Efficient Unbiased Sparsification

Leighton Barnes[*][†] , Timothy Chow[*], Emma Cohen[*], Keith Frankston[*],
Benjamin Howard[*], Fred Kochman[*], Daniel Scheinerman[*], and Jeffrey VanderKam[*]

[*] Center for Communications Research, Princeton, NJ 08540
[†] corresponding author: l.barnes@idaccr.org

Abstract—An unbiased m-*sparsification* of a vector $p \in \mathbb{R}^n$ is a random vector $Q \in \mathbb{R}^n$ with mean p that has at most $m < n$ nonzero coordinates. Unbiased sparsification compresses the original vector without introducing bias; it arises in various contexts, such as in federated learning and sampling sparse probability distributions. Ideally, unbiased sparsification should also minimize the expected value of a divergence function $\mathrm{Div}(Q, p)$ that measures how far away Q is from the original p. If Q is optimal in this sense, then we call it *efficient*. Our main results describe efficient unbiased sparsifications for divergences that are either permutation-invariant or additively separable. Surprisingly, the characterization for permutation-invariant divergences is robust to the choice of divergence function, in the sense that our class of optimal Q for squared Euclidean distance coincides with our class of optimal Q for Kullback–Leibler divergence, or indeed any of a wide variety of divergences.

Index Terms—federated learning, sparsification, unbiased estimator, optimization

I. INTRODUCTION

Suppose we have a vector $p \in \mathbb{R}^n$, and (possibly because of memory or bandwidth limitations) we want to approximate it with a vector $Q \in \mathbb{R}^n$ with at most m nonzero entries, where $m < n$. The construction of Q is allowed to be randomized, and we want Q to be the "best possible approximation" of p. "Best possible approximation" will be defined precisely later, but at minimum, we want the expected value of Q to equal p. Depending on our application, we may also desire the stronger property that the sum of the entries of Q should always equal the sum of the entries of p. How should we construct Q?

We call this task *efficient unbiased sparsification* (EUS)—sparsification, because the number of nonzero entries is reduced from n to m; unbiased, because the expected value of Q equals p; and efficient, in the statistical sense of diverging from p as little as possible. The problem of efficient unbiased sparsification, or something very close to it, arises in several different contexts.

1) In the context of *sampling sparse probability distributions*, we have a discrete probability distribution p on n outcomes, and we seek to randomly construct a sparse probability distribution Q that has at most m nonzero probabilities. We would like this construction to be unbiased, in the sense that the average over the potential sparse probability distributions should be the original distribution p, and for it to minimize the expected statistical divergence between Q and p. One natural choice of a divergence in this case would be the standard Kullback–Leibler (KL) divergence.

2) In *distributed statistical estimation* [1], [2], [3], [4], statistical samples are distributed across a number of client nodes that must send bandwidth-limited messages to a central server. The central server then performs statistical analyses using the compressed versions of the samples. By sparsifying potentially high-dimensional samples, they can be communicated more efficiently to the central server. The constraint that

the sparsification be unbiased preserves statistical properties such as the mean, while the efficiency of the sparsification ensures that the compressed versions of the samples are not too far from the original ones.

3) In the related field of *federated learning* [5], [6], client nodes are trying to jointly train a machine learning model. In order to facilitate the distributed training, minibatch gradients or model updates need to be communicated between nodes so that they can be aggregated into a combined model. For large machine learning models, however, a single gradient vector can be billions of parameters long, and sparsification strategies can be deployed in order to reduce the associated communication cost [7]. Notably, related works [8], [9], [10] describe optimization objectives and their associated optimal strategies that are similar to a special case of the present work. They consider a "soft" sparsification constraint, where the *expected* number of nonzero components can be at most m, instead of the "hard" constraint that we use, and they consider only the squared Euclidean distance as their divergence.

In a slightly different optimization problem that occurs in federated learning, n refers to the number of client nodes instead of the dimension of the updates, and bandwidth limitations force us to restrict the number m of clients allowed to communicate in each round. Some clients have more important updates, so the question arises of how to pick clients in a way that respects their importance, while minimizing the statistical distortion that restriction inevitably causes. This problem also leads to a similar optimization problem and is considered in [11].

4) In *sampling with specified marginals* [12], the goal is to randomly choose a subset of m items from a population of $n > m$ items, in such a way that the *inclusion probability* of item i ($1 \le i \le n$)—i.e., the probability that item i belongs to the chosen m-element subset—is proportional to some specified positive number p_i. It turns out that for some sets of numbers p_i, it is impossible to achieve this goal exactly, but we would still like to come as close as possible.

In this paper, we solve the EUS problem by setting up the associated optimization problems, explicitly giving algorithms that produce optimal random vectors Q, and, in some cases, by describing the distributions of all random vectors Q that optimize the objectives. We consider both permutation-invariant and additively separable divergences, which we will define shortly. Surprisingly, the characterization for permutation-invariant divergences is robust to the choice of divergence function, in the sense that our class of optimal Q for (say) squared Euclidean distance coincides with our class of optimal Q for (say) Kullback–Leibler divergence, or indeed any of a wide variety of divergences.

The space of unbiased m-sparsifications of a given $p \in \mathbb{R}^n$ may

be thought of as an infinite-dimensional "simplex" [13, Section III.8], which is a convex space in the sense that any mixture of two unbiased m-sparsifications is an unbiased m-sparsification. Expectation is linear, so we are minimizing a linear function over a convex space. That might sound promising, but infinite-dimensional objects are not so easy to work with. Indeed, it is not even obvious that any EUS exists. We proceed by reducing to a finite-dimensional, but not necessarily convex, problem. We then develop novel techniques for proving that our algorithms describe global minima for the finite-dimensional reductions.

A. Sampling with Specified Marginals

As preparation for describing our algorithm, we first look more closely at the problem of sampling with specified marginals. Assume first that $\sum_i p_i = m$, and that $p_i \leq 1$ for all i. Then there are many methods of randomly sampling an m-element subset T of $\{1, 2, \ldots, n\}$ in such a way that the inclusion probability of item i is equal to (and not just proportional to) p_i. One method is to partition a line segment of length m into n subintervals such that the length of subinterval i is p_i, and then choose $x \in [0, 1)$ uniformly at random, and finally let $i \in T$ if and only if $x + j$ lies in subinterval i for some integer j. We leave it to the reader to check that this method does indeed work. Tillé [12] gives many other methods, including one that maximizes entropy.

Now, let us re-examine the assumptions that $\sum_i p_i = m$ and $p_i \leq 1$ for all i. Given a probability distribution on m-element subsets, let s_i denote the inclusion probability of item i. Recall that in our original problem the marginal inclusion probabilities s_i were only required to be *proportional* to p_i, and not necessarily *equal* to p_i. By linearity of expectation, the sum of the s_i equals the expected total number of elements chosen—which in this case is precisely m. It follows that for there to exist a probability distribution on m-element subsets that achieves inclusion probabilities proportional to the given numbers p_i, a necessary condition is that if the p_i are rescaled so that they sum to m, then each rescaled p_i must be *equal* to s_i, and in particular must be at most 1. In other words, if for any i,

$$p_i > \frac{1}{m} \sum_{j=1}^{n} p_j, \tag{1}$$

then it is impossible to sample with the specified marginals. What is one supposed to do in this case?

Of course, one option is to simply issue an error message and give up. However, Tillé [12, Section 2.10] offers a different approach. If the largest p_i is "too big"—meaning that it satisfies (1)—then item i is automatically granted membership in our chosen set of m items. We are thus reduced to choosing $m - 1$ items from the remaining set of $n - 1$ candidates. Again, if the largest remaining p_i is "too big" then it is automatically included. This process is iterated until we reach a set of p_i for which sampling with the specified marginals becomes possible. Motivated by Tillé's procedure, we make the following definition.

Definition I.1. Given a sequence of positive real numbers arranged (without loss of generality) in weakly decreasing order $p_1 \geq p_2 \geq \cdots \geq p_n$, and a positive integer $m < n$, we say that index i is m-*heavy* if

$$\sum_{j=i+1}^{n} p_j \leq (m - i) p_i. \tag{2}$$

If i is not m-heavy then we say it is m-*light*.

Adding p_i to both sides of (2) shows that if i is m-heavy, then

$$\sum_{j=i}^{n} p_j \leq (m - i)p_i + p_i = (m - (i - 1))p_i \leq (m - (i - 1))p_{i-1};$$

i.e., $i - 1$ is also m-heavy. So there is some threshold h up to which all the i are m-heavy and beyond which all the i are m-light. In practice, the fastest way to locate this threshold is probably by binary search.

As an alternative formulation of the same concept, consider the constant of proportionality $r = \frac{m-h}{\sum_{i>h} p_i}$ such that $s_i = rp_i$ for light i. By the above definition i is heavy if and only if $rp_i \geq 1$. So for any i we will have $s_i = \min(rp_i, 1)$. But we must also satisfy $\sum_i s_i = m$. It is clear that

$$M(r) := \sum_{i=1}^{n} \min(rp_i, 1)$$

is a piecewise-linear, continuous, increasing function of r with $M(0) = 0$ and $M(r) = n$ iff $r \geq 1/\min(p_i)$, so there is a unique r such that $M(r) = m < n$. Then the heavy indices are precisely $\{i : p_i \geq 1/r\}$.

B. Efficient Unbiased Sparsification

In order to state our main results, we must first give more precise definitions of unbiased sparsification and divergences.

Definition I.2. Write $\mathrm{I}(v) = \{i \mid v_i \neq 0\}$ for the set of indices of nonzero coordinates of $v \in \mathbb{R}^n$, which we sometimes refer to as the *survivor set* of v. If $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ and m is a positive integer such that $|\mathrm{I}(p)| > m$, then a random vector $Q = (Q_1, \ldots, Q_n) \in \mathbb{R}^n$ is an *unbiased (random) m-sparsification* of p if

a) $|\mathrm{I}(Q)| \leq m$ (i.e., Q is m-sparse) and

b) $\mathbf{E}[Q] = p$ (i.e., Q is an unbiased estimate of p).

In some cases we may want Q to satisfy additional properties. For instance, if p lies in the probability simplex ($p_i > 0$ and $\sum_i p_i = 1$) then it is natural to require the same of Q. In this paper we consider only two cases: either p and Q are both constrained to lie in the simplex or they are both unconstrained in $\mathbb{R}^n$ (except for the above sparsity and unbiasedness constraints).

When we say that an unbiased m-sparsification Q of p is *efficient*, we mean that it minimizes $\mathbf{E}[\mathrm{Div}(Q, p)]$ among all unbiased m-sparsifications of p, where Div is a given *divergence* function. But what exactly is a divergence function? There are many different notions of divergence in the literature [14]; while we are not able to handle every such notion, our results cover two wide classes of functions.

Definition I.3. Let X be a convex subset of $\mathbb{R}^n$ and let $\mathrm{Div} : X \times X \to \mathbb{R}$ be a function. For fixed p, we write D for the function $D(q) := \mathrm{Div}(q, p)$.

1) Div is *convex* if for every fixed p, D is a convex function of q. We say Div is *strictly convex* if D is twice differentiable[1] and its Hessian matrix is positive definite everywhere.

---

[1]"Twice differentiable" means that the second-order Fréchet derivative exists everywhere [15, Chapter VIII, Section 12]. In the literature, F is not always assumed to be twice differentiable, but then strange things can occur [16] that we prefer to ignore in this paper. We should also emphasize that when we say that Div is "strictly convex", we require only that D is strictly convex for each fixed p, and not that Div is a strictly convex function jointly in q and p.
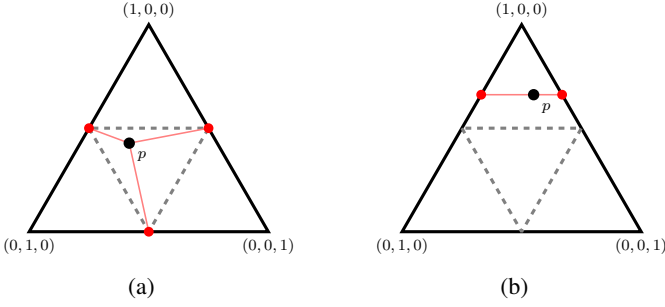
Fig. 1: Illustrations of US-PI over the probability simplex when $n = 3$ and $m = 2$.

2) Div is *additively separable* if for every fixed p there are functions $f_1, \ldots, f_n$ (possibly depending on p) such that

$$D(q) = \sum_{i=1}^{n} f_i(q_i). \tag{3}$$

3) Div is *permutation-invariant* if for every fixed p there is a function $F \colon X \to \mathbb{R}$ and $\alpha \in \mathbb{R}^n$ (both possibly depending on p) such that

$$\mathrm{Div}(q, p) = F(q) + \alpha \cdot (q - p) \tag{4}$$

and F is permutation-invariant in q; i.e., $F(q) = F(\sigma(q))$ where $\sigma(q) := (q_{\sigma(1)}, \ldots, q_{\sigma(n)})$ is any vector obtained from q via a permutation σ of the coordinates.

Examples of divergences that are strictly convex, additively separable, and permutation-invariant include squared Euclidean distance

$$\mathrm{Div}(q, p) = \|q - p\|^2 = \sum_i (q_i - p_i)^2,$$

and the Kullback–Leibler divergence

$$\mathrm{Div}(q, p) = D_{\mathrm{KL}}(q \parallel p) = \sum_i q_i \log(q_i / p_i).$$

More generally, a *Bregman divergence* [14] by definition has the form

$$\mathrm{Div}(q, p) = F(q) - F(p) - (\nabla F(p)) \cdot (q - p) \tag{5}$$

for some strictly convex function F called the *Bregman generator*. If F is additively separable (respectively, permutation-invariant in q) then the Bregman divergence is additively separable (respectively, permutation-invariant).

Similarly, any *f-divergence*

$$\mathrm{Div}(q, p) = \sum_i q_i f(p_i / q_i)$$

(for convex f) is additively separable, but f-divergences are not typically permutation-invariant. Note that our definition of permutation-invariant may be somewhat counterintuitive; for instance, the (un-squared) Euclidean distance $\mathrm{Div}(q, p) = \|q - p\|$ is *not* permutation-invariant by our definition (nor is it additively separable).

In order to ensure that $\mathbf{E}[\mathrm{Div}(Q, p)]$ makes sense for sparse Q, we require that Div be defined (and finite) on the (closed) orthant containing p. Our results still apply in many cases when the derivatives of the divergence are infinite on the boundary (as in the case of Kullback–Leibler divergence).

Our first main result, given in section II, is that for convex, permutation-invariant divergences, efficient m-sparsifications of p over the probability simplex are given by the following simple algorithm.

Algorithm. *Unbiased Sparsification for Permutation-Invariant Divergences (US-PI).*
Without loss of generality, reorder the coordinates of $p \in \mathbb{R}_{>0}^n$ so that $p_1 \geq \cdots \geq p_n > 0$. Let $H = \{1, 2, \ldots, h\}$ be the m-heavy indices and let

$$l := \frac{1}{m - h} \sum_{j=h+1}^{n} p_j.$$

Sample $m - h$ of the indices $\{h + 1, h + 2, \ldots, n\}$ with specified marginals proportional to p_i ($h + 1 \leq i \leq n$), and call the sample set I. Return the vector $Q \in \mathbb{R}^n$ whose coordinates are given by

$$Q_i = \begin{cases} p_i, & \text{if } i \leq h; \\ l, & \text{if } i \in I; \\ 0, & \text{otherwise.} \end{cases}$$

Figure 1 illustrates the possible 2-sparsifications of $p \in \mathbb{R}_{>0}^3$ that may result from US-PI. In (a), p has no heavy indices and so US-PI yields three possible values for the sparsifications of p: $(1/2, 1/2, 0)$, $(1/2, 0, 1/2)$, and $(0, 1/2, 1/2)$. In (b), the first index of p is heavy, and so the algorithm yields only two possible values for the sparsifications of p: $(p_1, 1 - p_1, 0)$ and $(p_1, 0, 1 - p_1)$.

This characterization has two remarkable features: first, the efficient sparsifications are *independent* of the divergence that is being optimized, beyond its convexity and permutation-invariance; second, we find that the same random variable Q minimizes the expected divergence even if we drop the simplex constraint $\sum_i Q_i = 1$. (Indeed, we conjecture that dropping the constraint $Q_i \geq 0$ also does not change the optimal Q.)

Our second main result, given in section III, is a similar characterization of the efficient sparsifications of $p \in \mathbb{R}^n$ in the case where the divergence is convex and additively separable (but not necessarily permutation-invariant). In this case the efficient sparsifications *do* depend on the choice of divergence, and they do not typically satisfy $\sum_i Q_i = \sum_i p_i$. On the other hand, for additively separable divergences we can show that the efficient sparsifications of p necessarily take on values in the same orthant as p.

II. PERMUTATION-INVARIANT DIVERGENCES

Our first step is to use convexity to reduce the problem to *concentrated* distributions, meaning they are supported on a finite (and bounded) number of values. This allows us to give a finite-dimensional parametrization of the problem.

Unfortunately, this parametrization is no longer convex as stated. We nevertheless show that critical points of the Lagrangian must correspond to efficient sparsifications.

Finally, we show that if Div is permutation-invariant then the solutions corresponding to the random variable given by US-PI are indeed critical points, yielding our desired result.

A. Facet Concentration

Here we fix $p \in \mathbb{R}^n_{>0}$ and assume that $\mathrm{Div}(q, p) = D(q)$ is convex. We also assume (without loss of generality) that $\sum_i p_i = 1$, and we are interested in unbiased sparsifications Q of p which also take values in the simplex $\Delta = \{x \in \mathbb{R}^n \mid x_i \geq 0, \ \sum_i x_i = 1\}$.

Throughout, the symbols I and J will represent subsets of $\{1, 2, \ldots, n\}$ of cardinality m. For each I, let

$$\Delta^I = \{x \in \mathbb{R}^n_{\geq 0} \mid \mathrm{I}(x) = I\}$$
$$= \{x \in \mathbb{R}^n_{\geq 0} \mid x_i > 0 \text{ for } i \in I, \ x_i = 0 \text{ for } i \notin I\}$$

be the (open) facet of Δ consisting of the points whose nonzero coordinates are those with indices in I. We note that the Δ^I are pairwise-disjoint convex bodies. An unbiased sparsification[2] on Δ is a random variable Q taking values in $\cup_I \Delta^I$ such that $\mathbf{E}[Q] = p$.

Lemma 1 (Facet concentration). *Assume that* Div *is convex, and let* Q *be an* m-*sparsification of* $p \in \mathbb{R}^n_{>0}$. *For each* I *with* $\Pr(Q \in \Delta^I) > 0$, *write* $q^I := \mathbf{E}[Q | Q \in \Delta^I]$. *Then the sparsification* Q' *such that* $\Pr(Q' \in \Delta^I) = \Pr(Q \in \Delta^I)$ *and* $\Pr(Q' = q^I | Q' \in \Delta^I) = 1$ *satisfies* $\mathbf{E}[\mathrm{Div}(Q', p)] \leq \mathbf{E}[\mathrm{Div}(Q, p)]$.

We call such a Q *(facet-)concentrated*.

Proof. See Appendix A-A. $\square$

We can parametrize the concentrated sparsifications in terms of the facet probabilities

$$x_I = \Pr(Q \in \Delta^I) \qquad x_I \in \mathbb{R}_{\geq 0}$$

along with the support points

$$y^I = \mathbf{E}[Q \mid Q \in \Delta^I] \qquad y^I \in \Delta^I.$$

So now we have a finite-dimensional problem:

Problem. *Sparse Concentrated Distribution Optimization (SCDO).*
For strictly convex $D : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$,

$$\text{minimize} \quad f(x_I, y^I) := \mathbf{E}[D(Q)] = \sum_I x_I D(y^I) \quad \text{subject to}$$

$$x_I \geq 0$$

$$S(x_I, y^I) := \sum_I x_I = 1$$

$$G_i(x_I, y^I) := \sum_I x_I y^I_i = p_i \quad \text{for all } i$$

$$y^I_i = 0 \quad \text{for all } i \notin I$$

$$\text{and} \quad y^I_i \geq 0 \quad \text{for all } I, i.$$

We write $Q \sim \mathcal{C}(x_I, y^I)$ to denote that the random variable Q has the concentrated distribution corresponding to $(x_I, y^I)_I$, i.e., Q takes on value $y^I \in \Delta^I$ with probability x_I. Note that there may be many choices of $(x_I, y^I)_I$ corresponding to the same random variable Q; in particular, if $x_I = 0$ then the choice of y^I has no effect on the resulting distribution.

---

[2]Note that here our Q is implicitly constrained to have *exactly* m nonzero coordinates, whereas earlier we allowed *at most* m nonzero coordinates. This assumption serves to prevent the argument from becoming needlessly complicated. It can be shown that it is never optimal to use fewer than m nonzero coordinates.

B. Optimality of Critical Points

As written, the objective function of SCDO is convex, but its constraints are not. Therefore we cannot use techniques from convex optimization straight out of the box. Fortunately, in our case we find that a critical point of the Lagrangian still suffices to give a global optimum.

Lemma 2. *Suppose that* D *is smooth and that* Q *is a concentrated unbiased sparsification of* p. *Suppose that for all* $(x_I, y^I)_I$ *with* $Q \sim \mathcal{C}(x_I, y^I)$ *(i.e., where* Q *is* $y^I \in \Delta^I$ *with probability* x_I*) there exist* $\nu, \lambda_i \in \mathbb{R}$, *and* $\mu_I \in \mathbb{R}_{\geq 0}$ *such that*

$$\nabla f(x_I, y^I) = \nu \nabla S(x_I, y^I) + \sum_{i=1}^n \lambda_i \nabla G_i(x_I, y^I) + \sum_{I : x_I = 0} \mu_I \nabla x_I.$$

Then Q *is an efficient unbiased sparsification of* p.

Proof. See Appendix A-B. $\square$

C. Solving SCDO for Permutation-Invariant Divergences

Note that all of our results up to this point have only relied on the convexity of $D(q) = \mathrm{Div}(q, p)$. We now assume that Div is also permutation-invariant, so that $D(q) = F(q) + \alpha \cdot (q - p)$, where $\alpha \in \mathbb{R}^n$ and $F(q) = F(\sigma(q))$ is invariant under permutations σ of the coordinates of q. First note that for any unbiased sparsification Q of p,

$$\mathbf{E}[D(Q)] = \mathbf{E}[F(Q)] + \alpha \cdot (\mathbf{E}[Q] - p) = \mathbf{E}[F(Q)],$$

so it is equivalent to minimize $\mathbf{E}[F(Q)]$.

Without loss of generality assume that $p_1 \geq p_2 \geq \cdots \geq p_n > 0$. We define a family of "preservative" unbiased sparsifications. Later we will show that preservative unbiased sparsifications are efficient.

Definition II.1. Let $H = \{1, 2, \ldots, h\}$ be the set of m-heavy indices, as defined in Definition I.1. Let $\ell = \frac{1}{m-h} \sum_{i > h} p_i$, and recall that $p_h > \ell \geq p_{h+1}$ and $h < m$. Denote

$$\tilde{y}^I_i := \begin{cases} p_i & i \in I \cap H \\ \ell & i \in I \setminus H \\ 0 & i \notin I. \end{cases}$$

We say a concentrated unbiased m-sparsification Q of p is *preservative* if $Q \sim \mathcal{C}(\tilde{x}_I, \tilde{y}^I)$ for some $\tilde{x}_I$ with $\tilde{x}_I = 0$ for all $I \not\supset H$.

Note that the $\tilde{x}_I$ in a preservative unbiased m-sparsification of p must satisfy the unbiasedness constraints $\sum_{I \supset H} \tilde{x}_I \tilde{y}^I = p$. Plugging in the definition of $\tilde{y}^I$, we find that the unbiasedness constraints are equivalent to constraints on the marginal inclusion probabilities:

$$s_i := \sum_{I \ni i} \tilde{x}_I = \begin{cases} 1 & i \in H \\ \frac{p_i}{\ell} & i \notin H. \end{cases}$$

In other words, the preservative Q are precisely those which are produced by US-PI: the random variable $I = \mathrm{I}(Q)$ always contains all of the heavy indices and contains each light index i with probability proportional to p_i.

Theorem 1. *If* Div *is convex and permutation-invariant, then the preservative unbiased* m-*sparsifications* Q *of* $p \in \mathbb{R}^n_{>0}$ *are efficient sparsifications.*

Proof. See Appendix A-C and Appendix A-D. $\square$

III. Additively Separable Divergences

Now we turn to the case where $D(Q) = \sum_i f_i(Q_i)$ is strictly convex and additively separable, but not necessarily permutation-invariant. We also remove the constraint that p and Q should lie in the probability simplex (although we still write $p \in \mathbb{R}^n_{>0}$ without loss of generality.) Our proof takes a similar route to the the permutation-invariant case. We begin by strengthening Lemma 1 to *coordinate concentration*, meaning that if Q is an EUS (with no constraint on the sum of the coordinates of Q) then for each coordinate i there is only one possible nonzero value that Q_i can take. While Lemma 1 already allowed us to reduce our problem to a finite (but exponentially large) number of dimensions, coordinate concentration reduces it further to a *convex*[3] optimization problem in n variables—the inclusion probabilities s_i.

This allows us to solve this convex optimization problem via a straightforward application of Lagrange multipliers and the Karush–Kuhn–Tucker (KKT) conditions [17, Section 5.5.3]. The details of coordinate concentration and the remainder of the additively separable case are deferred to Appendix B.

We can characterize the efficient sparsifications of p with respect to a separable divergence $\mathrm{Div}(q, p) = \sum_i f_i(q_i)$ as follows:

Algorithm. *Unbiased Sparsification for Additively Separable Divergences (US-AS).*

Given $\mathrm{Div}(q, p) = D(q) = \sum_i f_i(q_i)$ strictly convex, define $g_i(x) = xf_i'(x) - f_i(x) + f_i(0)$. Let $\lambda > 0$ be the unique value such that

$$\sum_i \min\left(1, \frac{p_i}{g_i^{-1}(\lambda)}\right) = m.$$

Declare index i to be *heavy* if $g_i(p_i) \geq \lambda$ (and *light* otherwise), and let $h < m$ be the number of heavy indices. Sample $m - h$ of the light indices with specified marginals $s_i := p_i/g_i^{-1}(\lambda) < 1$, and call the sample set I. Return the vector $Q \in \mathbb{R}^n$ whose coordinates are given by

$$Q_i = \begin{cases} p_i, & \text{if } i \text{ is heavy;} \\ g_i^{-1}(\lambda), & \text{if } i \in I; \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 2. *Let* Div *be a strictly convex, additively separable divergence defined on* $\mathbb{R}^n$. *Then the efficient (with respect to* Div*) unbiased m-sparsifications of* $p \in \mathbb{R}^n_{>0}$ *are precisely those produced by US-AS.*

In general the sparsifications produced by this procedure do *not* satisfy the stronger constraint $\sum_i Q_i = \sum_i p_i$, but it is not hard to verify that if D is also permutation-invariant then this procedure aligns with US-PI and the resulting random variable does satisfy that constraint.

Note also that this result shows that (for additively separable divergences) an efficient sparsification of $p \in \mathbb{R}^n_{>0}$ must necessarily be nonnegative, whereas the proof of Theorem 1 required this constraint to be enforced artificially.

[3]In fact, the function we find ourselves needing to optimize has the form of an *f-divergence* [14].

IV. Generalizations

A. Weaker Notion of Sparsification

If we relax our definition of an m-sparsification from requiring $|\mathrm{I}(Q)| \leq m$ to merely $\mathbf{E}[|\mathrm{I}(Q)|] \leq m$ (as in [11], [8], [9], [10]) then we find that the efficient sparsifications are still characterized by the same marginal inclusion probabilities s_i, albeit with more leeway in the survivor sampling procedure.

B. Non-Positive Targets

What if our target vector p is allowed to have negative (or complex, or other vector-valued) entries? For this question to make sense we must assume that Div is defined on the cone generated by the coordinates of p. For instance, for $p \in \mathbb{R}^{n \times k}$ it is perfectly sensible to ask (as in the second federated learning example in section I) for the random variable Q in $\mathbb{R}^{n \times k}$ with at most $m < n$ nonzero *rows* which minimizes the expected squared Euclidean distance between p and Q. We can always flip the signs on the p-inputs of $\mathrm{Div}(q, p)$ to treat p as being in $\mathbb{R}^n_{\geq 0}$ and use the modified divergence in the optimization problem above. Flipping signs on the inputs preserves additive separability, so if Div is additively separable then US-AS still applies. But flipping signs may destroy permutation-invariance, so even if Div is permutation-invariant US-PI may not apply. If we are lucky and the sign-flipped divergence is still permutation-invariant (as in the example of suared Euclidean distance) then US-PI does apply— but note that the resulting sparsification only preserves the sum of the absolute values $\sum_i |Q_i| = \sum_i |p_i|$.

C. Nonzero Default Value

The federated learning literature is also interested in the case where the "default" value of Q_i may be some nonzero value $z = (z_1, \ldots, z_n)$ (typically the same for all i); that is, where $\mathrm{I}(Q) := \{i : Q_i \neq z_i\}$. In this case we can just replace Q, p, and Div with $\tilde{Q} = Q - z$, $\tilde{p} = p - z$, and $\widetilde{\mathrm{Div}}(\tilde{q}, \tilde{p}) = \mathrm{Div}(\tilde{q}+z, \tilde{p}+z) = \mathrm{Div}(q, p)$. Again, this transformation preserves additive separability but may destroy permutation-invariance and positivity of p. If one is going to allow a constant nonzero default value $z = (z_0, \ldots, z_0)$ then it is also interesting to optimize the expected divergence over the choice of z_0 (for given, fixed p).

D. Open Questions

We noted above that in the non-permutation-invariant case imposing the additional constraint $\sum_i Q_i = \sum_i p_i$ will change the optimal solution. What is the new optimum?

We also noted that the proof in section II only shows that the output of US-PI is efficient for permutation-invariant divergences among *nonnegative* sparsifications, whereas the proof in section III shows that the same output is efficient among *all* sparsifications as long as the divergence is also additively separable. We conjecture that the additional condition of additive separability is not actually necessary here.

Finally, we are also interested in knowing the answer for divergences which are neither permutation-invariant nor additively separable, for instance, the squared *Mahalanobis distance* given by $F(x) = x^T A x$ for a positive definite matrix A.

References

[1] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," *Advances in Neural Information Processing Systems*, pp. 2328 – 2336, 2013.

[2] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," *Advances in Neural Information Processing Systems*, pp. 2726 – 2734, 2014.

[3] M. Braverman, A. Garg, T. Ma, and H. L. Nguyen, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 1011 – 1020, 2016.

[4] L. P. Barnes, Y. Han, and A. Özgür, "Lower bounds for learning distributions under communication constraints via Fisher information," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 9583 – 9612, 2020.

[5] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv:1610.492v2*, 2017.

[6] P. K. et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1-2, pp. 1 – 210, 2021.

[7] L. P. Barnes, H. A. Inan, B. Isik, and A. Özgür, "rTop-k: A statistical estimation approach to distributed SGD," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 897 – 907, 2020.

[8] J. Konečný and P. Richtárik, "Randomized distributed mean estimation: accuracy vs. communication," *Frontiers in Applied Mathematics and Statistics*, vol. 4, no. 62, 2018.

[9] H. Wang, S. Sievert, Z. Charles, S. Liu, S. Wright, and D. Papailiopoulos, "ATOMO: Communication-efficient learning via atomic sparsification," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[10] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[11] W. Chen, S. Horváth, and P. Richtárik, "Optimal client sampling for federated learning," *Transactions on Machine Learning Research*, 2022.

[12] Y. Tillé, *Sampling Algorithms*. Springer, 2006.

[13] A. Barvinok, *A Course in Convexity*. American Mathematical Society, 2002.

[14] S. Amari, "Divergence, optimization and geometry," in *Neural Information Processing: 16th International Conference, ICONIP 2009, Part I*, ser. Lecture Notes in Computer Science, vol. 5863, C. S. Leung, M. Lee, and J. H. Chan, Eds. Springer, 2009, pp. 185–193.

[15] J. Dieudonné, *Foundations of Modern Analysis*. Academic Press, 1969.

[16] R. M. Dudley, "On second derivatives of convex functions," *Mathematica Scandinavica*, vol. 41, pp. 159–174, 1977.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[18] M. Ghomi, "The problem of optimal smoothing for convex functions," *Proceedings of the American Mathematical Society*, vol. 130, no. 8, pp. 2255–2259, March 2002.

[19] D. J. H. Garling, *Inequalities: A Journey into Linear Analysis*. Cambridge University Press, 2007.

APPENDIX A
PROOFS

A. Proof of Lemma 1

Suppose $Q \in \Delta$ is an unbiased sparsification of p. Let π denote the concentration map; i.e. $\pi(Q)$ is the random variable such that $\pi(Q) = \mathbf{E}[Q \mid Q \in \Delta^I] = q^I \in \Delta^I$ with probability $\Pr(Q \in \Delta^I)$. Then clearly $\mathbf{E}(\pi(Q)) = \mathbf{E}(Q) = p$ so, $\pi(Q)$ is also an unbiased sparsification of p. By Jensen's inequality, $\mathbf{E}[D(\pi(Q))] \leq \mathbf{E}[D(Q)]$ as desired.

B. Proof of Lemma 2

By way of contradiction, suppose $Q \sim \mathcal{C}(\alpha_I, u^I)$ is a concentrated unbiased sparsification of p which satisfies the premise of the theorem but fails to be efficient. That means, using Lemma 1, there is some other concentrated unbiased sparsification $Q' \sim \mathcal{C}(\beta_I, v^I)$ where $\mathbf{E}[D(Q')] < \mathbf{E}[D(Q)]$.

Since we need only find one contradictory (α_I, u^I), we choose to take $u^I = v^I$ whenever $\alpha_I = 0$ or $\beta_I = 0$. That is, for any facet Δ^I where Q has probability $\alpha_I = 0$ of appearing, we set the corresponding point u^I (which value Q never actually takes) equal to the point v^I which Q' may (or may never) take in Δ^I, and vice versa. (If both $\alpha_I = \beta_I = 0$, pick $u^I = v^I \in \Delta^I$ arbitrarily.)

For $0 \leq t \leq 1$, let Q_t be the random variable corresponding to the convex mixture of distributions of Q and Q', where for any measurable set $A \subset \cup_I \Delta^I$ we have

$$\Pr(Q_t \in A) = (1 - t)\Pr(Q \in A) + t\Pr(Q' \in A).$$

Thus $Q_0 = Q$ and $Q_1 = Q'$. Note that since Q and Q' are both unbiased sparsifications of p, it is clear that Q_t is also an unbiased sparsification of p. When $0 < t < 1$, the random variable Q_t is not necessarily concentrated, as it can take on up to two different values in Δ^I rather than just one. Let $g(t) := \mathbf{E}[D(Q_t)]$, and note that

$$g(t) = (1 - t)\mathbf{E}[D(Q)] + t\mathbf{E}[D(Q')]$$

is an affine function of t. In particular, $g'(0) = \mathbf{E}[D(Q')] - \mathbf{E}[D(Q)] < 0$.

Let $\pi(Q_t)$ denote the concentration of Q_t as above, and let $h(t) := \mathbf{E}[D(\pi(Q_t))]$. We have $h(t) \leq g(t)$ for all t by convexity of D. Also note that $h(0) = g(0)$ and $h(1) = g(1)$, since both Q and Q' are already concentrated.

Now we show that $h : [0,1] \rightarrow \mathbb{R}$ is a smooth function. Since Q_t takes value u^I with probability $(1 - t)\alpha_I$ and v^I with probability $t\beta_I$, we can calculate that $\pi(Q_t) \sim \mathcal{C}(x_I(t), y^I(t))$, where

$$x_I(t) := (1 - t)\alpha_I + t\beta_I,$$

$$y^I(t) := \begin{cases} \frac{(1-t)\alpha_I u^I + t\beta_I v^I}{(1-t)\alpha_I + t\beta_I} & \text{if } \alpha_I, \beta_I > 0 \\ v^I = u^I & \text{if } \alpha_I = 0 \text{ or } \beta_I = 0. \end{cases}$$

It is clear that $x_I(t)$ and $y^I(t)$ are smooth in t for any given I. We have

$$h(t) = \mathbf{E}[D(\pi(Q_t))] = \sum_{I:\alpha_I + \beta_I > 0} x_I(t)D(y^I(t)).$$

Since D is smooth and the maps $t \mapsto x_I(t)$ and $t \mapsto y^I(t)$ are smooth, we know that h is smooth.

Now,

$$\begin{aligned} h'(0) &= \lim_{t \to 0^+} \frac{h(t) - h(0)}{t} \\ &= \lim_{t \to 0^+} \frac{h(t) - g(0)}{t} \quad \text{(since } h(0) = g(0)) \\ &\leq \lim_{t \to 0^+} \frac{g(t) - g(0)}{t} \quad \text{(since } h(t) \leq g(t) \text{ and } t > 0) \\ &= g'(0) \\ &= \mathbf{E}[D(Q')] - \mathbf{E}[D(Q)] \\ &< 0. \end{aligned}$$

Since $\gamma(t) := (x_I(t), y^I(t))_I$ is smooth, we can apply the chain rule:

$$\begin{aligned} h'(0) &= (f \circ \gamma)'(0) \\ &= (\nabla f)(\gamma(0)) \cdot \gamma'(0) \\ &= (\nabla f)((\alpha_I, u^I)_I) \cdot \gamma'(0) \\ &= \left(\nu \nabla S + \sum_{i=1}^{n} \lambda_i \nabla G_i + \sum_{I:\alpha_I = 0} \mu_I \nabla x_I \right) \cdot \gamma'(0). \end{aligned}$$

But $\gamma(t)$ satisfies all the constraints, and hence $\gamma'(0)$ is perpendicular to the gradients ∇S and ∇G_j. Furthermore, we also have that $\gamma'(0)$ is nonnegative on all the x_I components for which $\alpha_I = 0$, due to the inequality constraints $x_I \geq 0$. Since every such $\mu_I \geq 0$, we conclude that $(\mu_I \nabla x_I) \cdot \gamma'(0) \geq 0$. Hence $h'(0) \geq 0$, but this contradicts $h'(0) < \mathbf{E}[D(Q')] - \mathbf{E}[D(Q)] < 0$.

C. Proof of Theorem 1

For now, assume that F is smooth. Following Lemma 2, for every $(x_I, y^I)_I$ corresponding to a preservative sparsification Q, we will find multipliers $\nu, \lambda_i \in \mathbb{R}$ and $\mu_I \in \mathbb{R}_{\geq 0}$ such that

$$\nabla f(x_I, y^I) = \nu \nabla S(x_I, y^I) + \sum_{i=1}^{n} \lambda_i \nabla G_i(x_I, y^I) + \sum_{I:x_I = 0} \mu_I \nabla x_I,$$

where

$$f(x_I, y^I) = \sum_I x_I F(y^I),$$

$$S(x_I, y^I) = \sum_I x_I,$$

$$\text{and} \quad G_i(x_I, y^I) = \sum_{I \ni i} x_I y_i^I$$

as in SCDO. Note that if $Q \sim \mathcal{C}(x^I, y^I)$ is preservative then $y^I = \tilde{y}^I$ whenever $x_I > 0$. It is straightforward to compute the various gradients:

	$f(x_I, y^I)$	$S(x_I, y^I)$	$G_j(x_I, y^I)$	x_J
$\frac{\partial}{\partial x_I}$	$F(y^I)$	1	y_j^I	$\delta_{J=I}$
$\frac{\partial}{\partial y_i^I}$	$x_I \frac{\partial F}{\partial q_i}(y^I)$	0	$x_I \delta_{j=i}$	0

Let $I_0 = \{1, 2, \ldots, m\} \supseteq H$, so that

$$\tilde{y}^{I_0} = (\overbrace{p_1, p_2, \ldots, p_h}^{h}, \overbrace{\ell, \ell, \ldots, \ell}^{m-h}, \overbrace{0, 0, \ldots, 0}^{n-m}).$$

Note that $p_h \geq \ell$. Because F is permutation-invariant, we have $\frac{\partial F}{\partial q_i}(\tilde{y}^{I_0}) = \frac{\partial F}{\partial q_j}(\tilde{y}^{I_0})$ whenever $\tilde{y}_i^{I_0} = \tilde{y}_j^{I_0}$. Therefore we can write

$$\nabla F(\tilde{y}^{I_0}) = (\overbrace{a_1, a_2 \ldots, a_h}^{h}, \overbrace{b, b, \ldots, b}^{m-h}, \overbrace{c, c, \ldots, c}^{n-m}).$$

Now we determine the λ_j. A brief look at the gradient table shows that we must have

$$\lambda_j = \frac{\partial F}{\partial q_j}(\tilde{y}^I) \quad \text{if } j \in I \text{ and } x_I > 0.$$

In particular, as I varies over those m-element sets containing j where $x_I > 0$, we need $\frac{\partial F}{\partial q_j}(\tilde{y}^I)$ to be constant. Indeed, for any I with $x_I > 0$,

$$\frac{\partial F}{\partial q_j}(\tilde{y}^I) = \begin{cases} a_j & j \in H \subset I \\ b & j \in I \setminus H \\ 0 & j \notin I \end{cases}$$

does not depend on $I \ni j$. Hence we have

$$\lambda = (\overbrace{a_1, a_2, \ldots, a_h}^{h}, \overbrace{b, b, \ldots, b}^{n-h}).$$

With this λ we have matched ∇f on the $\frac{\partial}{\partial y_i^I}$ components, but we have not yet matched ∇f on the $\frac{\partial}{\partial x_I}$ components. For that, we will need to use a multiple ν of ∇S as well as nonnegative multiples μ_I of the ∇x_I wherever $x_I = 0$.

We want to show that if $(x_I, y^I)_I$ is preservative then there exist ν and $\mu_I \geq 0$ (with $\mu_I = 0$ whenever $x_I > 0$) such that:

$$F(y^I) = \nu + \lambda \cdot y^I + \mu_I.$$

If $x_I > 0$ then $\mu_I = 0$ and preservativity requires that $I \supset H$ and $y^I = \tilde{y}^I$. Permutation-invariance tells us that $F(\tilde{y}^I) = F(\tilde{y}^{I_0})$ and it is easy to see that $\lambda \cdot \tilde{y}^I = \lambda \cdot \tilde{y}^{I_0}$, so the constraints for $x_I > 0$ are satisfied by setting

$$\nu = F(\tilde{y}^I) - \lambda \cdot \tilde{y}^I = F(\tilde{y}^{I_0}) - \lambda \cdot \tilde{y}^{I_0}.$$

If $x_I = 0$ then $y^I \in \Delta^I$ is unconstrained by preservativity, and in order to have $\mu_I \geq 0$ we must show that

$$F(y^I) - \lambda \cdot y^I \geq \nu = F(\tilde{y}^{I_0}) - \lambda \cdot \tilde{y}^{I_0}. \tag{6}$$

If $I \supseteq H$, then $\lambda \cdot y^I = \nabla F(\tilde{y}^I) \cdot y^I$ (because λ and $\nabla F(\tilde{y}^I)$ agree wherever y^I is nonzero), so (6) holds by convexity of F:

$$F(y^I) \geq F(\tilde{y}^I) + \nabla F(\tilde{y}^I) \cdot (y^I - \tilde{y}^I)$$
$$= F(\tilde{y}^I) + \lambda \cdot (y^I - \tilde{y}^I)$$
$$\implies F(y^I) - \lambda \cdot y^I \geq F(\tilde{y}^I) - \lambda \cdot \tilde{y}^I = \nu.$$

It remains to show (6) for $I \not\supseteq H$. First note that F is Schur convex since it is symmetric and convex. Hence, $\frac{\partial F}{\partial q_i}(y) \geq \frac{\partial F}{\partial q_j}(y)$ whenever $y_i \geq y_j$. Since the components of $\tilde{y}^{I_0}$ are decreasing, we know that the components of $\nabla F(\tilde{y}^{I_0})$ are decreasing:

$$a_1 \geq \cdots \geq a_h \geq b \geq c.$$

In particular, λ is decreasing.

Let σ be a permutation such that $\sigma(y^I)$ is decreasing. Then $\sigma(y^I) \in \Delta^{I_0}$. Furthermore, since λ is decreasing and $y^I \geq 0$, σ is the permutation which maximizes the inner product $\lambda \cdot \sigma(y^I)$. Thus we have:

$$F(y^I) - \lambda \cdot y^I = F(\sigma(y^I)) - \lambda \cdot y^I \quad \text{(F is perm.-inv.)}$$
$$\geq F(\sigma(y^I)) - \lambda \cdot \sigma(y^I) \quad \text{(λ is decreasing)}$$
$$\geq F(\tilde{y}^{I_0}) - \lambda \cdot \tilde{y}^{I_0} = \nu \quad (\sigma(y) \in \Delta^{I_0}).$$

We have met the premise of Lemma 2 and therefore shown that any preservative Q is efficient.

D. Removing the Smoothness Condition

Now we turn the general case, where F is convex and permutation-invariant but not necessarily smooth. We note that any optimal Q has the same total sum as p, hence we may restrict the domain of F to the simplex $A = \{x \in \mathbb{R}^n_{\geq 0} \mid \sum_{i=1}^n x_i \leq \sum_{i=1}^n p_i\}$.

We first show that F can be approximated by a smooth $\widetilde{F}$ on the domain A, where $\widetilde{F}$ is also convex and permutation-invariant. The main idea is to shrink the domain A a little to A', to give "wiggle-room", and then convolve F with smooth density function θ with small support, yielding a smooth approximation G to F which is defined on A'. The smooth approximation G remains convex because it is a mixture of translates of F which are all convex themselves. This portion of our argument is taken from §2 of [18]. We then choose an affine contraction $R : A \to A'$ which only moves points a slight distance. We define $H(x) = G \circ R$, which is smooth and convex, and approximately equal to F. However, H is not permutation-invariant, so we define $\widetilde{F}(x) = \frac{1}{n!}\sum_{\sigma \in S_n} H(\sigma(x))$ to be the average of H over all permutations. The permutation-invariant $\widetilde{F}$ is convex, and is even closer to F than was H, since F is permutation-invariant.

For $\delta > 0$ let $B_\delta(0) = \{y \in \mathbb{R}^n \mid \|y\|_2 < \delta\}$ be the ball of radius δ centered at 0. Let $A_\delta = \{x \in A \mid x + y \in A \text{ for all } y \in B_\delta(0)\}$. Note that A_δ is convex. Let δ_0 be small enough so that $|F(x_1) - F(x_2)| < \epsilon/2$ whenever $x_1, x_2 \in A$ and $|x_1 - x_2| < \delta_0$. Let $a \in A$ be the mean of A. Consider the affine contraction $R_t(x) = a + t(x - a)$ where $0 < t < 1$. There exists some t such that $\|R_t(x) - x\| < \delta_0$ for all $x \in A$. In particular, we have $|F(R_t(x)) - F(x)| < \epsilon/2$ for all $x \in A$. Furthermore, there exists $\delta_1 > 0$ such that the image of R_t is contained in A_{δ_1}.

Suppose that A_δ is nonempty (this is true if δ is sufficiently small). Let θ be a smooth density function supported in $B_\delta(0)$. For $x \in A_\delta$, define $G_\delta(x) = \int_{y \in B_\delta(0)} F(x - y)\theta(y)dy$. Then G_δ is both convex and smooth. Furthermore, $G_\delta(x) - F(x) = \left(\int_{y \in B_\delta(0)} F(x-y)\theta(y)dy\right) - F(x) = \int_{y \in B_\delta(0)}(F(x-y) - F(x))\theta(y)dy$. In particular, we have that $|G_\delta(x) - F(x)| < \epsilon/2$ for all $x \in A_\delta$ for $\delta \leq \delta_0$.

Now take $\delta = \min(\delta_0, \delta_1)$. Suppose $x \in A$. We know that $R_t(x) \in A_{\delta_1} \subset A_\delta$. Hence $R_t(x)$ lies in the domain of G_δ. We know that $|G_\delta(R_t(x)) - F(R_t(x))| < \epsilon/2$. Furthermore, $|F(R_t(x)) - F(x)| < \epsilon/2$, and so $|G_\delta(R_t(x)) - F(x)| < \epsilon$. Now $H := G_\delta \circ R_t$ is convex, smooth, and $|H(x) - F(x)| < \epsilon$ for all $x \in A$. Finally we define $\widetilde{F}(x) = \frac{1}{n!}\sum_{\sigma \in S_n} H(\sigma(x))$. Then $\widetilde{F}$ is convex, smooth, and permutation-invariant. Furthermore, $|\widetilde{F}(x) - F(x)| = \frac{1}{n!}\left|\sum_{\sigma \in S_n} H(\sigma(x)) - F(\sigma(x))\right| < \epsilon$.

There is a sequence $\widetilde{F}_k : A \to \mathbb{R}$ such that each $\widetilde{F}_k$ is smooth, convex, and permutation-invariant, where $\widetilde{F}_k(x) \to F(x)$ as $k \to \infty$. Suppose that Q' is any unbiased sparsification. Let Q be preservative; i.e. any Q as defined by US-PI. Then Q is simultaneously optimal for all the $\widetilde{F}_k$. The domain A is compact and so $\mathbf{E}[\widetilde{F}_k(Q')] \to \mathbf{E}[F(Q')]$ and $\mathbf{E}[\widetilde{F}_k(Q)] \to \mathbf{E}[F(Q)]$ as $k \to \infty$. Since $\mathbf{E}[\widetilde{F}_k(Q')] \geq \mathbf{E}[\widetilde{F}_k(Q)]$ for all k, we have $\mathbf{E}[F(Q')] \geq \mathbf{E}[F(Q)]$. Hence Q is optimal for F.

APPENDIX B
ADDITIVELY SEPARABLE DIVERGENCES

A. Coordinate Concentration

Lemma 3 (Coordinate concentration). *Assume that* Div *is strictly convex and additively separable. Let Q be an efficient unbiased*

m-sparsification of $p \in \mathbb{R}^n_{>0}$. Then for all i, there exists $q_i > 0$ such that $\Pr(Q_i = q_i \mid Q_i \neq 0) = 1$. That is, any efficient unbiased m-sparsification of p is concentrated on a unique nonzero value in each coordinate.

Proof. Consider any unbiased m-sparsification Q of p. For any i, if $\Pr(Q_i \neq 0) = 0$ then $p_i = \mathbf{E}[Q_i] = 0$, whereas we are assuming that $p_i > 0$. So it is legitimate to condition on $Q_i \neq 0$, since this is an event with positive probability, and hence we can define $q_i := \mathbf{E}[Q_i \mid Q_i \neq 0]$. Note that $p_i = \mathbf{E}[Q_i] = q_i \Pr[Q_i \neq 0]$, so in particular $p_i > 0$ implies $q_i > 0$.

Fix any index i, and let Q' be the random variable obtained as follows: first sample $\hat{q} \leftarrow Q$; if $\hat{q}_i = 0$ return $\hat{q}$; otherwise replace the i-th coordinate of $\hat{q}$ with q_i and return the result. For simplicity of notation, in what follows, we assume without loss of generality that $i = 1$.

It is easy to check that $\mathbf{E}[Q'] = \mathbf{E}[Q]$ and $|\mathrm{I}(Q')| = |\mathrm{I}(\hat{q})| = m$, so Q' is also an unbiased m-sparsification of p. (But note that $\sum_j Q'_j$ is in general not equal to $\sum_j Q_j$, so even if Q takes values in the probability simplex, Q' usually does not).

By hypothesis, Div is additively separable. Because Div is strictly convex, the functions f_i in Equation (3) are strictly convex. By linearity of expectation,

$$\mathbf{E}[D(Q') \mid Q_1 \neq 0]$$
$$= \mathbf{E}[D(q_1, Q_2, Q_3, \ldots, Q_n) \mid Q_1 \neq 0]$$
$$= \mathbf{E}[f_1(q_1) + f_2(Q_2) + f_3(Q_3) + \cdots + f_n(Q_n)) \mid Q_1 \neq 0]$$
$$= \mathbf{E}[f_1(q_1) \mid Q_1 \neq 0] + \mathbf{E}[f_2(Q_2) \mid Q_1 \neq 0]$$
$$\quad + \cdots + \mathbf{E}[f_n(Q_n) \mid Q_1 \neq 0]$$
$$= f_1(\mathbf{E}[Q_1 \mid Q_1 \neq 0]) + \mathbf{E}[f_2(Q_2) \mid Q_1 \neq 0]$$
$$\quad + \cdots + \mathbf{E}[f_n(Q_n) \mid Q_1 \neq 0].$$

Similarly,

$$\mathbf{E}[D(Q) \mid Q_1 \neq 0] = \mathbf{E}[f_1(Q_1) \mid Q_1 \neq 0]$$
$$+ \mathbf{E}[f_2(Q_2) \mid Q_1 \neq 0]$$
$$+ \cdots + \mathbf{E}[f_n(Q_n) \mid Q_1 \neq 0].$$

Since f_1 is convex, Jensen's inequality [19, Theorem 4.2.1] implies that

$$f_1(\mathbf{E}[Q_1 \mid Q_1 \neq 0]) \leq \mathbf{E}[f_1(Q_1) \mid Q_1 \neq 0]. \tag{7}$$

Comparing the expressions for $\mathbf{E}[D(Q') \mid Q_1 \neq 0]$ and $\mathbf{E}[D(Q) \mid Q_1 \neq 0]$, we see that (7) is equivalent to

$$\mathbf{E}[D(Q') \mid Q_1 \neq 0] \leq \mathbf{E}[D(Q) \mid Q_1 \neq 0]. \tag{8}$$

Now

$$\mathbf{E}[D(Q')] = \Pr(Q'_1 = 0)\mathbf{E}[D(Q') \mid Q'_1 = 0]$$
$$+ \Pr(Q'_1 \neq 0)\mathbf{E}[D(Q') \mid Q'_1 \neq 0]$$
$$= \Pr(Q_1 = 0)\mathbf{E}[D(Q) \mid Q_1 = 0]$$
$$+ \Pr(Q_1 \neq 0)\mathbf{E}[D(Q') \mid Q_1 \neq 0],$$

while

$$\mathbf{E}[D(Q)] = \Pr(Q_1 = 0)\mathbf{E}[D(Q) \mid Q_1 = 0]$$
$$+ \Pr(Q_1 \neq 0)\mathbf{E}[D(Q) \mid Q_1 \neq 0].$$

Therefore, (8) implies that

$$\mathbf{E}[D(Q')] \leq \mathbf{E}[D(Q)]. \tag{9}$$

But Q is efficient, so (9) must in fact be an equality. This forces (8) to be an equality, which in turn forces (7) to be an equality. But since f_1 is *strictly* convex, Jensen's inequality [19, Theorem 4.2.1] is an equality only if the random variable $(Q_1 \mid Q_1 \neq 0)$ is concentrated on its mean value.

$\square$

Coordinate concentration allows us to further reduce the task of finding efficient unbiased m-sparsifications to an n-dimensional problem, as follows. Let q_i be as in the statement of Lemma 3, and define $s_i := \Pr(Q_i \neq 0)$. Then the quantity we seek to minimize is

$$\mathbf{E}[D(Q)] = \sum_{i=1}^{n}\big(\Pr(Q_i = 0) \cdot f_i(0)$$
$$+ \Pr(Q_i \neq 0) \cdot \mathbf{E}[f_i(Q_i) \mid Q_i \neq 0]\big)$$
$$= \sum_{i=1}^{n}\big((1 - s_i)f_i(0) + s_i f_i(q_i)\big).$$

It follows directly from the definitions of s_i and q_i that $s_i q_i = \mathbf{E}[Q_i]$; on the other hand, unbiasedness means that $\mathbf{E}[Q_i] = p_i$. In the proof of Lemma 3, we noted that $s_i > 0$, so we may replace q_i with p_i/s_i. That is, we seek to minimize

$$\sum_{i=1}^{n}\big((1 - s_i)f_i(0) + s_i f_i(p_i/s_i)\big).$$

If we replace each function $f_i(x)$ with a constant shift $f_i(x) - c_i$ (where c_i can depend on p but not on x), then the value of the divergence just changes by a constant, which does not affect the optimization problem we are trying to solve. So by setting $c_i := f_i(0)$, we may assume without loss of generality that $f_i(0) = 0$ for all i. Hence we are reduced to finding unbiased m-sparsifications Q that minimize

$$\sum_{i=1}^{n} s_i f_i(p_i/s_i).$$

Now, the sum of the s_i is the expected number of nonzero entries of Q, which by definition is at most m. Therefore, we are led to consider the following optimization problem.

Problem. *Inclusion Probability Optimization (IPO).*
For strictly convex f_i with $f_i(0) = 0$,

$$\text{minimize} \quad \sum_{i=1}^{n} s_i f_i\left(\frac{p_i}{s_i}\right) \quad \text{subject to}$$
$$\sum_{i=1}^{n} s_i \leq m$$
$$\text{and} \quad 0 < s_i \leq 1 \quad \text{for all } i.$$

In Appendix B-B, we solve IPO. We show in particular that the optimal solution satisfies $\sum_i s_i = m$. We claim that we thereby characterize all efficient unbiased m-sparsifications. Why? Well, we have just argued that given any efficient unbiased m-sparsification Q, the quantities $s_i := \Pr(Q_i \neq 0)$ must yield an optimal solution to IPO. Conversely, given any optimal solution $\{s_i\}$ to IPO, we saw in subsection I-A that the conditions $\sum_i s_i = m$ and $0 < s_i \leq 1$ imply that it is possible to sample with the specified marginals $\{s_i\}$. Each way of sampling with

the specified marginals completely specifies a unique unbiased m-sparsification of p, which is guaranteed to be efficient since the $\{s_i\}$ constitute an optimal solution to IPO.

B. Solving IPO

As usual, we regard $p \in \mathbb{R}_{>0}^n$ as fixed. Let

$$F(s) := \sum_{i=1}^n s_i f_i \left(\frac{p_i}{s_i} \right)$$

be the function we are seeking to minimize. We show that F is strictly convex[4]. By direct computation,

$$\frac{\partial F(s)}{\partial s_i} = f_i \left(\frac{p_i}{s_i} \right) - \left(\frac{p_i}{s_i} \right) f_i' \left(\frac{p_i}{s_i} \right). \tag{10}$$

Then

$$\frac{\partial^2 F(s)}{\partial s_i^2} = -\frac{p_i}{s_i^2} f_i' \left(\frac{p_i}{s_i} \right) + \frac{p_i}{s_i^2} f_i' \left(\frac{p_i}{s_i} \right) + \left(\frac{p_i^2}{s_i^2} \right) f_i'' \left(\frac{p_i}{s_i} \right)$$

$$= \left(\frac{p_i^2}{s_i^2} \right) f_i'' \left(\frac{p_i}{s_i} \right) > 0,$$

where the final inequality follows because f_i is strictly convex and $s_i > 0$. So the Hessian is a diagonal matrix with strictly positive entries on the diagonal, and F is (strictly) convex.

Motivated by Equation 10, define $g_i(x) := x f_i'(x) - f_i(x)$, so that $\frac{\partial F}{\partial s_i}(s) = -g_i(p_i/s_i)$. A similar calculation to the one above shows that g_i is a strictly increasing function of $x > 0$. Moreover, $g_i(0) = 0$ because $f_i(0) = 0$, so $g_i(x) > 0$ for all $x > 0$.

The constraint that $s_i > 0$ is slightly awkward to deal with. Our approach is to pick some small $\epsilon > 0$ and replace the constraint $s_i > 0$ with $s_i \geq \epsilon$. We then show that for all sufficiently small ϵ, all optimal solutions are independent of ϵ and do not lie on $s_i = \epsilon$. Any feasible solution with $s_i > 0$ will be feasible for some $\epsilon > 0$, and hence its objective value cannot exceed the optimal value.

Following the standard recipe for convex optimization [17, Chapter 5], we define the Lagrangian

$$\mathcal{L}(s, \mu, \nu, \lambda) = F(s) + \sum_{i=1}^n \mu_i(s_i - 1) + \sum_{i=1}^n \nu_i(\epsilon - s_i)$$

$$+ \lambda \left(-m + \sum_{i=1}^n s_i \right),$$

where μ_i, ν_i, and λ are Lagrange multipliers. Let $\vec{e}_i$ denote the i-th unit vector, and let $\vec{1} := \sum_i \vec{e}_i$. Then

$$\nabla \mathcal{L} = (\nabla F)(s) + \sum_{i=1}^n \mu_i \vec{e}_i - \sum_{i=1}^n \nu_i \vec{e}_i + \lambda \vec{1}.$$

At an optimal point, the KKT conditions are (in addition to the condition that an optimal point be feasible)

$$-g_i \left(\frac{p_i}{s_i} \right) + \mu_i - \nu_i + \lambda = 0$$

$$\mu_i, \nu_i, \lambda \geq 0$$

$$\mu_i(s_i - 1) = 0$$

$$\nu_i(\epsilon - s_i) = 0$$

$$\lambda \left(-m + \sum_{i=1}^n s_i \right) = 0$$

---

[4]In fact, F is an *f-divergence* [14], and it is a standard fact that the (strict) convexity of f implies the (strict) convexity of F, but we give a proof anyway since it is short.

Given a proposed solution s, let $H := \{i: s_i = 1\}$ (the *heavy* indices), let $E := \{i: s_i = \epsilon\}$ (the *epsilon* indices), and let L denote the remaining (*light*) indices. We first claim that $L \neq \varnothing$. Suppose to the contrary that $L = \varnothing$. Now $m < n$ and $\sum_i s_i \leq m$, so H cannot comprise *all* the indices, and the remaining indices must be in E. But for all sufficiently small positive ϵ, $\sum_i s_i$ cannot be exactly equal to an integer m, so λ is forced to be zero. For any $i \in E$, we must have $\mu_i = 0$, so $g_i(p_i/\epsilon) + \nu_i = 0$, which implies that

$$g \left(\frac{p_i}{\epsilon} \right) = -\nu_i \leq 0. \tag{11}$$

But as we observed earlier, $g_i(p_i/\epsilon) > 0$ since $p_i > 0$. This contradiction shows that $L \neq \varnothing$.

For $i \in L$, $\mu_i = \nu_i = 0$, so

$$-g_i \left(\frac{p_i}{s_i} \right) + \lambda = 0.$$

Two inferences are immediate. First, $\lambda = g_i(p_i/s_i) > 0$, and hence

$$\sum_{i=1}^n s_i = m.$$

It follows that there cannot be any ϵ contribution to $\sum_i s_i$, so $E = \varnothing$. Second, $g_i(p_i/s_i) = \lambda$ is constant across all $i \in L$. By definition of L, we must have $1 > s_i = p_i/g_i^{-1}(\lambda)$ for all $i \in L$.

Conversely, for $i \in H$ we must have

$$g_i(p_i/s_i) - \lambda = \mu_i \geq 0$$

so $1 = s_i \leq p_i/g_i^{-1}(\lambda)$ (since g_i is increasing). That is, for all i, we have $s_i = \min(1, p_i/g_i^{-1}(\lambda))$.

The final constraint which we must satisfy (since $\lambda > 0$) is

$$m = \sum_i s_i = \sum_i \min(1, p_i/g_i^{-1}(\lambda)). \tag{12}$$

The right hand side is a continuous, decreasing function of λ with range $(0, n]$. Furthermore, it is strictly decreasing except where it is equal to $n > m$. Therefore there is a unique $\lambda > 0$ solving Equation 12 (which can easily be found by, say, binary search), which yields our desired optimum.