2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)

Approximation Algorithms for Stochastic Minimum-Norm Combinatorial Optimization

Sharat Ibrahimpur, Chaitanya Swamy
Department of Combinatorics and Optimization
University of Waterloo
Waterloo, Canada
{sharat.ibrahimpur,cswamy}@uwaterloo.ca

***Abstract*—Motivated by the need for, and growing interest in, modeling uncertainty in data, we introduce and study *stochastic minimum-norm optimization*. We have an underlying combinatorial optimization problem where the costs involved are *random variables* with given distributions; each feasible solution induces a random multidimensional cost vector, and given a certain objective function, the goal is to find a solution (that does not depend on the realizations of the costs) that minimizes the expected objective value. For instance, in stochastic load balancing, jobs with random processing times need to be assigned to machines, and the induced cost vector is the machine-load vector. The choice of objective is typically the maximum- or sum- of the entries of the cost vector, or in some cases some other ℓ_p norm of the cost vector. Recently, in the deterministic setting, Chakrabarty and Swamy [7] considered a much broader suite of objectives, wherein we seek to minimize the f-norm of the cost vector under a given *arbitrary monotone, symmetric norm* f. In stochastic minimum-norm optimization, we work with this broad class of objectives, and seek a solution that minimizes the *expected f-norm* of the induced cost vector.**

The class of monotone, symmetric norms is versatile and includes ℓ_p-norms, and Top_ℓ-norms (sum of ℓ largest coordinates in absolute value), and enjoys various closure properties; in particular, it can be used to incorporate *multiple* norm budget constraints, $f_\ell(x) \le B_\ell$, $\ell = 1, \ldots, k$.

We give a general framework for devising algorithms for stochastic minimum-norm combinatorial optimization, using which we obtain approximation algorithms for the stochastic minimum-norm versions of the load balancing and spanning tree problems. We obtain the following concrete results.

- **An $O(1)$-approximation for *stochastic minimum-norm load balancing on unrelated machines* with: (i) arbitrary monotone symmetric norms and job sizes that are Bernoulli random variables; and (ii) Top_ℓ norms and arbitrary job-size distributions.**
- **An $O(\log m/\log\log m)$-approximation for the general stochastic minimum-norm load balancing problem, where m is the number of machines.**
- **An $O(1)$-approximation for stochastic minimum-norm spanning tree with arbitrary monotone symmetric norms and arbitrary edge-weight distributions; this guarantee extends to the stochastic minimum-norm matroid basis problem.**

Two key technical contributions of this work are: (1) a structural result of independent interest connecting stochastic minimum-norm optimization to the simultaneous optimization of a (*small*) collection of expected Top_ℓ-norms; and (2) showing how to tackle expected Top_ℓ-norm minimization by leveraging techniques used to deal with minimizing the expected maximum, circumventing the difficulties posed by the non-separable nature of Top_ℓ norms.

A full version of the paper is available on the CS arXiv.

***Index Terms*—Approximation algorithms; stochastic optimization; norm minimization; load balancing; spanning trees**

I. Introduction

Uncertainty is a facet of many real-world decision environments, and a thriving and growing area of optimization, *stochastic optimization*, deals with optimization under uncertainty. Stochastic load-balancing and scheduling problems, where we have uncertain job sizes (a.k.a processing times), constitute a prominent and well-studied class of stochastic-optimization problems (see, e.g., [9], [10], [12], [14], [15], [23], [26]). In *stochastic load balancing*, we are given job-size distributions and we need to fix a job-to-machine assignment *without knowing the actual processing-time realizations*. This assignment induces a *random* load vector, and its quality is assessed by evaluating the expected objective value of this load vector, which one seeks to minimize; the objectives typically considered are: *makespan*—i.e., maximum load-vector entry—which leads to the stochastic makespan minimization problem [9], [10], [15], and the *ℓ_p-norm of the load vector* [24]. More generally, in a generic stochastic-optimization problem, we have an underlying combinatorial optimization problem and the costs involved are described by random variables with given distributions. We need to take decisions, i.e., find a feasible solution, given only the distributional information, and without knowing the realizations of the costs. (This is sometimes called *one-stage* stochastic optimization.) Each feasible solution induces a random cost vector, we have an objective that seeks to quantitatively measure the quality of the solution by aggregating the entries of the cost vector, and the goal is to find a feasible solution that minimizes the expected objective value of the induced cost vector. As another example in this setup, consider the *stochastic spanning tree* problem, which is the following basic stochastic network-design problem: we have a graph with random edge costs and we seek a spanning tree of low expected objective value, where the objective is applied to the cost vector that consists of the costs of edges in the tree.

The two most-commonly considered objectives in such settings (as also for deterministic problems) are: (a) the *maximum* cost-vector entry (which adopts an egalitarian view); and (b) the *sum* of the cost-vector entries (i.e., a utilitarian view that considers the total cost incurred). These objectives give rise to

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00094

various classical problems: besides the makespan minimization problem in load balancing, other examples include (deterministic/stochastic) bottleneck spanning tree (max- objective), minimum spanning tree (sum- objective), and the k-center (max- objective) and k-median (sum- objective) clustering problems. Recognizing that the max- and sum- objectives tend to skew solutions in different directions, other ℓ_p-norms of the cost vector (e.g., ℓ_2-norm) have also been considered in certain settings as a means of interpolating between, or trading off, the max- and sum- objectives. For instance, ℓ_p-norms have been investigated for both deterministic and stochastic load balancing [3], [4], [22], [24] and for deterministic k-clustering [11]. Very recently, Chakrabarty and Swamy [7] introduced a rather general model to unify these various problems (including minimizing ℓ_p-norms), that they call *minimum-norm optimization*: given an *arbitrary* monotone, symmetric norm f, find a solution that minimizes the f-norm of the induced cost vector.

A. Our contributions

In this work, we introduce and study *stochastic minimum-norm optimization*, which is the stochastic version of min-norm optimization: given a stochastic-optimization problem and an arbitrary monotone, symmetric norm f, we seek a feasible solution that minimizes the *expected f-norm of the induced cost vector*. As a model, this combines the versatility of (deterministic) min-norm optimization with the more realistic setting of uncertain data, thereby giving us a unified way of dealing with the various objective functions typically considered for (deterministic and stochastic) optimization problems in the face of uncertain data. We consider problems where there is a certain degree of independence in the underlying costs, so that the components of the underlying cost vector are always *independent* (and nonnegative) random variables.

Our chief contribution is a framework that we develop for designing algorithms for stochastic minimum-norm combinatorial optimization problems, using which we devise approximation algorithms for the stochastic minimum-norm versions of load balancing (Theorems 20 and 19) and spanning trees (Theorem 22). We only assume that we have a *value oracle* for the norm f; this is weaker than the optimization-oracle or first-order-oracle access required to f in [7], [8].

Stochastic minimum-norm optimization can be motivated from two distinct perspectives. The class of monotone, symmetric norms is quite rich and broad. In particular, it contains all ℓ_p-norms, as also another fundamental class of norms called Top_ℓ*-norms*: $\mathsf{Top}_\ell(x)$ is the sum of the ℓ largest coordinates of x (in absolute value). Notice that Top_ℓ norms provide another means of interpolating between the min-max (Top_1) and min-sum (Top_m) problems (where m is number of coordinates). One of the motivating factors for [7] to consider the (far-reaching) generalization of arbitrary monotone, symmetric norms (in the deterministic setting) was that it allows one to capture the optimization problems associated with these various objectives under one umbrella, and thereby come up with a unified set of techniques for handling these optimization problems. These same benefits also apply in the stochastic setting, making stochastic minimum-norm optimization an appealing model to study.

Another motivation comes from the fact that the class of monotone, symmetric norms is closed under various operations, including taking nonnegative combinations, and taking the maximum of any finite collection of monotone, symmetric norms. A noteworthy and non-evident consequence of these closure properties is that they allow us to incorporate budget constraints $f_\ell(x) \leq B_\ell$, $\ell = 1, \ldots, k$ involving *multiple* monotone, symmetric norms $f_1, \ldots, f_k$ using the min-norm optimization model: we can simply define another (monotone, symmetric) norm $g(x) := \max\{f_\ell(x)/B_\ell : \ell = 1, \ldots, k\}$, and the (single) budget constraint $g(x) \leq 1$ can be captured by the problem of minimizing $g(x)$. Multiple norm budget constraints may arise, or be useful, for example, when no single norm may be a clear choice for assessing the solution quality. Moreover, such constraints, and, in particular, the above means of capturing them, can be especially useful in stochastic settings, as they can provide us with more fine-grained control of the underlying random cost vector, which can help offset the risk associated with uncertainty; e.g., the constraint $\mathbf{E}\big[\max\{\mathsf{Top}_\ell(Y)/B_\ell : \ell = 1, \ldots, k\}\big] \leq 1$, enforces a fair bit of control on the random cost vector Y and provides safeguards against the costs being too high.

To elaborate on our framework and results, it is useful to highlight and appreciate two distinct high-level challenges that arise in dealing with stochastic min-norm optimization. We delve into more details in Section II. First, how do we reason about the expectation of an *arbitrary* monotone, symmetric norm? A useful structural result proved in [7] is that any monotone symmetric norm f can be expressed as the maximum of a collection of *ordered norms*, where an ordered norm is simply a nonnegative combination of Top_ℓ-norms (Theorem 3). While this does give a more concrete way of thinking about the expected f-norm, the challenge nevertheless in the stochastic setting is that one now needs to reason about the expectation of the maximum of a collection of random variables, where each random variable is the ordered norm of our cost vector Y. *One of our chief insights is that the expectation of the maximum of a collection of ordered norms is within an $O(1)$-factor of the maximum of the expected ordered norms* (Theorem 13), i.e., interchanging the expectation and maximum operators only loses an $O(1)$ factor! The crucial consequence is that this provides us with a ready means for reasoning about $\mathbf{E}\big[f(Y)\big]$, namely by controlling $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ for all indices ℓ (see Theorem 14). We believe that this structural result about the expectation of a monotone, symmetric norm is of independent interest.

This brings us to the second challenge: how do we deal with a specific norm, such as the Top_ℓ norm? To our knowledge, there is no prior work on any stochastic Top_ℓ-norm minimization problem, and we need to control *all* expected Top_ℓ norms. Our approach is based on carefully identifying certain statistics of the random vector Y that provide a convenient handle on

$\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ (see Section IV). (These statistics also play a role in establishing our above result on the expectation of a monotone, symmetric norm.) For a specific application (e.g., stochastic $\{$load balancing, spanning trees$\}$), we formulate an LP encoding (loosely speaking) that the statistics of our random cost vector match the statistics of the cost vector of an optimal solution. The main technical component is then to devise a rounding algorithm that rounds the LP solution while losing only a small factor in these statistics, and we utilize *iterative-rounding* ideas to achieve this.

Combining the above ingredients leads to our approximation guarantees for stochastic min-norm $\{$load balancing, spanning trees$\}$. Our strongest and most-sophisticated results are for stochastic min-norm load balancing with: (i) arbitrary monotone symmetric norms and Bernoulli job sizes (Theorem 19); and (ii) Top_ℓ norms and arbitrary job-size distributions (Theorem 18); *in both cases, we obtain constant-factor approximations*. (We emphasize that we have not attempted to optimize constants, and instead chosen to keep exposition simple and clean.) We also obtain an $O(\log m/\log\log m)$-approximation for general stochastic min-norm load balancing (Theorem 20). We remark that dealing with Bernoulli distributions is often considered to be a stepping stone towards handling general distributions (see, e.g., [15]), and so we believe that our techniques will eventually lead to a constant-factor approximation for (general) stochastic min-norm load balancing. For stochastic spanning trees, wherein edge costs are random, we obtain an $O(1)$-approximation for *arbitrary monotone symmetric norms and arbitrary distributions* (Theorem 22).

Related work. As mentioned earlier, stochastic load balancing is a prominent combinatorial-optimization problem that has been investigated in the stochastic setting under various ℓ_p norms. Kleinberg, Rabani, and Tardos [15] investigated *stochastic makesepan minimization* (i.e., minimize expected maximum load) in the setting of identical machines (i.e., the processing time of a job is the same across all machines), and were the first to obtain an $O(1)$-approximation for this problem. We utilize the tools that they developed to reason about the expected makespan. Their results were improved for specific job-size distributions by [9]. Almost two decades later, Gupta et al. [10] obtained the first $O(1)$-approximation for stochastic makespan minimization on unrelated machines, and an $O\left(\frac{p}{\log p}\right)$-approximation for minimizing the expected ℓ_p-norm of the load vector. Our $O(1)$-approximation result for all Top_ℓ-norms substantially generalizes the makespan-minimization (i.e., Top_1-norm) result of [10]. The latter guarantee was improved to a constant by Molinaro [24] via a careful and somewhat involved use of the so-called L-function method of Latala [17].

Our results and techniques are incomparable to those of [24]. At a high level, Molinaro argues that $\mathbf{E}\left[\|Y\|_p\right]$, where Y is the m-dimensional machine-load vector, can be bounded by controlling $\left(\sum_{i\in[m]}\mathbf{E}\left[Y_i^p\right]\right)^{1/p}$, and uses a notion of effective size due to Latala [17] to obtain a handle on $\mathbf{E}\left[Y_i^p\right]$ in terms of the X_{ij} random variables of the jobs assigned to machine i. Essentially, "effective size" of a random variable maps the random variable to a deterministic quantity that one can work with instead (but its definition and utility depends on a certain scale parameter). Previously, for stochastic makespan minimization, Kleinberg et al. [15] and Gupta et al. [10] utilized a notion of effective size due to Hui [13], which helps in controlling tail bounds of the machine loads. Molinaro leverages the full power of Latala's notion of effective size by applying it at multiple scales, which allows him to obtain an $O(1)$-approximation for ℓ_p norms. A pertinent question that perhaps arises is: given the success yielded by the various notions of effective sizes for ℓ_∞ and other ℓ_p norms, can one come up with a notion of effective size that one can use for a general monotone symmetric norm f? This however seems unlikely. A concrete reason for this can be gleaned from the modeling power of general monotone, symmetric norms that arises due to their closure properties. Recall that one can encode multiple monotone, symmetric-norm budget constraints via one monotone, symmetric norm f (by taking a maximum of the scaled constituent norms). A notion of effective size for f would (remarkably) translate to one deterministic quantity that simultaneously yields some control for all the norms involved in the budget constraints; this is unreasonable to expect, even when the constituent norms are ℓ_p norms.

Examples of other well-known combinatorial optimization problems that have been investigated in the stochastic setting include stochastic knapsack and bin packing [9], [15], [18], stochastic shortest paths [18]. The works of [18]–[20] consider expected-utility-maximization versions of various combinatorial optimization problems. In a sense, this can be viewed as a counterpart of stochastic min-norm optimization, where we have a *concave* utility function, and we seek to maximize the expected utility of the underlying random value vector induced by our solution. Their results are obtained by a clever discretization of the probability space; this does not seem to apply to stochastic minimization problems.

Top_ℓ- and ordered-norms have been proposed in the location-theory literature, as a means of interpolating between the k-center and k-median clustering problems, and have been studied in the Operations Research literature [16], [25], but largely from a modeling perspective. Recently they have received much interest in the algorithms and optimization communities—partly, because Top_ℓ norms yield an alternative (to ℓ_p norms) natural means of interpolating between the ℓ_∞- and ℓ_1- objectives—and this work has led to strong algorithmic results for Top_ℓ-norm- and ordered-norm- minimization in deterministic settings [1], [2], [5]–[8].

II. Technical overview and organization

We discuss here the various challenges that arise in stochastic min-norm optimization, and give an overview of the technical ideas we develop to overcome these challenges with pointers to the relevant sections for more details.

As noted in [7], even for *deterministic* min-norm optimization, the simple approach of minimizing $f(\vec{v})$, where $\vec{v}$ ranges over the cost of fractional solutions (e.g., convex combinations of integer solutions), fails badly since this convex program often has large integrality gap. In the stochastic setting, a further problem with this approach is that the random variable $f(Y)$ will typically have exponential-size support (note that Y follows a product distribution) making it computationally challenging to evaluate $\mathbf{E}\big[f(Y)\big]$.

As noted earlier, we face two types of challenges in tackling stochastic min-norm optimization. The first is posed by the generality of an arbitrary monotone, symmetric norm. While the work of [7] shows that Top_ℓ-norms are fundamental building blocks of monotone symmetric norms, and suggests a way forward for dealing with stochastic min-norm optimization, as we elaborate below, the stochastic nature of the problem throws up various new issues, that create significant difficulties with leveraging the tools in [7] to reason about the *expectation* of a monotone symmetric norm.

Second, stochastic min-norm optimization is complicated even for various *specific* norms. As mentioned earlier, $O(1)$-approximation algorithms were obtained only quite recently for stochastic min-norm load balancing under the ℓ_∞ norm (i.e., stochastic makespan-minimization) [10], and other ℓ_p norms [24]; there is no prior work on stochastic Top_ℓ-load balancing (or any other stochastic Top_ℓ-norm optimization problem). The difficulties arise again due to the underlying stochasticity; even with the ℓ_∞ norm, which is the most specialized of the aforementioned norms, one needs to bound the expectation of the maximum of a collection of random variables, a quantity that is not convenient to control. The importance of Top_ℓ norms in the deterministic setting indicates that stochastic Top_ℓ-norm optimization is a key special case that one needs to understand, but the non-separable nature of Top_ℓ norms adds another layer of difficulty.

Expectation of a monotone, symmetric norm (Section V). Recall that Chakrabarty and Swamy [7] show that for any monotone, symmetric norm $f : \mathbb{R}^m_{\geq 0} \to \mathbb{R}_{\geq 0}$, there is a collection $\mathcal{C} \subseteq \mathbb{R}^m_{\geq 0}$ of weight vectors with non-increasing coordinates such that $f(x) = \max_{w\in\mathcal{C}} w^T x^{\downarrow}$ for any $x \in \mathbb{R}^m_{\geq 0}$ (see Theorem 3), where $x^{\downarrow}$ is the vector x with coordinates sorted in non-increasing order. The quantity $w^T x^{\downarrow}$ is called an (w-) *ordered norm*, and can be expressed as a nonnegative combination $\sum_{\ell=1}^m (w_\ell - w_{\ell+1})\mathsf{Top}_\ell(x)$ of Top_ℓ norms (where $w_{m+1} := 0$). This structural result is quite useful for deterministic min-norm optimization, since it yields an easier-to-work-with objective, and more significantly, it *immediately* shows that controlling all Top_ℓ norms suffices to control $f(x)$; these properties are leveraged by [7] to devise approximation algorithms for deterministic min-norm optimization.

However, it is rather unclear how this structural result helps with the *stochastic* setting. If Y is the random cost vector of our solution (with independently-distributed coordinates) one can now rewrite the objective function $\mathbf{E}\big[f(Y)\big]$ as $\mathbf{E}\big[\sup_{w\in\mathcal{C}} w^T Y^{\downarrow}\big]$ but dealing with the latter quantity entails reasoning about the expectation of the maximum of a collection of (positively correlated) random variables, which is a potentially onerous task. Taking cue from the deterministic setting, a natural result to aim for in the stochastic setting is that, *bounding all expected* Top_ℓ *norms enables one to bound* $\mathbf{E}\big[f(Y)\big]$. But unlike the deterministic setting, reformulating the objective as $\mathbf{E}\big[\sup_{w\in\mathcal{C}} w^T Y^{\downarrow}\big]$ does not yield any apparent dividends, and it is not at all clear that such a result is actually true. The issue again is that it is difficult to reason about $\mathbf{E}\big[\sup \ldots\big]$, and interchanging the expectation and $\sup$ terms is not usually a viable option (it is not hard to see that $\mathbf{E}\big[\max\{Z_1, \ldots, Z_k\}\big]$ may in general be $\Omega(k)$ times $\max\{\mathbf{E}\big[Z_1\big], \ldots, \mathbf{E}\big[Z_k\big]\}$).

One of our chief contributions is to prove that the analogue mentioned above for the stochastic setting *does indeed hold*, i.e., controlling $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ for all $\ell \in [m]$ allows one to control $\mathbf{E}\big[f(Y)\big]$. The key to this, and our main technical result here, is that, somewhat surprisingly and intriguingly, $\mathbf{E}\big[f(Y)\big] = \mathbf{E}\big[\sup_{w\in\mathcal{C}} w^T Y^{\downarrow}\big]$ *is at most a constant factor larger than* $\sup_{w\in\mathcal{C}} \mathbf{E}\big[w^T Y^{\downarrow}\big]$ (Theorem 13). The quantity $\sup_{w\in\mathcal{C}} \mathbf{E}\big[w^T Y^{\downarrow}\big] = \sup_{w\in\mathcal{C}} w^T \mathbf{E}\big[Y^{\downarrow}\big]$ has a nice interpretation: it is simply $f\big(\mathbf{E}\big[Y^{\downarrow}\big]\big)$, and so this can be restated as $\mathbf{E}\big[f(Y)\big] = O(1) \cdot f\big(\mathbf{E}\big[Y^{\downarrow}\big]\big)$. This result provides us with the same mileage that [7] obtain in the deterministic setting. Since $\mathbf{E}\big[w^T Y^{\downarrow}\big]$ is $\sum_{\ell=1}^m (w_\ell - w_{\ell+1})\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$, as in the deterministic setting, this immediately implies that controlling $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ for all $\ell \in [m]$ allows us to control $\mathbf{E}\big[f(Y)\big]$, thereby providing a foothold for reasoning about the fairly general stochastic min-norm optimization problem. In particular, we infer that if $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] \leq \alpha \cdot \mathbf{E}\big[\mathsf{Top}_\ell(W)\big]$ for all $\ell \in [m]$ (and we can restrict to ℓs that are powers of 2 here), then $\mathbf{E}\big[f(Y)\big] = O(\alpha) \cdot \mathbf{E}\big[f(W)\big]$ (Theorem 14). We believe that our structural result for the expectation of a monotone, symmetric norm is of independent interest, and should also find application in other stochastic settings involving monotone, symmetric norms.

Our structural result showing that $\mathbf{E}\big[f(Y)\big] = O(1) \cdot f\big(\mathbf{E}\big[Y^{\downarrow}\big]\big)$ is obtained by carefully exploiting the structure of monotone, symmetric norms. A key component of this is identifying suitable *statistics* of the random vector Y, indexed by $\ell \in [m]$ such that: (a) the statistics for index ℓ lead to a convenient proxy function for estimating $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ within $O(1)$ factors; (b) the statistics are related to the expectations of some random variables that are tightly concentrated around their means; and (c) $\Pr\big[f(Y) > \sigma \cdot f\big(\mathbf{E}\big[Y^{\downarrow}\big]\big)\big]$ is governed by the probability that these random variables deviate from their means. Together these properties imply the desired bound on $\mathbf{E}\big[f(Y)\big]$. Next, we elaborate on these statistics.

Proxy functions and statistics (Section IV). Since f is a symmetric function, it is not hard to see that $f(Y)$ depends only on the *"histogram"* $\big\{N^{>\theta}(Y)\big\}_{\theta\in\mathbb{R}_{\geq 0}}$, where $N^{>\theta}(Y)$ is the number of coordinates of Y larger than θ. But this dependence is quite non-linear; its precise form is given by the structural result for monotone, symmetric norms, and by noting that we can write $\mathsf{Top}_\ell(Y) = \int_0^\infty \min\big\{\ell, N^{>\theta}(Y)\big\} d\theta$. Despite these non-linearities, we show that the expected histogram curve $\big\{\mathbf{E}\big[N^{>\theta}(Y)\big]\big\}_{\theta\in\mathbb{R}_{\geq 0}}$ (see Fig. 1 in Section IV) controls $\mathbf{E}\big[f(Y)\big]$.

To show this, and also compress $\big\{\mathbf{E}\big[N^{>\theta}(Y)\big]\big\}_{\theta\in\mathbb{R}_{\geq 0}}$ to a finite, manageable number of statistics, we consider first the Top_ℓ-norm. While $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] = \int_0^\infty \mathbf{E}\big[\min\{\ell, N^{>\theta}(Y)\}\big]d\theta$, interestingly, we prove that *this is only a constant-factor smaller than* $\gamma_\ell(Y) := \int_0^\infty \min\{\ell, \mathbf{E}\big[N^{>\theta}(Y)\big]\}d\theta$ (Theorem 10); i.e., interchanging E and min only leads to an $O(1)$-factor loss. Defining $\tau_\ell(Y)$ to be the smallest θ such that $\mathbf{E}\big[N^{>\theta}(Y)\big] < \ell$, which can be viewed as an estimate of the ℓ-th largest entry of Y, we can more compactly write $\gamma_\ell(Y) = \ell\tau_\ell(Y) + \int_{\tau_\ell(Y)}^\infty \mathbf{E}\big[N^{>\theta}(Y)\big]d\theta$. The statistics of interest to us are the quantities $\tau_\ell = \tau_\ell(Y)$ (and implicitly $\mathbf{E}\big[N^{>\tau_\ell}(Y)\big]$) for $\ell = 1, \ldots, m$. They enable us to bound the tail probability $\Pr[f(Y) > \sigma \cdot f(\mathbf{E}\big[Y^{\downarrow}\big])]$ by $\exp(-\Omega(\sigma))$ which implies that $\mathbf{E}\big[f(Y)\big] = O(1) \cdot f(\mathbf{E}\big[Y^{\downarrow}\big])$. Roughly speaking, this follows because we show (by exploiting the proxy function $\gamma_\ell(Y)$) that $\Pr[f(Y) > \sigma \cdot f(\mathbf{E}\big[Y^{\downarrow}\big])]$ is at most the probability that $N^{>\tau_\ell}(Y) > \Omega(\sigma) \cdot \ell$ for some $\ell \in [m]$, and Chernoff bounds show that $N^{>\tau_\ell}(Y) = \sum_{i=1}^m \Pr[Y_i > \tau_\ell]$ is tightly concentrated around $\mathbf{E}\big[N^{>\tau_\ell}(Y)\big] < \ell$.

Approximation algorithms for stochastic minimum-norm optimization. In Sections VI and VII, we apply our framework to design the first approximation algorithms for the stochastic minimum-norm versions of load balancing, and spanning tree (and matroid basis) problems respectively. These sections can be read independently of each other.

Let $\mathcal{O}^*$ denote the random cost vector resulting from an optimal solution. Applying our framework entails bounding $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ in terms of $\mathbf{E}\big[\mathsf{Top}_\ell(\mathcal{O}^*)\big]$ for all $\ell \in [m]$. For algorithmic tractability, we only work with indices $\ell \in \mathsf{POS} = \mathsf{POS}_m := \{2^i : i = 0, 1, \ldots, \lfloor \log_2 m \rfloor\}$: since $\mathsf{Top}_\ell(\vec{v}) \leq \mathsf{Top}_{\ell'}(\vec{v}) \leq 2\mathsf{Top}_\ell(\vec{v})$ holds for any $\vec{v} \in \mathbb{R}^m_{\geq 0}$ and any $\ell \leq \ell' \leq 2\ell$, it is easy to see that with a factor-2 loss, this still yields a bound on $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]/\mathbf{E}\big[\mathsf{Top}_\ell(\mathcal{O}^*)\big]$ for *all* $\ell \in [m]$. At a high level, we "guess" within, say a factor of 2, $\mathbf{E}\big[\mathsf{Top}_\ell(\mathcal{O}^*)\big]$ or certain associated quantities such as $\tau_\ell(\mathcal{O}^*)$, for all $\ell \in \mathsf{POS}$. This guessing takes polynomial time since it involves enumerating a vector with $O(\log m)$ *monotone* (i.e., non-increasing or non-decreasing) coordinates, each of which lies in a logarithmically bounded range. We then write an LP to obtain a fractional solution whose associated cost-vector $\overline{Y}$ roughly speaking satisfies $\mathbf{E}\big[\mathsf{Top}_\ell(\overline{Y})\big] \leq O\big(\mathbf{E}\big[\mathsf{Top}_\ell(\mathcal{O}^*)\big]\big)$ for all $\ell \in \mathsf{POS}$. The chief technical ingredient is to devise a rounding procedure to obtain an integer solution while preserving the $\mathbf{E}\big[\mathsf{Top}_\ell(.)\big]$-costs (up to some factor) for all $\ell \in \mathsf{POS}$. We exploit *iterative rounding* to achieve this, by capitalizing on the fact that (loosely speaking) our matrix of tight constraints has bounded column sums (see Theorem 5).

For spanning trees (Section VII)—edge costs are random and we seek to minimize the expected norm of the edge-cost vector of the spanning tree—the implementation of the above plan is quite direct. We guess $\tau_\ell^* := \tau_\ell(\mathcal{O}^*)$ for all $\ell \in \mathsf{POS}$, and our LP imposes the constraints $\mathbf{E}\big[N\big(\overline{Y}^{>\tau_\ell^*}\big)\big] \leq \ell$ for all $\ell \in \mathsf{POS}$. Since the coordinates of $\overline{Y}$ correspond to individual edge costs, and we know their distributions, it is easy to impose the above constraint (see $(\mathsf{Tree}(\vec{t}))$). Iterative rounding works out quite nicely here since after normalizing the above constraints to obtain unit right-hand-sides, each column sum is $O(1)$. Thus, we obtain a *constant-factor approximation*; also, everything extends to the setting of general matroids.

Our results for load balancing (Section VI) are the most technically sophisticated results in the paper. We obtain *constant-factor approximations* for: (i) *arbitrary* monotone, symmetric norms with Bernoulli job processing times; and (ii) Top_ℓ-norm with *arbitrary* distributions. We also obtain an $O(\log m/\log\log m)$-approximation for the most general setting, where both the norm and the job-processing-time distributions may be arbitrary.

The cost-vector Y in load balancing corresponds to the (random) loads on the machines. Each component Y_i is thus an *aggregate* of some random variables: $Y_i = \sum_{j \text{ assigned to } i} X_{ij}$, where X_{ij} is the (random) processing time of job j on machine i. The complication that this creates is that we do not have direct access to (the distribution for) the random variable Y_i, making it difficult to calculate (or estimate) $\Pr[Y_i > \theta]$ (and hence $\mathbf{E}\big[N(Y^{>\theta})\big]$). in fact (This is #P-hard, even for Bernoulli X_{ij}s [15], but if we know the jobs assigned to i, then we can use a dynamic program can to obtain a $(1+\varepsilon)$-approximation to $\Pr[Y_i > \theta]$.)

We circumvent these difficulties by leveraging some tools from the work of [10], [15] on stochastic makespan minimization, in conjunction with an *alternate* proxy that we develop for $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$. We show that, for a suitable choice of θ, $\sum_{i=1}^m \mathbf{E}\big[Y_i^{\geq\theta}\big]$ is a good proxy for $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ (see Theorem 9), where $Y_i^{\geq\theta}$ is the random variable that is 0 if the event $\{Y_i < \theta\}$ happens, and Y_i otherwise. Complementing this, the insight we glean from [10], [15] is that we can estimate $\mathbf{E}\big[Y_i^{\geq\theta}\big]$ within $O(1)$ factors using quantities that can be obtained from the distributions of the X_{ij} random variables (see Section III-A). This involves analyzing the contribution of job j (to $\mathbf{E}\big[Y_i^{\geq\theta}\big]$) differently based on whether X_{ij} is "small" (truncated) or "large" (exceptional), and utilizing the notion of *effective size* of a random variable [13], [15] to bound the contribution from small jobs. We guess the θ values—call them t_ℓ^*—corresponding to $\mathbf{E}\big[\mathsf{Top}_\ell(\mathcal{O}^*)\big]$ for all $\ell \in \mathsf{POS}$, and write an LP for finding a fractional assignment where we enforce constraints encoding that these t_ℓ^* values are compatible with the correct guesses.

III. Preliminaries

For an integer $m \geq 1$, we use $[m]$ to denote the set $\{1, \ldots, m\}$. For any integer $m \geq 1$, we define $\mathsf{POS}_m = \{2^i : i \in \mathbb{Z}_{\geq 0}, 2^i \leq m\}$; we drop the subscript m when it is clear from the context. For $x \in \mathbb{R}$, define $x^+ := \max\{x, 0\}$. For an event A we use the indicator random variable $\mathbb{1}_A$ to denote if event A happens. For any vector $x \in \mathbb{R}^m_{\geq 0}$ we use $x^{\downarrow}$ to denote the vector x with its coordinates sorted in non-increasing order.

Throughout, we use symbols Y and W to denote *random vectors*. The coordinates of these random vectors are always *independent*, nonnegative random variables. We denote this by saying that the random vector follows a product distribution. We reserve Z to denote a scalar nonnegative random variable. Given Z and $\theta \in \mathbb{R}_{\geq 0}$, define the *truncated* random variable $Z^{<\theta} := Z \cdot \mathbb{1}_{Z<\theta}$, which has support in $[0, \theta)$. Analogously, define the *exceptional* random variable $Z^{\geq\theta} := Z \cdot \mathbb{1}_{Z\geq\theta}$, whose support lies in $\{0\} \cup [\theta, \infty)$. The following form of Chernoff bounds will be useful.

Lemma 1 (Chernoff bound)**.** *Let $Z_1, \ldots, Z_k$ be independent $[0, 1]$ random variables, and $\mu \geq \sum_{j\in[k]} \mathbf{E}\left[Z_j\right]$. For any $\varepsilon > 0$, we have $\mathbf{Pr}\left[\sum_{j\in[k]} Z_j > (1+\varepsilon)\mu\right] \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu}$. If $\varepsilon > 1$, then we also have the simpler bound $\mathbf{Pr}\left[\sum_{j\in[k]} Z_j > (1+\varepsilon)\mu\right] \leq e^{-\varepsilon\mu/3}$.*

Monotone, symmetric norms: A function $f : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is a *norm* if it satisfies: (a) $f(x) = 0$ iff $x = 0$; (b) (homogeneity) $f(\lambda x) = |\lambda| f(x)$, for all $\lambda \in \mathbb{R}$, $x \in \mathbb{R}^m$; and (c) (triangle inequality) $f(x + y) \leq f(x) + f(y)$ for all $x, y \in \mathbb{R}^m$. Since our cost vectors are always nonnegative, we only consider nonnegative vectors in the sequel. A *monotone, symmetric* norm f is a norm that satisfies: $f(x) \leq f(y)$ for all $0 \leq x \leq y$ (monotonicity); and $f(x) = f(x^{\downarrow})$ for all $x \in \mathbb{R}^m_{\geq 0}$ (symmetry). In the sequel, whenever we say norm, we always mean a monotone, symmetric norm. We will often assume that f is normalized, i.e., $f(1, 0, \ldots, 0) = 1$.

The following two types of monotone, symmetric norms will be especially important to us.

Definition 2. For any $\ell \in [m]$, the Top_ℓ *norm* is defined as follows: for $x \in \mathbb{R}^m_{\geq 0}$, $\mathsf{Top}_\ell(x)$ is the sum of the ℓ largest coordinates of x, i.e., $\mathsf{Top}_\ell(x) = \sum_{i=1}^{\ell} x_i^{\downarrow}$.

More generally, for any $w \in \mathbb{R}^m_{\geq 0}$ satisfying $w_1 \geq w_2 \geq \cdots \geq w_m \geq 0$—we call such a w a non-increasing vector—the *w-ordered norm* (or simply ordered norm) of a vector $x \in \mathbb{R}^m_{\geq 0}$ is defined as $\|x\|_w := w^T x^{\downarrow}$. Observe that $\|x\|_w = \sum_{\ell=1}^{m} (w_\ell - w_{\ell+1}) \mathsf{Top}_\ell(x)$, where $w_\ell := 0$ for $\ell > m$.

Top_ℓ-norm minimization yields a natural way of interpolating between the min-max (Top_1) and min-sum (Top_m) problems. The following result by [7], which gives a foothold for working with an arbitrary monotone, symmetric norm, further highlights their importance.

Theorem 3 (Structural result for monotone, symmetric norms [7])**.** *Let $f : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ be a monotone, symmetric norm.*

(a) There is a collection $\mathcal{C} \subseteq \mathbb{R}^m_{\geq 0}$ of non-increasing vectors such that $f(x) = \sup_{w\in\mathcal{C}} w^T x^{\downarrow}$ for all $x \in \mathbb{R}^m_{\geq 0}$. If f is normalized, we have $\sup_{w\in\mathcal{C}} w_1 = 1$.

(b) Hence, if $x, y \in \mathbb{R}^m_{\geq 0}$ are such that $\mathsf{Top}_\ell(x) \leq \alpha \mathsf{Top}_\ell(y) + \beta$ for all $\ell \in [m]$, where $\alpha \geq 0$, then $f(x) \leq \alpha f(y) + \beta f(1, 0, \ldots, 0)$.

We will often need to enumerate vectors with monotone integer coordinates. We use the following standard result, lifted from [7], to obtain a polynomial bound on the number of vectors that need to be enumerated.

Claim 4. *There are at most $(2e)^{\max\{M,k\}}$ non-increasing sequences of k integers chosen from $\{0, \ldots, M\}$.*

Iterative rounding: Our algorithms are based on rounding fractional solutions to LP-relaxations that we formulate for the stochastic min-norm versions of load balancing and spanning trees. The rounding algorithm needs to ensure that the various budget constraints that we include in our LP to control quantities associated with expected Top_ℓ norms (for multiple indices ℓ) are roughly preserved. The main technical tool involved in achieving this is *iterative rounding*, as expressed by the following theorem, which follows from a result in [21].

Theorem 5 (Follows from Corollary 11 in [21])**.** *Let $\mathcal{M} = (\mathcal{U}, \mathcal{I})$ be a matroid with rank function r, and $\mathcal{Q} := \{z \in \mathbb{R}^{\mathcal{U}}_{\geq 0} : z(\mathcal{U}) = \mathsf{r}(\mathcal{U}),\ \ z(F) \leq \mathsf{r}(F)\ \forall F \subseteq \mathcal{U}\}$ be its base polytope. Let $\overline{z}$ be a feasible solution to the following multi-budgeted matroid LP:*

$$\min \quad c^T z \qquad \text{s.t.} \qquad Az \leq b, \quad z \in \mathcal{Q}$$

where $A \in \mathbb{R}^{k\times\mathcal{U}}_{\geq 0}$. Suppose that for any $e \in \mathrm{supp}(\overline{z})$, we have $\sum_{i\in[k]} A_{i,e} \leq \nu$ for some parameter ν. We can round $\overline{z}$ in polynomial time to obtain a basis B of $\mathcal{M}$ satisfying:
(a) $c(B) \leq c^T\overline{z}$; (b) $A\chi^B \leq b + \nu\mathbf{1}$, where $\mathbf{1}$ is the vector of all 1s; and (c) B is contained in the support of $\overline{z}$.

A. Bounding $\mathbf{E}\left[\boldsymbol{S}^{\geq\boldsymbol{\theta}}\right]$ when $\boldsymbol{S}$ is a sum of independent random variables

In stochastic load balancing, each coordinate of the random cost vector is a "composite" random variable that is the sum of independent random variables. In such settings, where we do not have direct access to the distributions of individual coordinates, we develop a proxy function for estimating $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ that involves computing $\mathbf{E}\left[Y_i^{\geq\theta}\right]$ for $i \in [m]$. We discuss here to compute $\mathbf{E}\left[S^{\geq\theta}\right]$ for a composite random variable $S = \sum_{j\in[k]} Z_j$, where $Z_1, \ldots, Z_k \geq 0$ are independent random variables. By gleaning suitable insights from [10], [15], we show how to estimate this quantity given access to the distributions of the Z_j random variables. First, observe that for any $j \in [k]$ if the event $\{Z_j \geq \theta\}$ happens, then the event $\{S \geq \theta\}$ happens. So after some simple manipulations, we can show that

$$\mathbf{E}\left[S^{\geq\theta}\right] = \Theta\bigg(\sum_{j\in[k]} \mathbf{E}\left[Z_j^{\geq\theta}\right] + \mathrm{E}\Big[\big(\sum_{j\in[k]} Z_j^{<\theta}\big)^{\geq\theta}\Big]\bigg).$$

The first term above is easily computed from the Z_j-distributions. So to get a handle on $\mathbf{E}\left[S^{\geq\theta}\right]$ it suffices to control the second term, which measures the contribution from the sum of the *truncated* random variables $Z_j^{<\theta}$ to $\mathbf{E}\left[S^{\geq\theta}\right]$. This requires a nuanced notion called *effective size*, a concept that originated in queuing theory [13].

Definition 6 (Effective size)**.** For a nonnegative random variable Z and a parameter $\lambda > 1$, the λ-effective size $\beta_\lambda(Z)$ of Z is defined as $\log_\lambda \mathbf{E}\left[\lambda^Z\right]$. Also, define $\beta_1(Z) := \mathbf{E}\left[Z\right]$.

The usefulness of effective sizes follows from Lemmas 7 and 8, which indicate that $\mathbf{E}\left[\left(\sum_j Z_j^{<\theta}\right)^{\geq\theta}\right]$ can be estimated by controlling the effective sizes of some random variables related to the $Z_j^{<\theta}$ random variables.

Lemma 7. *Let Z be a nonnegative random variable and $\lambda \geq 1$. If $\beta_\lambda(Z) \leq b$, then $\mathbf{Pr}\left[Z \geq b + c\right] \leq \lambda^{-c}$, for any $c \geq 0$. Furthermore, if $\lambda \geq 2$, then $\mathbf{E}\left[Z^{\geq \beta_\lambda(Z)+1}\right] \leq \frac{\beta_\lambda(Z)+3}{\lambda}$.*

If Z and Z' are independent, then $\beta_\lambda(Z + Z') = \beta_\lambda(Z) + \beta_\lambda(Z')$. So by Lemma 7, if $\sum_{j\in[k]} \beta_\lambda(Z_j^{<\theta}/\theta) = O(1)$, then $\mathbf{E}\left[\left(\sum_{j\in[k]} Z_j^{<\theta}/\theta\right)^{>\Omega(1)}\right] = O(1)/\lambda$, or equivalently $\mathbf{E}\left[\left(\sum_{j\in[k]} Z_j^{<\theta}\right)^{>\Omega(\theta)}\right]$ is $O(\theta)/\lambda$.

A key contribution of Kleinberg et al. [15] is encapsulated by Lemma 8 below, which is obtained by combining various results from their work, and complements the above upper bound. It lower bounds $\mathbf{E}\left[S^{\geq\theta}\right]$, where S is the sum of independent, *bounded* random variables (e.g., truncated random variables), in terms of the sum of the λ-effective sizes of these random variables.

Lemma 8. *Let $S = \sum_{j\in[k]} Z_j$ be a sum of independent $[0, \theta]$-bounded random variables. Let $\lambda \geq 1$ be an integer. Then, $\mathbf{E}\left[S^{\geq\theta}\right] \geq \theta \cdot \left(\sum_{j\in[k]} \beta_\lambda\left(Z_j/4\theta\right) - 6\right)/4\lambda$.*

IV. Proxy functions and statistics for controlling $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ and $\mathbf{E}\left[f(Y)\right]$

In this section, we identify various statistics of a random vector Y following a product distribution on $\mathbb{R}^m_{\geq 0}$, which lead to two proxy functions that provide a convenient handle on $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ within $O(1)$ factors.

The first proxy function uses $\sum_{i\in[m]} \mathbf{E}\left[Y_i^{\geq\theta}\right]$ as a means to control $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$. Roughly speaking, if θ is such that $\sum_{i\in[m]} \mathbf{E}\left[Y_i^{\geq\theta}\right] = \ell\theta$, then we argue that $\ell\theta$ is a good proxy for $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$; Theorem 9 makes this statement precise. This proxy is helpful in settings where Y_i is a sum of independent random variables, such as stochastic min-norm load balancing (Section VI), wherein Y_i denotes the load on machine i. The second proxy function is based on a statistic that we denote $\tau_\ell(Y)$, which aims to capture the (expected) ℓ-th largest entry of Y. This statistic is defined using the *expected histogram curve* $\{\mathbf{E}\left[N^{>\theta}(Y)\right]\}_{\theta\in\mathbb{R}_{\geq 0}}$ (see Fig. 1), where $N^{>\theta}(Y)$ is the number of coordinates of Y that are larger than θ, and we show that this leads to an effective proxy $\gamma_\ell(Y)$ for $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ (Theorem 10). Collectively, these statistics and the $\gamma_\ell(Y)$ proxies (for all $\ell \in [m]$) are instrumental in showing that the expected histogram curve controls $\mathbf{E}\left[f(Y)\right]$, and hence that $\mathbf{E}\left[f(Y)\right]$ can be bounded in terms of the expected Top_ℓ norms of Y (see Section V). Also, these statistics are quite convenient to work with in designing algorithms for problems where the Y_is are "atomic" random variables with known distributions, such as stochastic min-norm spanning trees (see Section VII).

Theorem 9 (Proxy based on $\mathbf{E}\left[Y_i^{\geq\theta}\right]$ statistics). *Let $\theta \geq 0$.*
(a) If $\sum_{i\in[m]} \mathbf{E}\left[Y_i^{\geq\theta}\right] \leq \ell\theta$ holds, then $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right] \leq 2\ell\theta$.
(b) If $\sum_{i\in[m]} \mathbf{E}\left[Y_i^{\geq\theta}\right] > \ell\theta$ holds, then $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right] > \frac{\ell\theta}{2}$.

The proof of part (a) follows by writing each Y_i as $Y_i^{<\theta} + Y_i^{\geq\theta}$, and noting that $\mathbf{E}\left[\mathsf{Top}_\ell(Y_1^{<\theta}, \ldots, Y_m^{<\theta})\right] \leq \ell\theta$ and $\mathbf{E}\left[\mathsf{Top}_\ell(Y_1^{\geq\theta}, \ldots, Y_m^{\geq\theta})\right] \leq \sum_{i\in[m]} \mathbf{E}\left[Y_i^{\geq\theta}\right]$. The proof of part (b) follows by induction on $\ell + m$. We condition on the event $Y_m \geq \theta$. If $Y_m \geq \theta$, we lower bound $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ by $\mathbf{E}\left[Y_m | Y_m \geq \theta\right] + \mathbf{E}\left[\mathsf{Top}_{\ell-1}(Y_1, \ldots, Y_{m-1})\right]$, and otherwise, we lower bound $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$ by $\mathbf{E}\left[\mathsf{Top}_\ell(Y_1, \ldots, Y_{m-1})\right]$; in both cases, we use the induction hypothesis to lower bound the $\mathbf{E}\left[\mathsf{Top}_\ell(.)\right]$ or $\mathbf{E}\left[\mathsf{Top}_{\ell-1}(.)\right]$ terms.

Proxy function modeling the ℓth largest coordinate. Our second proxy function is derived by the simple, but quite useful, observation that for any $x \in \mathbb{R}^m_{\geq 0}$, we have $\mathsf{Top}_\ell(x) = \int_0^\infty \min\{\ell, N^{>\theta}(x)\} d\theta$, where $N^{>\theta}(x)$ is the number of coordinates of x that are greater than θ. Noting that $\mathbf{E}\left[\min\{\ell, N^{>\theta}(Y)\}\right] \leq \min\{\ell, \mathbf{E}\left[N^{>\theta}(Y)\right]\}$, we therefore obtain that $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right] \leq \gamma_\ell(Y) := \int_0^\infty \min\{\ell, \mathbf{E}\left[N^{>\theta}(Y)\right]\} d\theta$ (see Fig. 1). Interestingly, we show that $\gamma_\ell(Y) = O(1) \cdot \mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$.

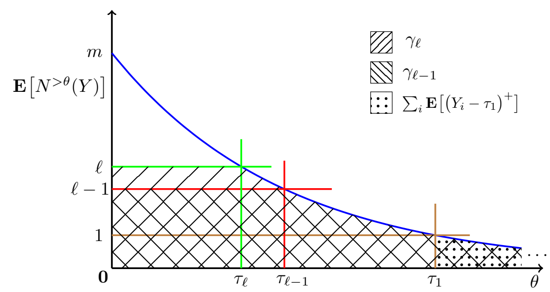

Fig. 1. The expected histogram curve $\{\mathbf{E}\left[N^{>\theta}(Y)\right]\}_{\theta\in\mathbb{R}_{\geq 0}}$.

Theorem 10. *For any $\ell \in [m]$, we have $\mathbf{E}\left[\mathsf{Top}_\ell(Y)\right] \leq \gamma_\ell(Y) \leq 4 \cdot \mathbf{E}\left[\mathsf{Top}_\ell(Y)\right]$, where $\gamma_\ell(Y) := \int_0^\infty \min\{\ell, \mathbf{E}\left[N^{>\theta}(Y)\right]\} d\theta$.*

The proof of the second inequality above relies on a rephrasing of $\gamma_\ell(Y)$ that makes it amenable to relate it to Theorem 9(b). For any $\ell \in [m]$, define $\tau_\ell(Y) := \inf\{\theta \in \mathbb{R}_{\geq 0} : \mathbf{E}\left[N^{>\theta}(Y)\right] < \ell\}$. This infimum is attained because $\mathbf{Pr}\left[Y_i \leq \theta\right]$ is a right-continuous function of θ, and $\mathbf{E}\left[N^{>\theta}(Y)\right] = m - \sum_{i\in[m]} \mathbf{Pr}\left[Y_i \leq \theta\right]$. Also, define $\tau_0(Y) := \infty$ and $\tau_\ell(Y) := 0$ for $\ell > m$. Since $\mathbf{E}\left[N^{>\theta}(Y)\right]$ is a non-increasing function of θ, we then have $\gamma_\ell(Y) = \ell\tau_\ell(Y) + \int_{\tau_\ell(Y)}^\infty \mathbf{E}\left[N^{>\theta}\right] d\theta$. We can further rewrite this in a way that quite nicely brings out the similarities between $\gamma_\ell(Y)$ and the (exact) proxy function for $\mathsf{Top}_\ell(x)$ used by [7] in the deterministic setting. Claim 11 gives a convenient way of casting the integral in the second term, which leads to expression for $\gamma_\ell(Y)$ stated in Lemma 12.

Claim 11. *For any $t \geq 0$, we have $\int_t^\infty \mathbf{E}\left[N^{>\theta}(Y)\right] d\theta = \sum_{i\in[m]} \mathbf{E}\left[(Y_i - t)^+\right]$.*

Lemma 12. *Consider any $\ell \in [m]$. We have $\gamma_\ell(Y) = \ell\tau_\ell(Y) + \sum_{i\in[m]} \mathbf{E}\left[(Y_i - \tau_\ell(Y))^+\right]$ and $\gamma_\ell(Y) = \min_{t\in\mathbb{R}_{\geq 0}}\left(\ell t + \sum_{i\in[m]} \mathbf{E}\left[(Y_i - t)^+\right]\right)$.*

Proof of Theorem 10: Let $\rho := \inf\{\theta : \sum_{i\in[m]} \mathbf{E}\big[Y_i^{\geq\theta}\big] \leq \ell\theta\}$. Note that $\sum_{i\in[m]} \mathbf{E}\big[Y_i^{>\rho}\big] \leq \ell\rho$. By Lemma 12, we have $\gamma_\ell(Y) \leq \ell\rho + \sum_{i\in[m]} \mathbf{E}\big[(Y_i-\rho)^+\big] \leq \ell\rho + \sum_{i\in[m]} \mathbf{E}\big[Y_i^{>\rho}\big] \leq 2\ell\rho$. For any $\theta < \rho$, Theorem 9(b) implies that $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] > \ell\theta/2$. Thus, $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] \geq \ell\rho/2$. ■

V. Expectation of a monotone, symmetric norm

Let Y follow a product distribution on $\mathbb{R}^m_{\geq 0}$, and $f : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ be an arbitrary monotone, symmetric norm. By Theorem 3 (a), there is a collection $\mathcal{C} \subseteq \mathbb{R}^m_{\geq 0}$ of weight vectors with non-increasing coordinates such that $f(x) = \sup_{w\in\mathcal{C}} w^T x^{\downarrow}$ for all $x \in \mathbb{R}^m_{\geq 0}$. For $w \in \mathcal{C}$, recall that we define $w_\ell := 0$ whenever $\ell > m$. We now prove one of our main technical results: *the expectation of a supremum of ordered norms is within a constant factor of the supremum of the expectation of ordered norms*. Formally, it is clear that $\mathbf{E}\big[f(Y)\big] = \mathbf{E}\big[\sup_{w\in\mathcal{C}} w^T Y^{\downarrow}\big] \geq \sup_{w\in\mathcal{C}} \mathbf{E}\big[w^T Y^{\downarrow}\big] = \sup_{w\in\mathcal{C}} w^T \mathbf{E}\big[Y^{\downarrow}\big] = f(\mathbf{E}\big[Y^{\downarrow}\big])$; we show that an inequality in the opposite direction also holds.

Theorem 13. *Let Y follow a product distribution on $\mathbb{R}^m_{\geq 0}$, and $f : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ be a monotone, symmetric norm. Then $f(\mathbf{E}\big[Y^{\downarrow}\big]) \leq \mathbf{E}\big[f(Y)\big] \leq 28 \cdot f(\mathbf{E}\big[Y^{\downarrow}\big])$.*

Theorem 13 is the backbone of our framework for stochastic minimum-norm optimization. Since $f(\mathbf{E}\big[Y^{\downarrow}\big]) = \sup_{w\in\mathcal{C}} \sum_{\ell\in[m]} (w_\ell - w_{\ell+1})\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$, Theorem 13 gives us a concrete way of bounding $\mathbf{E}\big[f(Y)\big]$, namely, by bounding all expected Top_ℓ norms. In particular, we obtain the following *corollary*, that we call *approximate stochastic majorization*. Recall that $\mathsf{POS}_m = \{2^i : i \in \mathbb{Z}_{\geq 0}, 2^i \leq m\}$.

Theorem 14 (**Approximate stochastic majorization**). *Let Y and W follow product distributions on $\mathbb{R}^m_{\geq 0}$. Let f be a monotone, symmetric norm on $\mathbb{R}^m$.*

(a) If $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] \leq \alpha \cdot \mathbf{E}\big[\mathsf{Top}_\ell(W)\big]$ for all $\ell \in [m]$, then $\mathbf{E}\big[f(Y)\big] \leq 28\alpha \cdot \mathbf{E}\big[f(W)\big]$.

(b) If $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] \leq \alpha \cdot \mathbf{E}\big[\mathsf{Top}_\ell(W)\big]$ for all $\ell \in \mathsf{POS}_m$, then $\mathbf{E}\big[f(Y)\big] \leq 2 \cdot 28\alpha \cdot \mathbf{E}\big[f(W)\big]$.

Theorem 14 (b) has immediate applicability for a given stochastic minimum-norm optimization problem such as stochastic min-norm load balancing. In our algorithms, we enumerate (say in powers of 2) all possible sequences $\{B_\ell\}_{\ell\in\mathsf{POS}_m}$ of estimates of $\big\{\mathbf{E}\big[\mathsf{Top}_\ell(\mathcal{O}^*)\big]\big\}_{\ell\in\mathsf{POS}_m}$, where $\mathcal{O}^*$ is the cost vector arising from an optimal solution, and find a solution (if one exists) whose cost vector satisfies (roughly speaking) these expected-Top_ℓ-norm estimates. A final step is to identify which of the solutions so obtained is a near-optimal solution. While a probabilistic guarantee follows easily since one can provide a randomized oracle for evaluating the objective $\mathbf{E}\big[f(Y)\big]$ (as $f(Y)$ enjoys good concentration properties), one can do better and *deterministically* identify a good solution. To this end, we show below (Lemma 15) that if $\{B_\ell\}_{\ell\in\mathsf{POS}_m}$ is a sequence that term-by-term well estimates $\big\{\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]\big\}_{\ell\in\mathsf{POS}_m}$, either from below or from above, then we can define a deterministic vector $\vec{b} \in \mathbb{R}^m_{\geq 0}$ such that $f(\vec{b})$ well-estimates $\mathbf{E}\big[f(Y)\big]$, from below or above respectively. We will apply parts (a) and (b) of Lemma 15 with the cost vectors arising from our solution and an optimal solution respectively, to argue the near-optimality of our solution. Lemma 15 utilizes the following definition. Let $K := 2^{\lfloor \log_2 m \rfloor}$ be the largest index in POS_m. Let $\{B_\ell\}_{\ell\in\mathsf{POS}_m}$ be a non-decreasing, nonnegative sequence. Define $B_0 := 0$, and $B_\ell = B_K$ for all $\ell > K$. The *upper envelope curve* $b : [0, m] \to \mathbb{R}_{\geq 0}$ of the sequence $\{B_\ell\}_{\ell\in\mathsf{POS}_m}$ is defined by $b(x) := \max\big\{y : (x, y) \in \mathrm{conv}(S)\big\}$, where $S = \{(\ell, B_\ell) : \ell \in \{0, m\} \cup \mathsf{POS}_m\}$ and $\mathrm{conv}(S)$ denotes the convex hull of S.

Lemma 15. *Let Y follow a product distribution on $\mathbb{R}^m_{\geq 0}$, and f be a monotone, symmetric norm on $\mathbb{R}^m$. Let $\{B_\ell\}_{\ell\in\mathsf{POS}_m}$ be a non-decreasing sequence such that $B_\ell \leq 2B_{\ell/2}$ for all $\ell \in \mathsf{POS}_m, \ell > 1$. Let $b : [0, m] \mapsto \mathbb{R}_{\geq 0}$ be the upper envelope curve of $\{B_\ell\}_{\ell\in\mathsf{POS}_m}$. Define $\vec{b} := \big(b(i) - b(i-1)\big)_{i\in[m]}$.*

(a) If $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] \leq \alpha B_\ell$ for all $\ell \in \mathsf{POS}_m$ (where $\alpha > 0$), then $\mathbf{E}\big[f(Y)\big] \leq 2 \cdot 28\alpha \cdot f(\vec{b})$.

(b) If $B_\ell \leq 2 \cdot \mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ for all $\ell \in \mathsf{POS}_m$, then $f(\vec{b}) \leq 6 \cdot \mathbf{E}\big[f(Y)\big]$.

Proof sketch of Theorem 13

We now discuss the proof of the second (and main) inequality of Theorem 13. Since Y and f are fixed throughout, we drop the dependence on these in most items of notation in the sequel. We may assume without loss of generality that f is normalized (i.e., $f(1, 0, \ldots, 0) = 1$) as scaling f to normalize it scales both $\mathbf{E}\big[f(Y)\big]$ and $f(\mathbf{E}\big[Y^{\downarrow}\big])$ by the same factor. It will be easier to work with the proxy function $\gamma_\ell(Y)$ for $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big]$ defined in Section IV. Recall that $\gamma_\ell = \ell\tau_\ell + \int_{\tau_\ell}^{\infty} \mathbf{E}\big[N^{>\theta}(Y)\big] d\theta$, where τ_ℓ is the smallest θ such that $\mathbf{E}\big[N^{>\theta}(Y)\big] < \ell$. Define $\tau_0 := \infty$ and $\gamma_0 := 0$ for notational convenience. Define $\mathsf{LB}' := \sup_{w\in\mathcal{C}} \sum_{\ell\in[m]} (w_\ell - w_{\ell+1})\gamma_\ell$. Given Theorem 10, it suffices to show that $\mathbf{E}\big[f(Y)\big] \leq 7 \cdot \mathsf{LB}'$.

The intuition and the roadmap of the proof are as follows. We have $f(Y) = \sup_{w\in\mathcal{C}} \sum_{\ell\in[m]} (w_\ell - w_{\ell+1})\mathsf{Top}_\ell(Y)$. Plugging in $\mathsf{Top}_\ell(Y) = \int_0^{\infty} \min\big\{\ell, N^{>\theta}(Y)\big\} d\theta \leq \ell\tau_\ell + \int_{\tau_\ell}^{\infty} N^{>\theta}(Y) d\theta$, we obtain that

$$f(Y) \leq \sup_{w\in\mathcal{C}} \Bigg[\sum_{\ell\in[m]} (w_\ell - w_{\ell+1})\Big(\ell\tau_\ell + \int_{\tau_\ell}^{\infty} N^{>\theta}(Y) d\theta\Big)\Bigg]. \tag{1}$$

Comparing (1) and LB' (after expanding out the γ_ℓ terms), syntactically, the only difference is that the $N^{>\theta}(Y)$ terms appearing in (1) are replaced by their expectations in LB'; however the dependence of $f(Y)$ on the $N^{>\theta}(Y)$ terms is quite non-linear, due to the sup operator. The chief insight that the $N^{>\theta}(Y)$ terms that really matter are those for $\theta = \tau_\ell$ for $\ell \in [m]$, and that the $N^{>\tau_\ell}(Y)$ quantities are *tightly concentrated* around their expectations. This allows us to, in essence, replace the $N^{>\theta}(Y)$ terms for $\theta = \tau_\ell, \ell \in [m]$ with their expectations (roughly speaking) when we consider

$\mathbf{E}\big[f(Y)\big]$, incurring a constant-factor loss, and thereby argue that $\mathbf{E}\big[f(Y)\big] = O(\mathsf{LB}')$.

To elaborate, since $N^{>\theta}(Y)$ is non-increasing in θ, we can upper bound each $\int_{\tau_\ell}^{\infty} N^{>\theta}(Y)d\theta$ expression in terms of $N^{>\tau_i}(Y)$, for $i = 2,\ldots,\ell$, and $\int_{\tau_1}^{\infty} N^{>\theta}(Y)d\theta$. Consequently, the RHS of (1) can be upper bounded in terms of the $N^{>\tau_\ell}(Y)$ quantities for $\ell = 2,\ldots,m$, and a term that depends on $\int_{\tau_1}^{\infty} N^{>\theta}(Y)d\theta = \sum_{i\in[m]}(Y_i - \tau_1)^+$. It is easy to charge the expectation of the latter term directly to LB'. We argue that the contribution from the $N^{>\tau_\ell}(Y)$ quantities for $\ell = 2,\ldots,m$, can be upper bounded by $f_{-1}(Y) := \sup_{w\in\mathcal{C}}\big(\sum_{\ell\in[m]}(w_\ell - w_{\ell+1})\ell\tau_\ell + \sum_{\ell=2}^{m} w_\ell(\tau_{\ell-1} - \tau_\ell)N^{>\tau_\ell}(Y)\big)$. We bound $\mathbf{E}\big[f_{-1}(Y)\big]$ (and hence $\mathbf{E}\big[f(Y)\big]$) by $O(\mathsf{LB}')$ by proving that the tail probability $\Pr[f_{-1}(Y) > \sigma\cdot\mathsf{LB}']$ decays exponentially with σ. The *crucial observation* here is that $\Pr[f_{-1}(Y) > \sigma\cdot\mathsf{LB}']$ is at most the probability that $N^{>\tau_\ell}(Y) > \Omega(\sigma)\cdot\ell$ for some index $\ell = 2,\ldots,m$. Each $N^{>\tau_\ell(Y)}$ is tightly concentrated around its expectation, which is at most ℓ, and so this probability is $O\big(e^{-\Omega(\sigma)}\big)$. ■

A. Controlling $\mathbf{E}\big[f(Y)\big]$ using the $\tau_\ell(Y)$ statistics

It is implicit from the proof of Theorem 13 that the τ_ℓ statistics control $\mathbf{E}\big[f(Y)\big]$, and we leverage this in settings where we have direct access to the distributions of the Y_i random variables. Corollary 16 makes this relationship explicit.

Corollary 16. *Let $\vec{\tau}(Y) := \big(\tau_\ell(Y)\big)_{\ell\in[m]}$. We have that $\mathbf{E}\big[f(Y)\big] = \Theta\big(\sum_{i\in[m]}\mathbf{E}\big[(Y_i - \tau_1(Y))^+\big] + f(\vec{\tau}(Y))\big)$.*

In our applications, we work with estimates $\{t_\ell\}_{\ell\in\mathsf{POS}_m}$ of the τ_ℓ^* values, for $\ell\in\mathsf{POS}_m$, of the cost vector arising from an optimal solution, and seek (and output) a solution whose cost vector Y minimizes $\sum_{i\in[m]}\mathbf{E}\big[(Y_i - t_1)^+\big]$ subject to the constraint that (roughly speaking) $\tau_\ell(Y) = O(t_\ell)$, for all $\ell\in\mathsf{POS}_m$. Two types of approximations arise here, one in estimating the τ_ℓ^*s and the other due to the process of finding Y. We state a more-robust version of Corollary 16 that incorporates such approximation factors, and is therefore particularly convenient to utilize in applications. We need the following definition: given a non-increasing sequence $\{v_\ell\}_{\ell\in\mathsf{POS}_m}$, we define its *expansion* to be the vector $v'\in\mathbb{R}^m$ given by $v'_i := v_{2^{\lfloor\log_2 i\rfloor}}$ for all $i\in[m]$.

Theorem 17. *Let Y follow a product distributions on $\mathbb{R}^m_{\geq 0}$. Let f be a normalized, monotone, symmetric norm on $\mathbb{R}^m$. Let $\{t_\ell\}_{\ell\in\mathsf{POS}_m}$ be a nonnegative, non-increasing sequence, and $\vec{t'}$ be its expansion.*

(a) Suppose that $\tau_{\beta\ell}(Y) \leq \alpha t_\ell$ for all $\ell\in\mathsf{POS}_m$, where $\alpha,\beta\geq 1$. Then, we have $\mathbf{E}\big[f(Y)\big] \leq 2\cdot 7\cdot(\alpha+2)\beta\cdot\big(\sum_{i\in[m]}\mathbf{E}\big[(Y_i - \alpha t_1)^+\big] + f(\vec{t'})\big)$.

(b) Suppose that $\tau_\ell(Y) \leq t_\ell \leq 2\tau_\ell(Y) + \kappa$ for all $\ell\in\mathsf{POS}_m$, where $\kappa\geq 0$. Then, we have $\sum_{i\in[m]}\mathbf{E}\big[(Y_i - t_1)^+\big] + f(\vec{t'}) \leq 32\cdot\mathbf{E}\big[f(Y)\big] + m\kappa$.

VI. Load Balancing

We now apply our framework to devise approximation algorithms for *stochastic minimum norm load balancing on unrelated machines*. We are given n *stochastic* jobs that need to be assigned to m unrelated machines. Throughout, we use $[n]$ and $[m]$ to denote the set of jobs and machines respectively; we use j to index jobs, and i to index machines. For each $i\in[m], j\in[n]$, we have a nonnegative r.v. X_{ij} that denotes the processing time of job j on machine i, whose distribution is specified in the input. Jobs are independent, so X_{ij} and $X_{i'j'}$ are independent whenever $j\neq j'$; however, X_{ij} and $X_{i'j}$ could be correlated. A feasible solution is an assignment $\sigma: [n]\to[m]$ of jobs to machines. This induces a random load vector load_σ where $\mathsf{load}_\sigma(i) := \sum_{j:\sigma(j)=i} X_{ij}$ for each $i\in[m]$; note that load_σ follows a product distribution on $\mathbb{R}^m_{\geq 0}$. The goal is to find an assignment σ that minimizes $\mathbf{E}\big[f(\mathsf{load}_\sigma)\big]$ for a given monotone symmetric norm f.

We often use $j\mapsto i$ as a shorthand for denoting $\sigma(j) = i$, when σ is clear from the context. We use σ^* to denote an optimal solution, and $\mathsf{OPT} := \mathbf{E}\big[f(\mathsf{load}_{\sigma^*})\big]$ to denote the optimal value. Let $\mathsf{POS} = \mathsf{POS}_m := \{1,2,4,\ldots,2^{\lfloor\log_2 m\rfloor}\}$.

Overview. Recall that Theorem 14 underlying our framework shows that in order to obtain an $O(\alpha)$-approximation for stochastic f-norm load balancing, it suffices to find an assignment σ that satisfies $\mathbf{E}\big[\mathsf{Top}_\ell(\mathsf{load}_\sigma)\big] \leq \alpha\,\mathbf{E}\big[\mathsf{Top}_\ell(\mathsf{load}_{\sigma^*})\big]$ for all $\ell\in\mathsf{POS}$. First, in Section VI-A, we consider the simpler problem where we only have one expected-Top_ℓ budget constraint, or equivalently, where f is a Top_ℓ-norm, and obtain an $O(1)$-approximation algorithm in this case.

Theorem 18. *There is a constant-factor approximation algorithm for stochastic Top_ℓ-norm load balancing on unrelated machines with arbitrary job size distributions.*

Section VI-A introduces many of the techniques that we build upon and refine in Section VI-B, where we deal with an arbitrary monotone, symmetric norm. We devise a constant-factor approximation when job sizes are *Bernoulli random variables*, and we give an $O(\log m/\log\log m)$-approximation for the most general setting, (i.e., arbitrary norm and arbitrary distributions). We remark that when working with an arbitrary norm f, our algorithms generate a polynomial number of candidate solutions. This collection of solutions is in fact *independent* of the norm f; the norm f is used only in the final step to select one of these solutions as the desired near-optimal solution.

Theorem 19. *There is a constant-factor approximation algorithm for stochastic f-norm load balancing on unrelated machines when job sizes are Bernoulli random variables.*

Theorem 20. *There is an $O(\log m/\log\log m)$-approximation algorithm for the general stochastic minimum-norm load balancing problem on unrelated machines, where the underlying monotone, symmetric norm and job-size distributions are arbitrary.*

A. Stochastic Top_ℓ-norm load balancing

We now prove Theorem 18. The key to our approach is Theorem 9, which implies that, for t chosen suitably, $\sum_i\mathbf{E}\big[\mathsf{load}_\sigma(i)^{\geq t}\big]$ acts as a convenient proxy for

$\mathbf{E}\big[\mathsf{Top}_\ell(\mathsf{load}_\sigma)\big]$. Theorem 9(b) shows that if $t \geq \frac{2\,\mathsf{OPT}}{\ell}$, then $\sum_i \mathbf{E}\big[\mathsf{load}_{\sigma*}(i)^{\geq t}\big] \leq \ell t$. We write a linear program, $(\mathsf{LP}(\ell,t))$, to find such an assignment (roughly speaking), and round its solution to obtain an assignment σ such that $\sum_i \mathbf{E}\big[\mathsf{load}_\sigma(i)^{\geq t}\big] = O(\ell t)$; by Theorem 9(a), this implies that $\mathbf{E}\big[\mathsf{Top}_\ell(\mathsf{load}_\sigma)\big] = O(\ell t)$. Hence, if we work with $t = O\big(\frac{\mathsf{OPT}}{\ell}\big)$ such that $(\mathsf{LP}(\ell,t))$ is feasible—which we can find via binary search—then we obtain an $O(1)$-approximation.

More precisely, for a given parameter $t \geq 0$, our LP seeks a fractional assignment satisfying $\sum_i \mathbf{E}\big[\mathsf{load}(i)^{\geq t}\big] = O(\ell t)$. As usual, we have z_{ij} variables indicating if job j is assigned to machine i, so z belongs to the assignment polytope $\mathcal{Q} := \{z \in \mathbb{R}^{m\times n}_{\geq 0}: \ \sum_{i\in[m]} z_{ij} = 1 \quad \forall i \in [m]\}$.

As alluded to in Section III-A, $\mathbf{E}\big[\mathsf{load}(i)^{\geq t}\big]$ can be controlled by separately handling the contribution from exceptional jobs $X_{ij}^{\geq t}$ and truncated jobs $X_{ij}^{<t}$. Our LP enforces that both these contributions (across all machines) are at most ℓt, thereby ensuring that $\sum_i \mathbf{E}\big[\mathsf{load}(i)^{\geq t}\big] = O(\ell t)$. Constraint (2) directly encodes that the total contribution from exceptional jobs is at most ℓt. To handle the contribution from truncated jobs, we utilize Lemma 8 here, which uses the notion of effective sizes. For each machine i, let $L_i := \sum_{j:j\mapsto i} X_{ij}^{<t}/t$ denote the scaled load on machine i due to the truncated jobs assigned to it. We use an auxiliary variable ξ_i to model $\mathbf{E}\big[L_i^{\geq 1}\big]$, so that $t\xi_i$ models $\mathbf{E}\big[(\sum_{j:j\mapsto i} X_{ij}^{<t})^{\geq t}\big]$. Since L_i is a sum of independent $[0,1]$-random variables, Lemma 8 yields various lower bounds on $\mathbf{E}\big[L_i^{\geq 1}\big]$; these are incorporated by constraints (3). Apriori, we do not know which is the *right* choice of λ in Lemma 8, so we simply include constraints for a sufficiently large collection of λ values so that one of them is close enough to the right choice. Finally, constraint (4) ensures that $\sum_i \mathbf{E}\big[(\sum_{j:j\mapsto i} X_{ij}^{<t})^{\geq t}\big] \leq \ell t$, thereby bounding the contribution from the truncated jobs. We obtain the following feasibility program $(\mathsf{LP}(\ell,t))$.

$$\sum_{i\in[m],j\in[n]} \mathbf{E}\big[X_{ij}^{\geq t}\big] z_{ij} \leq \ell t \tag{2}$$

$$\frac{\sum_{j\in[n]} \beta_\lambda\left(X_{ij}^{<t}/4t\right) z_{ij} - 6}{4\lambda} \leq \xi_i \quad \forall i \in [m], \forall \lambda \in [100m] \tag{3}$$

$$\sum_{i\in[m]} \xi_i \leq \ell \tag{4}$$

$$\xi \geq 0, \quad z \in \mathcal{Q}. \tag{5}$$

Claim 21. $(\mathsf{LP}(\ell,t))$ *is feasible for any* t *satisfying* $\mathbf{E}\big[\mathsf{Top}_\ell\left(\mathsf{load}_{\sigma^*}\right)\big] \leq \ell t/2$.

Rounding algorithm and analysis sketch. Suppose that $(\mathsf{LP}(\ell,t))$ is feasible, and $(\overline{z},\overline{\xi})$ is a feasible fractional solution. We set up an auxiliary LP, where we carefully choose, for each machine i, a suitable budget constraint from among the constraints (3). We round the fractional assignment $\overline{z}$ to obtain an assignment σ such that: (i) these budget constraints for the machines are satisfied approximately; and (ii) the total contribution from the exceptional jobs (across all machines) remains at most ℓt. The rounding step amounts to rounding a fractional solution to an instance of the *generalized assignment problem* (GAP), for which we can utilize the algorithm of [27], or use the iterative-rounding result from Theorem 5.

The budget constraint that we include for a machine is tailored to ensure that the total $\beta_\lambda\big(X_{ij}^{<t}/4t\big)$-effective load on a machine under the assignment σ is not too large; via Lemma 7, this will imply a suitable bound on $\mathbf{E}\big[L_i^{\geq\Omega(1)}\big]$, where $L_i = \sum_{j:j\mapsto i} X_{ij}^{<t}/4t$. Ideally, for each machine i we would like to choose constraint (3) for $\lambda_i = 1/\overline{\xi}_i$. This yields $\sum_j \beta_{\lambda_i}\big(X_{ij}^{<t}/4t\big)\overline{z}_{ij} \leq 4\lambda_i\overline{\xi}_i + 6 = 10$. So if this budget constraint is approximately satisfied in the rounded solution, say with RHS equal to some *constant* b, then Lemma 7 roughly gives us $\mathbf{E}\big[L_i^{\geq b+1}\big] \leq (b+3)/\lambda_i = (b+3)\overline{\xi}_i$. This in turn implies that

$$\sum_i \mathbf{E}\Big[\Big(\sum_{j:j\mapsto i} X_{ij}^{<t}\Big)^{\geq 4(b+1)t}\Big] = 4t\cdot\sum_i \mathbf{E}\big[L_i^{\geq b+1}\big] \leq 4t(b+3)\sum_i \overline{\xi}_i \leq 4(b+3)\ell t,$$

where the last inequality follows due to (4). The upshot is that $\sum_i \mathbf{E}\big[(\sum_{j:j\mapsto i} X_{ij}^{<t})^{\geq\Omega(t)}\big] = O(\ell t)$; coupled with the fact that $\sum_j \mathbf{E}\big[X_{\sigma(j),j}^{\geq t}\big] \leq \ell t$, we obtain that $\sum_i \mathbf{E}\big[\mathsf{load}_\sigma(i)^{\geq\Omega(t)}\big] = O(\ell t)$, and hence $\mathbf{E}\big[\mathsf{Top}_\ell(\mathsf{load}_\sigma)\big] = O(\ell t)$. The slight complication is that $1/\overline{\xi}_i$ need not be an integer in $[100m]$, so we modify the choice of λ_is appropriately to deal with this.

Finally, we can find $t = O\big(\frac{\mathsf{OPT}}{\ell}\big)$ such that $(\mathsf{LP}(\ell,t))$ is feasible via binary search, since, even for an arbitrary norm f, it is easy to obtain a (polytime computable) upper bound UB such that $\frac{\mathsf{UB}}{m} \leq \mathsf{OPT} \leq \mathsf{UB}$.

We remark that, whereas we work with a more general norm than ℓ_∞, our entire approach—polynomial-size LP-relaxation, rounding algorithm, and analysis—is in fact simpler and cleaner than the one used in [10] for the special case of Top_1 norm. Our savings can be traced to the fact that we leverage the notion of effective size in a more powerful way, by utilizing it at multiple scales to obtain lower bounds on $\mathbf{E}\big[\big(\sum_{j:j\mapsto i} X_{ij}^{<t}/t\big)^{\geq 1}\big]$ (Lemma 8).

B. Stochastic f-norm load balancing

We now focus on stochastic load balancing when f is a general monotone symmetric norm. Recall that σ^* denotes an optimal solution, and $\mathsf{OPT} = \mathsf{OPT}_f$ is the optimal value.

As noted earlier, Theorem 14 guides our strategy: we seek an assignment σ that simultaneously satisfies $\mathbf{E}\big[\mathsf{Top}_\ell\left(\mathsf{load}_\sigma\right)\big] = O\left(\alpha\,\mathbf{E}\big[\mathsf{Top}_\ell\left(\mathsf{load}_{\sigma^*}\right)\big]\right)$ for all $\ell \in \mathsf{POS}_m$ for some small factor α.

The LP-relaxation $(\mathsf{LP}(\ell,t))$ generalizes easily. Since we need to simultaneously work with all Top_ℓ norms, we now work with a guess t_ℓ of $2\,\mathbf{E}\big[\mathsf{Top}_\ell(\mathsf{load}_{\sigma^*})\big]/\ell$, for every $\ell \in \mathsf{POS}$. We have the usual z_{ij} variables encoding a fractional assignment. For each $\ell \in \mathsf{POS}$, there is a different definition of truncated r.v. $X_{ij}^{<t_\ell}$ and exceptional r.v. $X_{ij}^{\geq t_\ell}$. Correspondingly, for each index $\ell \in \mathsf{POS}$, we have a separate set of constraints (2)–(4) involving (the z_{ij}s), a variable $\xi_{i,\ell}$ (that represents ξ_i for the index ℓ, i.e., $\mathbf{E}\big[(\sum_{j:j\mapsto i} X_{ij}^{<t_\ell}/t_\ell)^{\geq 1}\big]$), and the guess t_ℓ. For technical reasons, we include additional constraints that enforce that a job j cannot be assigned to a machine i if $\mathbf{E}\big[X_{ij}^{\geq t_1}\big] > t_1$; observe that this is valid for the optimal integral solution whenever $t_1 \geq 2\,\mathbf{E}\big[\mathsf{Top}_1\left(\mathsf{load}_{\sigma^*}\right)\big]$.

Let $(\mathsf{LP}(\vec{t}))$ denote the resulting LP relaxation. Claim 21 easily generalizes to yield that if $\vec{t}$ is such that $t_\ell \geq 2\mathbf{E}\big[\mathsf{Top}_\ell\left(\mathsf{load}_{\sigma^*}\right)\big]/\ell$ for all $\ell \in \mathsf{POS}$, then, $(\mathsf{LP}(\vec{t}))$ is feasible.

Designing an LP-rounding algorithm is however substantially more challenging now. In the Top_ℓ-norm case, we set up an auxiliary LP by extracting a single budget constraint for each machine that served to bound the contribution from the truncated jobs on that machine. This LP is quite easy to round (e.g., using Theorem 5) because each z_{ij} variable participates in exactly one constraint (thereby trivially yielding an $O(1)$ bound on the column sum for each variable). Since we now have to simultaneously control multiple Top_ℓ-norms, for each machine i, we will now need to include a budget constraint *for every index* $\ell \in \mathsf{POS}$ so as to bound the contribution from the truncated jobs for index ℓ (i.e., $\mathbf{E}\big[\big(\sum_{j:j\mapsto i} X_{ij}^{<t_\ell}\big)^{\geq\Omega(t_\ell)}\big]$). Additionally, we will now need, for each $\ell \in \mathsf{POS}$, a separate constraint to bound the total contribution from the exceptional jobs for index ℓ. Consequently, every z_{ij} variable now participates in *multiple* budget constraints, which makes it difficult to argue a bounded-column-sum property for this variable and thereby leverage Theorem 5. Thus, while we can set up an auxiliary LP containing these various budget constraints, rounding a fractional solution to this LP so as to approximately satisfy these various budget constraints presents a significant technical hurdle.

We show how to partly overcome this obstacle in two cases. When job sizes are Bernoulli random variables, we show that an auxiliary LP as above yields a constraint matrix with $O(1)$-bounded column sums. Consequently, this auxiliary LP can be rounded with an $O(1)$ violation of all budget constraints (using Theorem 5), which then (with the right choice of $\vec{t}$) leads to an $O(1)$-approximation algorithm (Theorem 19). For the general case, we argue that we can set up an auxiliary LP that imposes a *weaker* form of budget constraints involving expected truncated job sizes, and does have $O(1)$ column sums. Via a suitable use of Chernoff bounds, this then leads to an $O(\log m/\log\log m)$-approximation for general stochastic f-norm load balancing (Theorem 20).

We do not know how to overcome the impediment discussed above in setting up the auxiliary LP for the general setting. We leave the question of determining the integrality gap of $(\mathsf{LP}(\vec{t}))$, as also the technical question of setting up a suitable auxiliary LP, in the general case, as intriguing open problems.

The final step involved is choosing the "right" $\vec{t}$ vector. Using Claim 4, one can argue that there are only a polynomial number of $\vec{t}$ vectors to consider. We now utilize Lemma 15, which shows that for each $\vec{t}$ (for which $\big(\mathsf{LP}(\vec{t})\big)$ is feasible), we can define a corresponding vector $\vec{b} = \vec{b}(\vec{t}) \in \mathbb{R}^m_{\geq 0}$ such that $f\big(\vec{b}(\vec{t})\big)$ acts as a good proxy for the objective value $\mathbf{E}\big[f(Y)\big]$ of the solution computed for $\vec{t}$, provided that $\mathbf{E}\big[\mathsf{Top}_\ell(Y)\big] = O\big(\mathsf{Top}_\ell(\vec{b})\big)$ for all $\ell \in \mathsf{POS}$. It follows that the solution corresponding to the smallest $f\big(\vec{b}(\vec{t})\big)$ is a near-optimal solution. Notice that we obtain a *deterministic* (algorithm and) approximation guarantee.

VII. Spanning Trees

We now apply our framework to devise an approximation algorithm for *stochastic minimum norm spanning tree*. We are given an undirected graph $G = (V, E)$ with stochastic edge-weights and we are interested in connecting the vertices of this graph to each other. For an edge $e \in E$, we have a nonnegative r.v. X_e that denotes its weight. Edge weights are independent. A feasible solution is a spanning tree $T \subseteq E$ of G. This induces a random weight vector $Y^T = (X_e)_{e\in T}$; note that Y^T follows a product distribution on $\mathbb{R}^{n-1}_{\geq 0}$ where $n := |V|$. The goal is to find a spanning tree T that minimizes $\mathbf{E}\big[f(Y^T)\big]$ for a given monotone, symmetric norm f.

Theorem 22. *There is an $O(1)$-approximation algorithm for stochastic f-norm spanning tree with arbitrary edge-weight distributions.*

Algorithm and analysis sketch. Let T^*, Y^*, and OPT denote an optimal solution, its random weight vector, and the optimal value, respectively. We drop the superscript in Y^T when T is clear from the context. We use Y_e to denote the coordinate in Y corresponding to edge e; we will ensure that Y_e is used only when $e \in T$.

The cost vector Y^T is inherently less complex than the load vector in load balancing, in that each coordinate Y_e^T is an "atomic" random variable whose distribution we can directly access. Thus, our approach is guided by Corollary 16 and Theorem 17, which show that to obtain an approximation guarantee for stochastic f-norm spanning tree, it suffices to find a spanning tree T such that the τ_ℓ statistics of Y^T are "comparable" to those of Y^*.

Our LP relaxation works with a non-increasing vector $\vec{t} \in \mathbb{R}^{\mathsf{POS}_{n-1}}_{\geq 0}$ that is intended to be a guess of $\big(\tau_\ell(Y^*)\big)_{\ell\in\mathsf{POS}}$. By Corollary 16, $\mathbf{E}\big[f(Y)\big] = \Theta\Big(\sum_e \mathbf{E}\big[(Y_e - \tau_1(Y))^+\big] + f(\vec{\tau}(Y))\Big)$. So our LP seeks a (fractional) spanning tree whose cost vector Y minimizes $\sum_e \mathbf{E}\big[(X_e - t_1)\big]^+$ subject to the constraint that $\tau_\ell(Y) \leq t_\ell$ for all $\ell \in \mathsf{POS}$. For $A \subseteq E$, let $\mathsf{comp}(A)$ denote the number of connected components of (V, A).

$$\min \quad \sum_{e\in E} \mathbf{E}\big[(X_e - t_1)^+\big] z_e \qquad (\mathsf{Tree}(\vec{t}))$$

$$\text{s.t.} \quad \sum_{e\in E} \mathbf{Pr}\big[X_e > t_\ell\big] z_e \leq \ell \qquad \forall \ell \in \mathsf{POS} \qquad (6)$$

$$z(A) \leq n - \mathsf{comp}(A) \qquad \forall A \subseteq E \qquad (7)$$

$$z \geq 0, \quad z(E) = n - 1. \qquad (8)$$

It is well known that the polytope defined by (7), (8) is the spanning-tree polytope, i.e., the convex hull of indicator vectors of spanning trees of G. (We remark that all our arguments extend to general matroids, wherein we seek a basis of minimum expected f-norm; the only change to the LP is that (7) is replaced by the constraint that z lie in the matroid base polytope.) The following claim is immediate.

Claim 23. $(\mathsf{Tree}(\vec{t}))$ *is feasible whenever* $t_\ell \geq \tau_\ell\left(Y^*\right)$ *for all* $\ell \in \mathsf{POS}$. *Moreover, the optimal LP value,* $\mathsf{Tree\text{-}OPT}\left(\vec{t}\right)$*, is at most* $\sum_{e \in T^*} \mathbf{E}\left[\left(X_e - \tau_1\left(Y^*\right)\right)^+\right]$.

Suppose that $(\mathsf{Tree}(\vec{t}))$ is feasible, and let $\overline{z}$ be an optimal LP solution. We round $\overline{z}$ to obtain a spanning tree T satisfying $\mathbf{E}\left[f(Y)\right] = O\left(value(\vec{t})\right)$, where we define $value(\vec{t}) := \mathsf{Tree\text{-}OPT}\left(\vec{t}\right) + f(\vec{t'})$, and $\vec{t'} \in \mathbb{R}^{n-1}_{\geq 0}$ denotes the expansion of $\vec{t}$. (Recall that $\vec{t'}$ is defined by setting $\vec{t'}_i := \vec{t}_{2^{\lfloor \log_2 i \rfloor}}$ for all $i \in [n-1]$.) Note that scaling constraint (6) by ℓ yields $O(1)$-bounded column sums in the resulting constraint matrix, and an additive $O(1)$ violation of the scaled constraint translates to an additive $O(\ell)$ violation of constraint (6) for index ℓ. We obtain T by invoking Theorem 5 with the (scaled constraint matrix and) parameter $\nu := \sum_{\ell \in \mathsf{POS}} \frac{1}{\ell} < 2$ to round $\overline{z}$.

Theorem 24. *The weight vector* $Y = Y^T$ *satisfies,*

(i) $\sum_{e \in T} \mathbf{E}\left[\left(Y_e - t_1\right)^+\right] \leq \mathsf{Tree\text{-}OPT}\left(\vec{t}\right)$*;*
(ii) for each $\ell \in \mathsf{POS}$*,* $\tau_{3\ell}\left(Y\right) \leq t_\ell$*; and*
(iii) $\mathbf{E}\left[f(Y)\right] = O(value(\vec{t}))$.

Proof: Parts (i) and (ii) follow from parts (a) and (b) of Theorem 5 respectively. Part (i) is immediate from part (a). For part (ii), since we invoke Theorem 5 with $\nu < 2$, by part (b), we have $\sum_{e \in T} \mathbf{Pr}\left[Y_e > t_\ell\right] < 3\ell$, and hence, $\tau_{3\ell}\left(Y\right) \leq t_\ell$. For part (iii), observe that $Y, \vec{t}$ satisfy the assumptions of Theorem 17(a) with $\alpha = 1$ and $\beta = 3$, so we obtain that $\mathbf{E}\left[f(Y)\right] \leq 126\left(\sum_{e \in T} \mathbf{E}\left[\left(Y_e - t_1\right)^+\right] + f(\vec{t'})\right)$, and the RHS expression is $O(value(\vec{t}))$ by part (i). ■

To finish up the proof of Theorem 22, using Claim 4, we argue that we can obtain a polynomial-size set $\mathcal{T} \subseteq \mathbb{R}^{\mathsf{POS}}_{\geq 0}$ containing a vector $\vec{t^*}$ such that $\tau^*_\ell \leq t^*_\ell \leq 2\tau^*_\ell + \frac{\mathsf{UB}}{m^2}$, where UB is such that $\frac{\mathsf{UB}}{m} \leq \mathsf{OPT} \leq \mathsf{UB}$. We return the solution returned by our algorithm for the vector that minimizes $value(\vec{t})$ among all vectors $\vec{t} \in \mathcal{T}$ for which $(\mathsf{Tree}(\vec{t}))$ is feasible. Let $\overline{Y}$ be the cost vector of the resulting solution. By Claim 23, $(\mathsf{Tree}(\vec{t^*}))$ is feasible, and so using Theorem 24, we have $\mathbf{E}\left[f(\overline{Y})\right] = O\left(value(\vec{t^*})\right)$. Finally, given the bounds on the t^*_ℓ values, using Theorem 17(b) and Claim 23, we obtain that $value(\vec{t^*}) = O(\mathsf{OPT}) + \frac{\mathsf{UB}}{m} = O(\mathsf{OPT})$.

REFERENCES

[1] S. Alamdari and D. B. Shmoys, "A bicriteria approximation algorithm for the k-center and k-median problems," in *Proceedings of the 15th WAOA*, 2017, pp. 66–75.

[2] A. Aouad and D. Segev, "The ordered k-median problem: Surrogate models and approximation algorithms," *Mathematical Programming*, vol. 177, pp. 55—-83, 2019.

[3] B. Awerbuch, Y. Azar, E. F. Grove, M. Kao, P. Krishnan, and J. S. Vitter, "Load balancing in the L_p norm," in *Proceedings of the 36th FOCS*, 1995, pp. 383–391.

[4] Y. Azar and A. Epstein, "Convex programming for scheduling unrelated parallel machines," in *Proceedings of the 37th STOC*, 2005, pp. 331–337.

[5] J. Byrka, K. Sornat, and J. Spoerhase, "Constant-factor approximation for ordered k-median," in *Proceedings of the 50th STOC*, 2018, pp. 620–631.

[6] D. Chakrabarty and C. Swamy, "Interpolating between k-median and k-center: Approximation algorithms for ordered k-median," in *Proceedings of the 45th ICALP*, 2018, pp. 29:1–29:14.

[7] ——, "Approximation algorithms for minimum norm and ordered optimization problems," in *Proceedings of the 51st STOC*, 2019, pp. 126–137.

[8] ——, "Simpler and better algorithms for minimum-norm load balancing," in *Proceedings of the 27th ESA*, 2019, pp. 27:1–27:12.

[9] A. Goel and P. Indyk, "Stochastic load balancing and related problems," in *Proceedings of the 40th FOCS*, 1999, pp. 579–586.

[10] A. Gupta, A. Kumar, V. Nagarajan, and X. Shen, "Stochastic load balancing on unrelated machines," in *Proceedings of the 29th SODA*, 2018, pp. 1274–1285.

[11] A. Gupta and K. Tangwongsan, "Simpler analyses of local search algorithms for facility location," *CoRR*, vol. abs/0809.2554, 2008.

[12] V. Gupta, B. Moseley, M. Uetz, and Q. Xie, "Greed works—online algorithms for unrelated machine stochastic scheduling," *Mathematics of Operations Research*, vol. 45, no. 2, pp. 497–516, 2020.

[13] J. Y. Hui, "Resource allocation for broadband networks," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1598–1608, 1988.

[14] S. Im, B. Moseley, and K. Pruhs, "Stochastic scheduling of heavy-tailed jobs," in *Proceedings of the 32nd STACS*, 2015, pp. 474–486.

[15] J. M. Kleinberg, Y. Rabani, and É. Tardos, "Allocating bandwidth for bursty connections," *SIAM J. Comput.*, vol. 30, no. 1, pp. 191–217, 2000.

[16] G. Laporte, S. Nickel, and F. S. da Gama, *Location science*. Springer, 2019.

[17] R. Latała, "Estimation of moments of sums of independent real random variables," *The Annals of Probability*, vol. 25, pp. 1502–1513, 1997.

[18] J. Li and A. Deshpande, "Maximizing expected utility for stochastic combinatorial optimization problems," *Mathematics of Operations Research*, vol. 44, no. 1, pp. 354–375, 2019.

[19] J. Li and Y. Liu, "Approximation algorithms for stochastic combinatorial optimization problems," *Journal of the Operations Research Society of China*, vol. 4, no. 1, pp. 1–47, 2016.

[20] J. Li and W. Yuan, "Stochastic combinatorial optimization via poisson approximation," in *Proceedings of the 45th STOC*, 2013, pp. 971–980.

[21] A. Linhares, N. Olver, C. Swamy, and R. Zenklusen, "Approximate multi-matroid intersection via iterative refinement," *Math. Program.*, vol. 183, no. 1, pp. 397–418, 2020.

[22] K. Makarychev and M. Sviridenko, "Solving optimization problems with diseconomies of scale via decoupling," *J. ACM*, vol. 65, no. 6, pp. 42:1–42:27, 2018.

[23] R. H. Möhring, A. S. Schulz, and M. Uetz, "Approximation in stochastic scheduling: the power of lp-based priority policies," *J. ACM*, vol. 46, no. 6, pp. 924–942, 1999.

[24] M. Molinaro, "Stochastic ℓ_p load balancing and moment problems via the L-function method," in *Proceedings of the 30th SODA*, 2019, pp. 343–354.

[25] S. Nickel and J. Puerto, *Location Theory: A Unified Approach*. Springer Berlin Heidelberg, 2006.

[26] M. Pinedo, "Offline deterministic scheduling, stochastic scheduling, and online deterministic scheduling," in *Handbook of Scheduling*. Chapman and Hall/CRC, 2004.

[27] D. B. Shmoys and É. Tardos, "An approximation algorithm for the generalized assignment problem," *Math. Program.*, vol. 62, pp. 461–474, 1993.