$\mathbb{E}\left(|t(\hat{Z}) - 1|\right) \leq f(n),$$

for some decreasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ and $f(n) \to 0$ as $n \to \infty$.

This condition further implies that $\mathbb{E}|t(\hat{Z})|$ is finite and $t(\hat{Z}) \overset{L^1}{\to} 1$. Many common statistics have such bounds. For example, the sample mean is a common sufficient operational statistic for...
distributions in the exponential family. When the sample comes from a sub-Gaussian distribution, the sample mean $t(\hat{Z})$ satisfies:

$$\mathbb{E}\left(|t(\hat{Z}) - 1|\right) \leq \frac{c}{\sqrt{n}} \text{ for some } c \geq 0.$$

When the sample comes from a sub-exponential distribution,

$$\mathbb{E}\left(|t(\hat{Z}) - 1|\right) \leq \frac{c}{\sqrt{n}} + \frac{d}{n} \text{ for some } c, d \geq 0.$$

Both are bounded by functions decreasing in $n$. Sub-Gaussian and sub-exponential classes contain many common distributions: the former contains Gaussian, Bernoulli and all bounded distributions, and the latter includes all sub-Gaussian distributions, and exponential and Poisson distributions. As another example, the sample maximum is a sufficient operational statistic for some bounded distributions (such as the uniform distribution). It is easy to verify that the sample maximum for any bounded distribution also has such bounds. Appendix D provides more details about these examples.

Let $a^o$ denote the smallest optimizer of $\mathbb{E}[\phi(a, Z)]$ and $\phi^o = \mathbb{E}[\phi(a^o, Z)]$, where $a^o$ and $\phi^o$ are assumed to be finite and $\phi^o$ is assumed to be positive to avoid triviality. We say that a solution is $\epsilon$-optimal if its payoff is at least $(1 - \epsilon)$ times the true optimal payoff.

**Theorem 3** If the payoff function $\phi$ is Lipschitz continuous with a Lipschitz constant $s$ and the sufficient operational statistic $t(\hat{Z})$ is bounded in mean by $f(n)$, then the solution in Theorem 2 is $(s|a^o|/\phi^o)f(n)$-optimal.

Theorem 3 shows that the performance loss is bounded by the product of a constant and the bound of the sufficient operational statistic. Stemming from the structure of the solution, this bound is not a high-probability bound, but a sure bound. Thus, the performance of the proposed solution concentrates at the same rate as the sufficient operational statistic concentrates to its mean. This result has several implications. First, the theorem links the concentration rate of the performance to that of the statistic. The latter is independent of the specific payoff function in the decision-making problem, meaning that the performance concentration rate can again be assessed by inspecting the sample distribution and the statistic. The statistics from many distributions have the desired concentration property. For example, for light-tailed distributions

---

2In fact, the sample mean for any distribution with finite variance satisfies $\mathbb{E}\left(|t(\hat{Z}) - 1|\right) = O(1/\sqrt{n})$ asymptotically, but the sub-Gaussian distribution has the finite-sample property.
(e.g., sub-Gaussian distributions), one may expect the performance to very quickly converge to the optimal performance; for heavy-tailed distributions (e.g., sub-exponential distributions), the performance still converges at a reasonable (but slower) rate. Sub-Gaussian and sub-exponential distributions cover many of the commonly used distributions in applications (such as Gaussian, Bernoulli, exponential, Poisson, and all bounded distributions). In addition to the performance guarantee, Theorem 3 also provides guidance for sample size determination to achieve a target performance; this can be accomplished by setting the performance bound to be the target and solving the corresponding sample size $n$.

4 Discussions and Ending Remarks

In this study, we investigate when a data-driven decision can be reduced to a function of a statistic based on the operational statistics literature. For a class of problems with homogeneous payoff functions, a property possessed by many economic models, we define the notion of sufficient operational statistics and derive a factorization rule to characterize the statistics through simply inspecting the sample distribution. The sufficient operational statistics greatly facilitate the analysis to obtain the optimal decision rules. Based on the statistics, we develop a simple solution procedure and demonstrate the procedure through a newsvendor example and an auction example.

In this solution procedure, the decision rule is sought within the class of degree-one homogeneous rules. The restriction to a class of decision rules is necessary, but the performance loss implied by the restriction needs to be understood. To this end, we explore the finite-sample performance of the solution and find that the performance of the solution concentrates to the true optimal performance at the same rate as the sufficient operational statistic concentrates to its mean.

We conclude this study by discussing some other aspects of sufficient operational statistics, as well as the limitations of this study and possible future research:

Benchmarking

One aspect of operational statistics that has been discussed in the literature is benchmarking against the traditional estimation-and-optimization approach. Prior studies (e.g., Liyanage and Shanthikumar 2005) have shown that the traditional approach is suboptimal and outperformed by operational statistics for the problems considered. The traditional approach uses the best unbiased estimator of the unknown distribution parameter to derive the optimal decision rule, so the decision rule is a function of the best unbiased estimator; further, by the Rao-Blackwell Theorem, the best
unbiased estimator is a function of a classical sufficient statistic. Thus, the decision rule obtained by the traditional approach is a function of a classical sufficient statistic. Analogously, the decision rule obtained by Theorem 2 is a function of a sufficient operational statistic. When sufficient operational statistics overlap sufficient statistics (i.e., the intersection region in Figure 1, mainly containing statistics characterized by Theorem 1), the traditional and the current decision rules are based on the same statistics. Since the current decision rule is optimal among all degree-one homogeneous rules, if the traditional decision rule is degree-one homogeneous, then the current decision rule for sure outperforms the traditional decision rule.

**Minimal Sufficiency**

In classical data reduction, a statistic is a minimal sufficient statistic if it is a function of any other sufficient statistic. In other words, minimal sufficient statistics achieve data reduction to the maximum extent without information loss. For example, the entire data are always sufficient statistics, but if a single-dimensional sufficient statistic exists, then this statistic clearly achieves a higher degree of data reduction than the entire data. In the decision-making context, the entire data are clearly also sufficient operational statistics. Correspondingly, Theorem 1 reduces the data to a single dimension without information loss for the decision-making.

In fact, a single-dimensional sufficient operational statistic always exists for the problems considered because, as previously mentioned, the optimal decision rule itself is a sufficient operational statistic. However, except for the statistics characterized by Theorem 1, one may not always be able to identify single-dimensional sufficient operational statistics by inspecting the sample distribution. Further, when Theorem 1 does not apply, identifying multi-dimensional sufficient operational statistics by inspecting the sample distribution generally appears intractable.

Finally, we note that even for single-dimensional sufficient operational statistics, some of them may achieve a higher degree of data reduction than others (i.e., some may be represented as functions of others). However, the data reduction within a single dimension usually depends on the specific payoff function in the decision-making problem. For example, consider the trivial extreme case where the payoff function is independent of the decision rule. In this case, any statistics are operationally sufficient, but the coarsest statistic is the one that maps the data to a constant value. Clearly, such payoff-function-specific minimal sufficiency is not a desired property. Hence, it seems to be more reasonable to adopt the following alternative definition for minimal sufficient operational statistics. We note that when multiple optimal decision rules exist in a decision-making problem, selecting the coarsest decision rule is often not the focus of investigation; as long as we reduce the
data to an optimal decision rule, further reducing the rule often does not seem very necessary. Thus, we say a statistic is a \textit{minimal sufficient operational statistic} if there exists an optimal decision rule that is a one-to-one function of the statistic. In other words, by identifying this statistic we have reduced the data to the optimal decision rule level; there is no further data reduction from the statistic to the optimal decision rule. If we adopt this definition, then any statistics identified by Theorem 1 are minimal sufficient operational statistics because by Theorem 2, the optimal decision rules based on these statistics are all one-to-one functions of the statistics.

\textbf{Limitations and Possible Further Research}

Data reduction usually requires a parametric setting: both the classical factorization theorem and the current factorization theorem essentially call for the guidance of a specific distribution form; otherwise, the factorization may not be performed. Thus, two relevant questions are: Can the analysis be extended to a parametric setting with more than one unknown distribution parameter? And can the analysis be extended to a nonparametric setting? For the first question, an extension to multiple parameters is possible, but often under more restrictive conditions. For example, in Appendix E, we examine a location-scale two-parameter setting and show that the result extends to density/mass functions of an exponential form \( f(x) = c \prod_i d_i^{a_i x + b_i} \) for some \( a_i, b_i, c, d_i \); e.g., \( ce^{-ax} \) and \( p^{a_1 x} (1 - p)^{a_2 x + b} \). The explanation for the more restrictive conditions is as follows. In classical data reduction, multiple unknown distribution parameters usually require multiple statistics to achieve sufficiency. For example, for a normal sample with both an unknown mean and variance, the sample mean and variance together constitute sufficient statistics. In the decision-making problem, however, the single decision compresses all information into one dimension, dropping a lot of information about the distribution and retaining only the information necessary for the optimal decision rule. This process usually requires a strong guidance from the specific payoff function of the decision-making problem and may not be accomplished by inspecting the sample distribution when multiple unknown parameters exist. Thus, the data reduction may still be accomplished, but often under more restrictive conditions.

For the second question, the data reduction in a nonparametric setting is generally not possible under the current notion of “full” sufficiency (i.e., the reduction to statistics that contain the \textit{full information} for the decision-making). However, some notion of “partial” sufficiency may be applied to identify statistics that contain the \textit{majority of information} needed for the decision-making. Prior studies have investigated the performance guarantees of various algorithms in some decision-making problems. However, for a general class of problems, it still awaits a thorough investigation whether it
is possible to retain some statistics that convey sufficient information to achieve a certain percentage of the optimal payoff. This deserves future research attention.

Acknowledgments

We are grateful to the department editor, Zuo-Jun (Max) Shen, the senior editor, and the two anonymous referees for their constructive comments and suggestions.

References


Appendices

A Proofs of Theorems 1 and 2

The proofs are based on the following lemma, which, to our best knowledge, has not been developed in any prior studies.

Lemma 1 For a random sample $Y$, if $t(\cdot)$ satisfies the condition in Theorem 1 with respect to $Y$, then $t(Y)$ is independent of $Y/t(Y)$.

Proof of Lemma 1. As discussed in the main text, without loss of generality, we consider only degree-1 homogeneous $t(\cdot)$. Consider the transformation from $Y$ to $W = [Y_1/t(Y), ..., Y_{n-1}/t(Y), t(Y)]$.

Clearly, $y_i = w_i w_n$ for $i = 1, ..., n-1$, and

$$
y_n = \arg_y [t(y_1, ..., y_{n-1}, y) = w_n] = \arg_y [t(y_1/w_n, ..., y_{n-1}/w_n, y/w_n) = 1] = w_n \arg_w [t(w_1, ..., w_{n-1}, w) = 1] = w_n s(w_1, ..., w_{n-1}),
$$

where $s(\cdot)$ represents some function of $w_1, ..., w_{n-1}$, and the second equality holds since $t(\cdot)$ is degree-1 homogeneous. We prove a continuous distribution case; the discrete distribution case follows essentially the same procedure without concerning the Jacobian determinant. For the given transformation, the Jacobian determinant

$$
\det(J) = \left| \begin{array}{cccc}
w_n & 0 & \cdots & w_1 \\
0 & w_n & \cdots & w_2 \\
\vdots & \vdots & \ddots & \vdots \\
w_n s_1(w_1, ..., w_{n-1}) & w_n s_2(w_1, ..., w_{n-1}) & \cdots & s(w_1, ..., w_{n-1})
\end{array} \right|,
$$

where $s_i(\cdot)$ is the derivative of $s(\cdot)$ with respect to its $i$th argument. By the Leibniz formula,

$$
\det(J) = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n j_{i,\sigma_i},
$$

where $\sigma$ is a permutation of the set $\{1, 2, ..., n\}$, $S_n$ is the set of all such permutations, and $\text{sgn}$ is the signature of $\sigma$ (it is $+1$ if $\sigma$ is obtained by an even number of switches and $-1$ otherwise). Since the Jacobian matrix is sparse, we can simply enumerate all non-zero terms in the Leibniz formula and then take a sum of them. Clearly, the term associated with $\sigma = \{1, 2, ..., n\}$ is the product of the elements along the diagonal, which is $w_{n-1} s(w_1, ..., w_{n-1})$. Note that for any non-zero term, $\sigma_i, i = 1, ..., n-1$ must be either $i$ or $n$ (since in row $i$, only $j_{i,i}$ and $j_{i,n}$ are non-zero), whereas
σ_n can take any value. If σ_i = n for some i = 1, ..., n − 1, then we must have σ_j = j for all
j = 1, ..., n − 1, j ≠ i and σ_n = i. In other words, to obtain a non-zero term, we can only permute
the set {1, 2, ..., n} by exchanging the value at some position i = 1, ..., n − 1 and the value at position
n. The corresponding term is −w_1 w_n^{n−2} w_n(s_1(w_1, ..., w_n−1)) = −w_n w_n−1 s_1(w_1, ..., w_n−1). Thus,

\[ \det(J) = w_n^{n−1} \left[ s(w_1, ..., w_n−1) − \sum_{i=1}^{n−1} w_i s_i(w_1, ..., w_n−1) \right]. \]

The joint pdf of W is:

\[ f_W(w) = f_Y(w_1 w_n, ..., w_n−1 w_n, w_n s(w_1, ..., w_n−1)) | \det(J)| \]
\[ = w_n^\alpha h(w_1, ..., w_n−1, s(w_1, ..., w_n−1)) g(w_n t(w_1, ..., w_n−1, s(w_1, ..., w_n−1)), \theta_0) | \det(J)| \]
\[ = w_n^\alpha h(w_1, ..., w_n−1, s(w_1, ..., w_n−1)) g(w_n, \theta_0) | \det(J)|. \]

Here, \( \theta_0 \) is the parameter of the sample Y. The second equality holds by the condition in Theorem 1,
that \( h(\cdot) \) is degree-\( \alpha \) homogeneous, and that \( t(\cdot) \) is degree-1 homogeneous. The third equality holds
by the definition of \( s(\cdot) \) in (4). Note that \( f_W(w) \) is multiplicatively separable in \( (w_1, ..., w_n−1) \) and
\( w_n \). Since \( W = [Y_1/t(Y), ..., Y_n−1/t(Y), Y_n/t(Y)] \), this multiplicative separability means that any
\( Y_i/t(Y), i = 1, ..., n−1 \) is independent of \( t(Y) \). We finally note that \( Y_n/t(Y) \) is degenerate and
is equal to \( s(Y_1/t(Y), ..., Y_n−1/t(Y)) \) from (4). If \( Y_i/t(Y), i = 1, ..., n−1 \) is independent of \( t(Y) \),
then \( Y_n/t(Y) \) must also be independent of \( t(Y) \). Therefore, \( Y/t(Y) \) is independent of \( t(Y) \).

The above analysis has implicitly assumed that the equation \( t(w_1, ..., w_n−1, w) = 1 \) has a single
root (or \( t(\cdot) \) is monotonic). If the equation has multiple roots, then we can follow the standard
procedure and partition the support of Y into regions with one-to-one transformations from Y to
W. We then follow the above analysis on each region. We note that these regions only differ in
the factor \( s(w_1, ..., w_n−1) \), but \( s(w_1, ..., w_n−1) \) is always multipatively separable from \( w_n \). Thus,
the result holds.

\[ \square \]

Proof of Theorem 1. The theorem follows from Theorem 2, where we show that if Lemma 1
holds, then the decision rule can be reduced to a function of the given statistic.

We further note that any one-to-one homogenous function of a sufficient operational statistic
characterized by Theorem 1 is also a sufficient operational statistic. To see this, suppose \( t(\cdot) \) is
a sufficient operational statistic characterized by Theorem 1 and \( \tilde{t}(\tilde{x}) = s(t(\tilde{x})) \) for all \( \tilde{x} \) and
some one-to-one homogenous function \( s(\cdot) \) with properly defined inverse \( s^{-1}(\cdot) \). \( \tilde{t}(\tilde{x}) \) is clearly
homogeneous. By Theorem 1, \( f_{\tilde{X}}(\tilde{x}) = h(\tilde{x}) g(t(\tilde{x}), \theta) = h(\tilde{x}) g(s^{-1}(\tilde{t}(\tilde{x})), \theta) = h(\tilde{x}) g(\tilde{t}(\tilde{x}), \theta) \),
where we define \( g(\tilde{t}(\tilde{x}), \theta) = g(s^{-1}(\tilde{t}(\tilde{x})), \theta) \). \( \tilde{t}(\tilde{x}) \) clearly satisfies Theorem 1 and hence is also a
sufficient operational statistic.

\[ \square \]
Proof of Theorem 2. By Lemma 1, for the base sample $\hat{Z}$, $t(\hat{Z})$ is independent of $\hat{Z}/t(\hat{Z})$. Again, without loss of generality, we consider only degree-1 homogeneous $t(\cdot)$. Thus,

$$\max_{r(\cdot)} E \left[ \phi \left( r(\hat{X}), X \right) \right] = \max_{r(\cdot)} E \left[ t(\hat{X})^k \phi \left( r \left( \frac{\hat{X}}{t(\hat{X})}, \frac{X}{t(\hat{X})} \right) \right) \right]$$

$$= \theta^k \max_{r(\cdot)} E \left[ t(\hat{Z})^k \phi \left( r \left( \frac{\hat{Z}}{t(\hat{Z})}, \frac{Z}{t(\hat{Z})} \right) \right) \right]$$

$$= \theta^k \max_{r(\cdot)} E_{\hat{Z}, Z} \left[ \max_a E_{\hat{Z}, Z} \left[ t(\hat{Z})^k \phi \left( a, \frac{Z}{t(\hat{Z})} \right) \right] \right]$$

$$= \theta^k \max_{\hat{Z}, Z} E \phi \left( at(\hat{Z}), Z \right).$$

Here, the first two equalities hold due to the degree-$k$ homogeneity of $\phi(\cdot)$ and degree-1 homogeneity of $r(\cdot)$ and $t(\cdot)$. The third equality holds because both $t(\hat{Z})$ and $Z$ are independent of $\hat{Z}/t(\hat{Z})$. The fourth equality holds because finding the optimal function $r \left( \frac{\hat{Z}}{t(\hat{Z})} \right)$ is equivalent to finding an optimal scalar $a$ for each realization of $\hat{Z}/t(\hat{Z})$ within the expectation. This optimal scalar is independent of $\hat{Z}/t(\hat{Z})$ because both $t(\hat{Z})$ and $Z$ in the scalar optimization are independent of $\hat{Z}/t(\hat{Z})$.

Let

$$a^* = \arg \max_a E \left[ \phi \left( at(\hat{Z}), Z \right) \right].$$

If multiple solutions exist, the any solution can be chosen. We now show that given any observed sample $\hat{x}$, $r^*(\hat{x}) = a^* t(\hat{x})$. By Lemma 1, $t(\hat{X})$ is independent of $\hat{X}/t(\hat{X})$. Thus,

$$\max_{r(\cdot)} E \left[ \phi \left( r(\hat{X}), X \right) \right] = \max_{r(\cdot)} E \left[ t(\hat{X})^k \phi \left( r \left( \frac{\hat{X}}{t(\hat{X})}, \frac{X}{t(\hat{X})} \right) \right) \right]$$

$$= \max_{r(\cdot)} E_{\hat{X}, X} \left[ \max_{\hat{X}} E_{\hat{X}, X} \left[ t(\hat{X})^k \phi \left( r \left( \frac{\hat{X}}{t(\hat{X})}, \frac{X}{t(\hat{X})} \right) \right) \right] \right].$$

In addition,

$$a^* = \arg \max_a E \left[ \phi \left( at(\hat{Z}), Z \right) \right] = \arg \max_a E \left[ \theta^k \phi \left( at(\hat{Z}), Z \right) \right] = \arg \max_a E \left[ t(\hat{X})^k \phi \left( a, \frac{X}{t(\hat{X})} \right) \right].$$

These two expressions mean that the optimal decision rule must satisfy $r(\hat{x}/t(\hat{x})) = a^*$ for any realized $\hat{x}/t(\hat{x})$. Thus, $r^*(\hat{x}) = a^* t(\hat{x})$. $\square$
B Proof of Performance Bound

Proof of Theorem 3. Let \( a^o \) be defined as in the main text. The difference between the true optimal expected payoff and the expected payoff generated by the decision rule in Theorem 2 is:

\[
\max_r E[\phi(r, X)] - E[\phi(r^*(\hat{X}), X)] = \theta^k E[\phi(a^o, Z)] - E[\phi(a^* t(\hat{X}), X)] \\
= \theta^k E[\phi(a^o, Z)] - \theta^k E[\phi(a^* t(\hat{Z}), Z)] \\
\leq \theta^k E[\phi(a^o, Z)] - \theta^k E[\phi(a^o t(\hat{Z}), Z)] \\
\leq \theta^k s |a^o| E \left( |1 - t(\hat{Z})| \right) \\
\leq \theta^k s |a^o| f(n),
\]

where the first equality follows from the definitions of \( a^o \) and \( r^*(X) \) and the degree-\( k \) homogeneity of \( \phi(\cdot) \), and the second from the degree-one homogeneity of \( t(\cdot) \). The first inequality follows from the optimality of \( a^* \) in equation (2), the second inequality follows from the Lipschitz continuity of \( \phi(\cdot) \), and the last from the statistic bound.

In addition, \( \max_r E[\phi(r, X)] = \theta^k \phi^o \). Thus,

\[
\left\{ \max_r E[\phi(r, X)] - E[\phi(r^*(\hat{X}), X)] \right\} / \max_r E[\phi(r, X)] \leq s |a^o| f(n) / \phi^o.
\]

Additional supporting information may be found in the online appendices for this article. □