A Polynomial-time Approximation Scheme for Fault-tolerant Distributed Storage

Constantinos Daskalakis* Anindya De† Ilias Diakonikolas‡ Ankur Moitra§

Rocco A. Servedio¶

October 6, 2013

Abstract

We consider a problem which has received considerable attention in systems literature because of its applications to routing in delay tolerant networks and replica placement in distributed storage systems. In abstract terms the problem can be stated as follows: Given a random variable X generated by a known product distribution over $\{0,1\}^n$ and a target value $0 \leq \theta \leq 1$, output a non-negative vector w, with $\|w\|_1 \leq 1$, which maximizes the probability of the event $w \cdot X \geq \theta$. This is a challenging non-convex optimization problem for which even computing the value $\mathbf{Pr}[w \cdot X \geq \theta]$ of a proposed solution vector w is #P-hard.

We provide an additive EPTAS for this problem which, for constant-bounded product distributions, runs in $\mathrm{poly}(n) \cdot 2^{\mathrm{poly}(1/\epsilon)}$ time and outputs an ϵ-approximately optimal solution vector w for this problem. Our approach is inspired by, and extends, recent structural results from the complexity-theoretic study of linear threshold functions. Furthermore, in spite of the objective function being non-smooth, we give a *unicriterion* PTAS while previous work for such objective functions has typically led to a *bicriterion* PTAS. We believe our techniques may be applicable to get unicriterion PTAS for other non-smooth objective functions.

1 Introduction

Many applications involve designing a system that will perform well in an uncertain environment. Sources of uncertainty include (for example) the demand when we are designing a server, the congestion when we are designing a routing protocol, and the failure of the system's own components when we are designing a distributed system. Such uncertainties are often modeled as stochastic variables, giving rise to non-linear and non-convex optimization problems. In this paper, we study a non-convex stochastic optimization problem that has received considerable attention in the systems literature [JDPF05, Fal03, LDT09, LDT10a, LDT10b, SRFS10] but has remained poorly understood.

The main motivation for studying this problem comes from *distributed storage* [DPR05, JB03, LDT09, SRFS10]. The goal in this literature is to develop methods for storing data among a set of faulty processors in a way that makes it possible to recover the data in its entirety despite processor failures. Clearly, to perform this task we need to use some form of redundancy, as otherwise a single processor failure could cause permanent loss of data. In particular, this task contains as subproblems both the choice of an error correcting code and the decision of how to allocate the encoded data into the failure-prone processors, resulting in an enormous design space.

An important observation that is used throughout the literature is that these two subproblems can be decoupled through the use of erasure codes (see, e.g., [Lub02, LMSS02, Mit04, Sho06]). Such codes can be used to encode the original data so that with high probability *any* large enough subset of encoded data can be used to reconstruct the original data. In view of this observation, we can formulate the distributed storage problem as a much simpler to state problem:

Suppose that our original data has size θ GB for some $0 \leq \theta \leq 1$, and we use an erasure code to generate 1 GB of encoded data. The goal is to allocate the data among n failure-prone nodes so as to maximize the probability that the original data can be recovered. The standard formulation of the problem [DPR05, JB03, LDT09, SRFS10] is that each node i has some known

*MIT. Email: costis@csail.mit.edu. Supported by a Sloan Foundation Fellowship, a Microsoft Research Faculty Fellowship, and NSF Awards CCF-0953960 (CAREER) and CCF-1101491.

†Simons Institute, University of California, Berkeley. Email: anindya@cs.berkeley.edu. Work done when the author was supported by Umesh Vazirani's Templeton Foundation Grant 21674.

‡University of Edinburgh. Email: ilias.d@ed.ac.uk. Part of this work was done while the author was a postdoc at the University of California, Berkeley supported by a Simons Fellowship. Supported in part by a SICSA PECE grant.

§MIT. Email: moitra@mit.edu. Part of this research was done while the author was a postdoc at the Institute for Advanced Study and was supported by NSF grant No. DMS-0835373 and by an NSF Computing and Innovation Fellowship.

¶Columbia University. Email: rocco@cs.columbia.edu. Supported by NSF grants CCF-1115703 and CCF-1319788.

probability $1 - p_i$ of failing, and that these failures are independent across different nodes. So, mathematically our goal is to solve the following problem, which we call Problem (P):

Input: An n-vector of probabilities $p = (p_1, \ldots, p_n) \in [0,1]^n$ and a parameter $\theta \in [0,1]$.

For $i \in [n]$ let μ_{p_i} be the distribution on $\{0,1\}$ with $\mu_{p_i}(1) = p_i$, and let the corresponding product distribution over $\{0,1\}^n$ be denoted by $\mathcal{D}_p = \bigotimes_{i=1}^n \mu_{p_i}$.

Output: A weight vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n_{\geq 0}$ satisfying $\|w\|_1 \leq 1$ (such a w is said to be a *feasible solution*). The goal is to maximize

$$\mathrm{Obj}(w) \stackrel{\text{def}}{=} \mathbf{Pr}_{X \sim \mathcal{D}_p}[w \cdot X \geq \theta].$$

A feasible solution that maximizes $\mathrm{Obj}(w)$ is said to be an *optimal solution*. We will denote by $\mathsf{OPT} = \mathsf{OPT}(p, \theta)$ the maximum value of any feasible solution.

In the above formulation w_i denotes the amount of data that we decide to store in the i-th storage node, and X_i is the indicator random variable of the event that the i-th storage node does not fail.

Before we proceed, we point out a connection of the optimization problem (P) above with the class of *Boolean halfspaces* or *Linear Threshold Functions (LTFs)* that will be crucially exploited throughout this paper. Recall that a Boolean function $f : \{0,1\}^n \to \{0,1\}$ is a halfspace if there exists a weight-vector $v \in \mathbb{R}^n$ and a threshold $t \in \mathbb{R}$ so that $f(x) = 1$ if and only if $v \cdot x \geq t$. Hence, the objective function value $\mathrm{Obj}(w)$ of a feasible weight-vector w (i.e., $w \in \mathbb{R}^n_{\geq 0}$ with $\|w\|_1 \leq 1$) can be equivalently expressed as $\mathrm{Obj}(w) = \mathbf{Pr}_{X \sim \mathcal{D}_p}[h_{w,\theta}(X) = 1]$, where $h_{w,\theta}(x) = \mathbf{1}_{\{x \in \{0,1\}^n : w \cdot x \geq \theta\}}$ is the halfspace with weight-vector w and threshold θ.

We remark that, even though the feasible set is continuous, it is not difficult to show that there exists a rational optimal solution. In particular, analogous to the linear-algebraic arguments [MTT61, Mur71, Rag88], we can also show that there always exists an optimal solution with bit-complexity polynomially bounded in n; in fact, one with at most $O(n^2 \log n)$ bits which is best possible [Has94]. (As a corollary, the supremum is always attained and problem (P) is well-defined.)

1.1 Previous and Related Work. Previous Work on the Problem.

The stochastic design problem (P) stated above was formulated explicitly in the work of Jain et al. [JDPF05]. That work was motivated by the problem of routing in Delay Tolerant Networks [JFP04]. These networks are characterized by a lack of consistent end-to-end paths, due to interruptions that may be either planned or unplanned, and selecting routing paths is considered to be one of the most challenging problems. The authors of [JDPF05] reduce the route selection problem to Problem (P) in a range of settings of interest, and study the structure of the optimal partition as well as its computational complexity, albeit with inconclusive theoretical results.

One of the special cases of the problem considered in [JDPF05] is the case where all the p_i's are equal, i.e., when $p_1 = \ldots = p_n = p$. Even in this case, the structure of the optimal solution is not well-understood. It is natural to expect that the optimal weight vector is obtained by equally splitting the allowed unit of weight over a subset of the indices, and setting the weights to be 0 on all other indices (in other words, set $w_1 = w_2 = \ldots = w_k = \frac{1}{k}$ and $w_{k+1} = \ldots = w_n = 0$, for some k). The authors of [JDPF05] consider the performance of this strategy for different values of p and θ, as do the papers [LDT09, LDT10a, LDT10b]. Surprisingly, such partitions are not necessarily optimal. For a counter-example, communicated to us by R. Kleinberg [Kle06], consider the setting where $n = 5$, $\theta = 5/12$ and $p = 1 - \epsilon$, for sufficiently small ϵ. In this case, the allocation vector $w = (1/4, 1/4, 1/6, 1/6, 1/6)$ performs better than the uniform weight vector over any subset of the coordinates $\{1, \ldots, 5\}$. There has also been work on a related distributed storage problem [SRFS10] that uses a slightly different model of node failures. In this model, instead of every node failing with probability p, a random subset of nodes of size pn is assumed to fail. In this setting, the conjecture that certain symmetric allocations are optimal is related to a conjecture of Erdos [Erd65] on the maximum number of edges in a k-uniform hypergraph whose (fractional) matching size is at most s (see [AFH+12] for a detailed discussion of the connection).

Related Work. Stochastic optimization is an important research area with diverse applications having its roots in the work of Dantzig [Dan55] and Beale [Bea55] that has been extensively studied since the 1950's (see e.g., [BL97] for a book on the topic). During the past couple of decades, there has been an extensive literature on efficient approximation algorithms for stochastic combinatorial optimization problems in various settings, see e.g., [KRT97b, DGV08, BGK11, Nik10, Swa11, LD11, LY13] and references therein.

In many of these works, one wants to select a subset of (discrete independent) random variables whose sum optimizes a certain non-linear function. For example, the objective function of our problem (P) corresponds to the *threshold probability maximization* problem [Nik10, LY13]. Note that, while the solution space in the aforementioned works is typically discrete and finite in nature, the solution space for our problem is continuous. In particular, it is not always possible to discretize the space without losing a lot in the objective function

value (see Section 1.3 for a detailed explanation of the difficulties in our setting).

Regarding *threshold probability maximization*, Li and Yuan [LY13] obtained *bicriterion* additive PTAS for stochastic versions of classical combinatorial problems, such as shortest paths, spanning trees, matchings and knapsacks. Roughly, they obtain a bicriterion guarantee because the function to be optimized does not have a bounded Lipschitz constant. In contrast, even though the $\mathbf{Pr}[w \cdot X \geq \theta]$ function that we are optimizing does not have a bounded Lipschitz constant, we are able to obtain a *unicriterion* PTAS by exploiting new structural properties of near-optimal solutions that we establish in this work, as described below. In terms of techniques, [LY13] use Poisson approximation and discretization as a main component of their results. We note that this approach is not directly applicable in our setting, since we are dealing with a *weighted* sum of Bernoulli random variables with arbitrary *real* weights and we are shooting for a unicriterion PTAS. We view the unicriterion guarantee that we achieve as an important contribution of the techniques in this work.

1.2 Our results. It seems unlikely that Problem (P) can be solved exactly in polynomial time. Note that (even for the very special case when each p_i equals $1/2$) (exactly) evaluating the objective function $\mathrm{Obj}(w)$ of a candidate solution w is $\#P$-hard. (This follows by a straightforward reduction from the counting version of knapsack, see e.g., Theorem 2.1 of [KRT97a] for a proof.) In fact, problem (P) is easily seen to lie in $NP^{\#P}$, and we are not aware of a better upper bound. We conjecture that the exact problem is intractable, namely $\#P$-hard.

The focus of this paper is on efficient approximation algorithms. As our main contribution, we give an additive EPTAS for (P) for the case that each p_i is bounded away from 0. That is, we give an algorithm that for every $\epsilon > 0$, outputs a feasible solution w such that $\mathrm{Obj}(w)$ is within an additive ϵ of the optimal value. An informal statement of our main result follows (see Theorem 3 for a detailed statement):

THEOREM 1. *[Main Result – informal statement] Fix any $\epsilon > 0$ and let $p = (p_1, \ldots, p_n)$ be any input instance such that $\min_i p_i \geq \epsilon^{\Omega(1)}$. There is a randomized algorithm which, for any such input vector p and any input threshold $0 \leq \theta \leq 1$, runs in $\mathrm{poly}(n) \cdot 2^{\mathrm{poly}(1/\epsilon)}$ time and with high probability outputs a feasible solution vector w whose value is within an additive ϵ of the optimal.*

1.3 Our techniques. Background. In recent years, there has been a surge of research interest in concrete complexity theory on various problems concerning halfspaces. These include constructions of low weight approximations of halfspaces [Ser07, DS09, DDFS12], PRGs for halfspaces [DGJ+10, MZ10], property testing algorithms [MORS10] and approximate reconstruction of halfspaces from low-degree Fourier coefficients [OS11, DDFS12] among others.

All these results use a "structure versus randomness" tradeoff for halfspaces which can be described roughly as follows: Consider the weights of a halfspace $\mathbf{1}_{\{x \in \{0,1\}^n : w \cdot x \geq \theta\}}$ in order of decreasing magnitude. If the largest-magnitude weight is "small" compared to the *2-norm* of the weight-vector w, then the Berry-Esséen theorem (a quantitative version of the Central Limit Theorem with explicit error bounds) implies that for independent $\{0,1\}$ random variables X_i (that are not too biased towards 0 or 1), the distribution of $w \cdot X$ will be well-approximated by the Gaussian distribution with the same mean and variance. This is a very useful statement because it implies that the discrete random variable $w \cdot X$ essentially inherits several nice properties of the Gaussian distribution (such as anti-concentration, strong tail bounds, and so on). On the other hand, if the largest-magnitude weight accounts for a significant fraction of the 2-norm, then the weight-vector obtained by erasing this weight has significantly smaller 2-norm, and we have "made progress;" intuitively, after a bounded number of steps of this sort, the 2-norm of the remaining weights will be extremely small, so the halfspace essentially depends only on the first few variables and should be "easy to handle" for that reason. These arguments can be made quantitatively precise using the notion of the "critical index" (introduced in [Ser07]; see Definition 2.2) which plays an important role in much of the work described above.

Our Contribution. In this paper we show how tools from the complexity-theoretic literature on halfspaces alluded to above can be leveraged in order to make algorithmic progress on our optimization problem (P). As we will explain below, several non-trivial technical issues arise in the context of problem (P) which require careful treatment.

At a high-level, in this work we adapt and enhance this technical machinery in order to obtain a structural understanding of the problem, which is then combined with algorithmic and probabilistic techniques to obtain a PTAS. Very roughly, we proceed as follows: We partition the space of *optimal* solution vectors v^* into a constant number of subsets, based on the value of the critical index of v^*. For each subset we apply a (different) algorithm which outputs a candidate (feasible) solution which is guaranteed to be ϵ-optimal, *assuming v^* belongs to the particular subset.* Since at least one subset contains an optimal solution, the best candidate solution will be ϵ-approximately optimal as desired.

Of course, we need to explain how to compute a candidate solution for each subset. A basic difficulty comes from the fact that our problem is not combinatorial. The space of feasible solutions is continuous and even though one can easily argue that there exists a rational

optimal solution with polynomially many bits, a priori we do not know anything more about its structure. We note that a natural first approach one would think to try in this context would be to appropriately "discretize" the weights (e.g., by using a geometric subdivision, etc) and then use dynamic programming to optimize in the discretized space. However, it is far from clear how to show that such a naive discretization works; one can easily construct examples of weight vectors w such that "rounding" the coefficients of w to an appropriate (inverse polynomial in n) granularity radically changes the value of the objective function[1].

To compute an approximately optimal solution for each case (i.e., for v^* in a particular subset as described above) one needs a better understanding of the structure of the optimal solutions. The reason why "rounding" the coefficients may substantially change the objective function value is because for certain weight vectors w the random variable $w \cdot X$ is very concentrated, i.e., it puts a substantial fraction of its probability mass in a small interval. If on the other hand, $w \cdot X$ is sufficiently *anti-concentrated*, i.e., it puts small mass on every small interval, then it is easy to argue that "rounding" does not affect the objective function by a lot. Known results [TV09] show that the anti-concentration of $w \cdot X$ depends strongly on the *additive structure* of w. While it is hopeless to show that all feasible weight-vectors are anti-concentrated, one could hope to show that there exists a *near-optimal* solution that has good anti-concentration. Essentially, this is what we do.

Our main structural theorem (Theorem 2) shows that, except in degenerate cases, there always exists an optimal solution whose "tail" has sufficiently large L_1-norm compared with the "head" [2]. We remark that, while results of a broadly similar flavor appear in many of these previous papers (see e.g., [Ser07, OS11, DS09]) there are a few crucial differences. First, the previous works compare the L_2 norms of the "head" and the "tail". Most importantly, all previous such results consist of re-expressing the LTFs in a "nice" form (which includes changing the value of the threshold θ). Indeed, the previous arguments which assert the existence of these nice forms do not control the value of the threshold as its exact value is immaterial. In contrast, for our problem the exact threshold in comparison to the L_1-norm of the weight vector is a crucial parameter. Our structural theorem says that every LTF has a well-structured equivalent version in which (1) the threshold stays exactly the same relative to the L_1-*norm* of the weights, and (2) L_1-*norm* of the "tail weights" is "large." Our proof of this theorem is based on *linear fractional programming*, which is novel in this context of structural results for LTFs. Conceptually, our structural theorem serves as a "pre-processing" step which ensures that the optimal weight-vector may be assumed to be well-structured; our algorithm crucially exploits this nice structure of the optimal solution to efficiently find a near-optimal solution.

[1]Moreover, we note that discretization of the space followed by standard approaches, e.g., along the lines of [CK05], seems to inherently lead to *bicriteria* guarantees.

[2]If the optimal weight vector only has nonzero coordinates in the L coordinates in the "head" (think of L as a constant – it will depend only on ϵ), then as we show we can find an optimal vector exactly in $\mathrm{poly}(n) \cdot 2^{\mathrm{poly}(L)}$ time by an enumeration-based approach.

2 Preliminaries

2.1 Simplifying assumptions about the problem instance.

It is clear that if $\theta = 0$ or $\theta = 1$ then it is trivial to output an optimal solution; hence throughout the rest of the paper we assume that $0 < \theta < 1$.

Without loss of generality we may make the following assumptions about the input $(p_1, \ldots, p_n)$:

(A1) $p_1 \geq \cdots \geq p_n$.

(A2) $p_1 < 1 - \epsilon$ and all $p_i \in \{\epsilon/(4n), \ldots, k\epsilon/(4n)\}$, where $k\epsilon/(4n)$ is the largest integer multiple of $\epsilon/(4n)$ that is less than $1 - \epsilon$. For the first claim, note that if $p_1 \geq 1 - \epsilon$ then the solution $w = (1, 0, \ldots, 0)$ has $\mathbf{Pr}_{X \sim \mathcal{D}_p}[w \cdot X \geq \theta] \geq 1 - \epsilon$ and hence $(1, 0, \ldots, 0)$ is an ϵ-optimal solution as desired. For the second claim, given an input vector of arbitrary values $p' = (p'_1, \ldots, p'_n) \in [0, 1 - \epsilon)^n$, if we round the p'_i values to integer multiples of $\epsilon/4n$ to obtain $p = (p_1, \ldots, p_n)$, then a simple coupling argument gives that for any event S, we have $\left|\mathbf{Pr}_{X \sim \mathcal{D}_p}[S] - \mathbf{Pr}_{X \sim \mathcal{D}_{p'}}[S]\right| \leq \epsilon/4$. Hence for our purposes, we may assume that the initial p_i values are "$\epsilon/(4n)$-granular" as described above.

We further make some easy observations about optimal solutions that will be useful later. First, it is clear that there exists an optimal solution w with $\|w\|_1 = 1$. (If $\|w\|_1 < 1$ then rescaling by $\|w\|_1$ gives a new feasible solution whose value is at least as good as the original one.) Second, by assumption (A1) there exists an optimal solution $w \in \mathbb{R}^n_+$ that satisfies $w_i \geq w_{i+1}$ for all $i \in [n-1]$. (If $w_i < w_{i+1}$ it is easy to see that by swapping the two values we obtain a solution whose value is at least as good as the original one.)

2.2 Tools from structural analysis of LTFs: regularity and the critical index.

DEFINITION 2.1. (REGULARITY) *Fix any real value $\tau > 0$. We say that a vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ is τ-*regular *if $\max_{i \in [n]} |w_i| \leq \tau \|w\|_2$. A linear form $w \cdot x$ is said to be τ-regular if w is τ-regular.*

Intuitively, regularity is a helpful notion because if w is τ-regular then the Berry-Esséen theorem can be used to show that for $X \sim \mathcal{D}_p$, the linear form $w \cdot X$ is distributed like a Gaussian (with respect to Kolmogorov distance)

up to an error of η, where η depends on the regularity parameter and the parameters $p_1, \ldots, p_n$ (see Corollary 6.1).

A key ingredient in our analysis is the notion of the "critical index" of a linear form $w \cdot x$. The critical index was implicitly introduced and used in [Ser07] and was explicitly used in [DS09, DGJ+10, OS11, DDFS12] and other works. Intuitively, the critical index of w is the first index i such that from that point on, the vector $(w_i, w_{i+1}, \ldots, w_n)$ is regular. A precise definition follows:

DEFINITION 2.2. (CRITICAL INDEX) *Given a vector $w \in \mathbb{R}^n$ such that $|w_1| \geq \cdots \geq |w_n| > 0$, for $k \in [n]$ we denote by σ_k the quantity $\sqrt{\sum_{i=k}^n w_i^2}$. We define the* τ-critical index $c(w, \tau)$ of w *as the smallest index $i \in [n]$ for which $|w_i| \leq \tau \cdot \sigma_i$. If this inequality does not hold for any $i \in [n]$, we define $c(w, \tau) = \infty$.*

Given a problem instance p satisfying (A1) and (A2) and a value ϵ, we define

(2.1)
$$L = \min\{n, \Theta(1/(\epsilon^2\gamma^2)\cdot(1/\gamma)\cdot(\log 1/(\epsilon\gamma))\cdot(\log(1/\epsilon))\},$$

where $\gamma = \min\{p_n, 1 - p_1\} \geq \epsilon/4n$. The idea behind this choice of L is that it is the cutoff for "having a large $(\epsilon\gamma)/200$-critical index."

2.3 A useful structural theorem about solutions. In Section 3 we prove that given any feasible solution, there is another feasible solution whose value is at least as good as the original one and which has a "heavy tail" with respect to the L_1 norm:

THEOREM 2. *Fix $K \in [n]$, $0 < \theta < 1$, and $w_1 \geq \cdots \geq w_n \geq 0$ such that $\sum_{i=1}^n w_i = 1$. Let $S = \{x \in \{0,1\}^n : w \cdot x \geq \theta\}$. Then there is a vector $v = (v_1, \ldots, v_n)$ such that*

(a) $\sum_{i=1}^n v_i = 1$ and $v_1 \geq \cdots \geq v_n \geq 0$;

(b) every $x \in S$ has $v \cdot x \geq \theta$; and

(c) either $v_{K+1} = \cdots = v_n = 0$ or else $\sum_{i=1}^k v_i \leq (K+2)^{(K+2)/2} \cdot \sum_{i=K+1}^n v_i$.

Applying Theorem 2 with $K = L$ as defined in (2.1), we get that there exists an optimal solution v^* that satisfies (a) and (b), and either $v^*_{L+1} = \cdots = v^*_n = 0$ or else $\sum_{i=1}^L v_i^* \leq (L+2)^{(L+2)/2} \cdot \sum_{i=L+1}^n v_i^*$. *Throughout the paper, we fix v^* to be such an optimal solution vector.*

2.4 Our approach and formal statement of the main result. At a high level, our approach is to consider three mutually exclusive and exhaustive cases for v^*:

- **Case 1:** v^* has $v^*_{L+1} = \cdots = v^*_n = 0$. In this case we say v^* is an *L-junta*. (Note that if $L = n$ then we are in this case; hence in Cases 2 and 3 we have that $L < n$.)
- **Case 2:** v^* is not an L-junta and $c(v^*, \epsilon\gamma/200) > L$. In this case we say that v^* is of *type* $L + 1$.
- **Case 3:** v^* is not an L-junta and $c(v^*, \epsilon\gamma/200) = K$ for some $K \in \{1, \ldots, L\}$. In this case we say that v^* is of *type* K.

We show (see Section 4) that in Case 1 it is possible to efficiently compute an *exactly* optimal solution. In both Cases 2 and 3 (see Sections 5 and 6 respectively) we show that it is possible (using two different algorithms) to efficiently construct a set of $N \leq \mathrm{poly}(n, 2^{\mathrm{poly}(L)})$ feasible solutions such that one of them (call it w') has $\mathrm{Obj}(w') \geq \mathsf{OPT} - \epsilon/2$. Running all three procedures, we thus obtain a set of $O(nN) = \mathrm{poly}(n, 2^{\mathrm{poly}(L)})$ candidate solutions such that one of them (call it $\widetilde{w}$) is guaranteed to have $\mathrm{Obj}(\widetilde{w}) \geq \mathsf{OPT} - \epsilon/2$. From this it is simple to obtain an ϵ-approximate optimal solution (see Section 7).

A precise version of our main result is given below, where by $\mathrm{bit}(\theta)$ we denote the bit-length of θ:

THEOREM 3. *[Main Result] There is a randomized algorithm with the following performance guarantee: It takes as input a vector of probabilities $p = (p_1, \ldots, p_n)$ satisfying (A1) and (A2), a threshold value $0 < \theta < 1$, and a confidence parameter $0 < \delta < 1$. It runs in $\mathrm{poly}(n, 2^{\mathrm{poly}(1/\epsilon, 1/\gamma)}, \mathrm{bit}(\theta)) \cdot \log(1/\delta)$ time, where $\gamma = \min\{p_n, 1-p_1\} \geq \epsilon/4n$. With probability $1-\delta$ it outputs a feasible solution $\widetilde{w}$ such that $\mathrm{Obj}(\widetilde{w}) \geq \mathsf{OPT} - \epsilon$, and an estimate $\widetilde{\mathrm{Obj}}(\widetilde{w})$ of $\mathrm{Obj}(\widetilde{w})$ that satisfies $|\widetilde{\mathrm{Obj}}(\widetilde{w}) - \mathrm{Obj}(\widetilde{w})| \leq \epsilon$.*

3 There exist well-structured optimal solutions: Proof of Theorem 2

Fix $K \in [n]$, $0 < \theta < 1$, and $w = (w_1, \ldots, w_n)$ with $w_1 \geq w_2 \geq \ldots w_n \geq 0$ and $\sum_{i=1}^n w_i = 1$. If $w_i = 0$ for all $i \in [K+1, n]$ it is clear that the weight-vector w satisfies conditions (a)-(c). So, we will henceforth assume that $W_T \stackrel{\text{def}}{=} \sum_{i=K+1}^n w_i > 0$.

We start by defining the following *linear–fractional program* $(\mathcal{LFP})$ over variables $u_1, \ldots, u_K$ and r. $(\mathcal{LFP})$ is defined by the following set of linear constraints:

(i) For all $x \in S$, it holds $\sum_{i=1}^K u_i x_i + \sum_{i=K+1}^n w_i x_i \geq r$.

(ii) For all $i \in [K-1]$, $u_i \geq u_{i+1}$; and $u_K \geq w_{K+1}$.

The (fractional) objective function to be maximized is

$$f_0(u_1, \ldots, u_K, r) = \frac{r}{\sum_{i=1}^K u_i + W_T}.$$

Observe that $(u_1, \ldots, u_K, r) = (w_1, \ldots, w_K, \theta)$ is a feasible solution, hence the maximum value of $(\mathcal{LFP})$ is at least θ.

We now proceed to turn $(\mathcal{LFP})$ into an essentially equivalent linear program $(\mathcal{LP})$, using the standard

Charnes–Cooper transformation [CC62]. The linear program $(\mathcal{LP})$ has variables t, $s_1, \ldots, s_K$ and δ and is defined by the following set of linear constraints:

(i) For all $x \in S$, it holds $\sum_{i=1}^{K} s_i x_i + \left(\sum_{i=K+1}^{n} w_i x_i\right) \cdot t \geq \delta$.

(ii) For all $i \in [K-1]$, $s_i \geq s_{i+1}$; and $s_K \geq w_{K+1} \cdot t$.

(iii) $\sum_{i=1}^{K} s_i + W_T \cdot t = 1$; and

(iv) $t \geq 0$.

The linear objective function to be maximized is δ.

The following standard claim (see e.g., [BV04]) quantifies the relation between the two aforementioned optimization problems:

CLAIM 3.1. *The optimization problems $(\mathcal{LFP})$ and $(\mathcal{LP})$ are equivalent.*

Proof. Let $(u_1^*, \ldots, u_K^*, r^*)$ be a feasible solution to $(\mathcal{LFP})$. It is straightforward to verify that the vector $(t^*, s_1^*, \ldots, s_K^*, \delta^*)$ with

$$t^* = \frac{1}{\sum_{i=1}^{K} u_i^* + W_T},$$

$s_i^* = t^* u_i^*$, for $i \in [K]$, and $\delta^* = t^* r^*$ is a feasible solution to $(\mathcal{LP})$ with the same objective function value. It follows that the linear program $(\mathcal{LP})$ is also feasible with maximum value at least θ. Moreover, the maximum value of $(\mathcal{LP})$ is greater than or equal to the maximum value of $(\mathcal{LFP})$.

Conversely, if $(t^*, s_1^*, \ldots, s_K^*, \delta^*)$ is a feasible solution to $(\mathcal{LP})$ with $t^* \neq 0$, then $(u_1^*, \ldots, u_K^*, r^*)$ with $u_i^* = s_i^*/t^*$ and $r^* = \delta^*/t^*$ is feasible for $(\mathcal{LFP})$, with the same objective function value

$$\delta^* = \frac{r^*}{\sum_{i=1}^{K} u_i^* + W_T}.$$

If $(t^*, s_1^*, \ldots, s_K^*, \delta^*)$ is a feasible solution to $(\mathcal{LP})$ with $t^* = 0$ and $(u_1^*, \ldots, u_K^*, r^*)$ is feasible to $(\mathcal{LFP})$ then

$$(\widetilde{u_1}, \ldots, \widetilde{u_K}, \widetilde{r}) = (u_1^*, \ldots, u_K^*, r^*) + \lambda(s_1^*, \ldots, s_K^*, \delta^*)$$

is feasible to $(\mathcal{LFP})$ for all $\lambda \geq 0$. Moreover, note that

$$\lim_{\lambda \to \infty} f_0(\widetilde{u_1}, \ldots, \widetilde{u_K}, \widetilde{r}) = \frac{\delta^*}{\sum_{i=1}^{K} s_i^*} = \delta^*.$$

So, we can find feasible solutions to $(\mathcal{LFP})$ with objective values arbitrarily close to the objective value of $(t^* = 0, s_1^*, \ldots, s_K^*, \delta^*)$. Therefore, the maximum value of $(\mathcal{LFP})$ is greater than or equal to the maximum value of $(\mathcal{LP})$.

Combining the above completes the proof of the claim.

We now proceed to analyze the linear program $(\mathcal{LP})$. We will show that there exists a feasible solution to $(\mathcal{LP})$ with properties that will be useful for us. Note that S is by definition non-empty. In particular, the all 1's vector belongs to S. Hence, because of constraint (iii), the optimal value δ^* of $(\mathcal{LP})$ is at most 1 (i.e., $(\mathcal{LP})$ is bounded). Consider a vertex $\mathbf{v}^* = (t^*, s_1^*, \ldots, s_K^*, \delta^*)$ of the feasible set of $(\mathcal{LP})$ maximizing the objective function δ. Claim 3.1 and the observation that the optimal value of $(\mathcal{LFP})$ is at least θ imply that $\delta^* \geq \theta$. We consider the following two cases:

[**Case I:** $t^* = 0$.] In this case, we select the desired vector $v = (v_1, \ldots, v_n)$ as follows: We set $v_i = s_i^*$ for all $i \in [K]$ and $v_i = 0$ for $i \in [K+1, n]$. Observe that condition (c) of the theorem statement is immediately satisfied. For condition (a), we note that constraint (ii) of $(\mathcal{LP})$ implies that $v_i \geq v_{i+1}$ for all $i \in [n-1]$, while constraint (iii) implies that $\sum_{i=1}^{n} v_i = \sum_{i=1}^{K} s_i^* = 1$. Finally, for Condition (b) note that by constraint (i) it follows that $\sum_{i=1}^{K} v_i x_i \geq \delta^* \geq \theta$. This completes the analysis of this case.

[**Case II:** $t^* \neq 0$.] In this case, we show that t^* cannot be very close to 0. It follows from basic LP theory that the vertex $\mathbf{v}^* = (t^*, s_1^*, \ldots, s_K^*, \delta^*)$ is the unique solution of a linear system $A' \cdot \mathbf{v}^* = b'$ obtained from a subset of tight constraints in $(\mathcal{LP})$. We record the following fact:

FACT 3.1. *Consider the linear program $(\mathcal{LP})$:*
(a) All the entries of the constraint matrix A are bounded from above by $\max\{1, W_T\}$.
(b) The constant vector b has entries in $\{0, 1\}$.
(c) Any coefficient not associated with the variable t is in $\{0, 1\}$.

As mentioned above $\mathbf{v}^*$ is the unique solution of a $(K+2) \times (K+2)$ linear system $A' \cdot \mathbf{v}^* = b'$, where (A', b') is obtained from (A, b) by selecting a subset of the rows. By Cramer's rule, we have that $t^* = \det(A'_t)/\det(A')$ where A'_t is obtained by replacing the column in A' corresponding to t^* with the vector b'. Since A'_t has only $0, 1$ entries, if $\det(A'_t) \neq 0$, then $\det(A'_t) \geq 1$. Since we assumed that $t^* \neq 0$, we will indeed have that $\det(A'_t) \geq 1$. Now observe that all the columns of A' except the one corresponding to t^* have entries bounded from above by 1. The column corresponding to t has all its entries bounded from above by W_T. By Hadamard's inequality we obtain

$$|\det(A')| \leq \prod_{i=1}^{K+2} \|A'_i\|_2 \leq (K+2)^{(K+2)/2} \cdot W_T.$$

By combining the above we get

$$t^* \geq (K+2)^{-(K+2)/2} \cdot (1/W_T).$$

We are now ready to define the vector $v = (v_1, \ldots, v_n)$. We select $v_i = s_i^*$ for $i \in [K]$ and $v_i = t^* w_i$

for $i \in [K+1, n]$. It is easy to verify that v satisfies conditions (a)-(c) of the theorem. Indeed, we use the fact that $\mathbf{v}^* = (t^*, s_1^*, \ldots, s_K^*, \delta^*)$ is feasible for $(\mathcal{LP})$.

Constraint (iii) of $(\mathcal{LP})$ yields $\sum_{i=1}^n v_i = \sum_{i=1}^K s_i^* + t^* \sum_{i=K+1}^n w_i = \sum_{i=1}^K s_i^* + t^* W_T = 1$ as desired. Constraint (ii) similarly implies that $v_1 \geq v_2 \geq \ldots v_n \geq 0$, which establishes condition (a).

We now proceed to establish condition (b). Let $x \in S$. We have that

$$\begin{aligned}\sum_{i=1}^n v_i x_i - \theta &\geq \sum_{i=1}^n v_i x_i - \delta^* \\ &= \sum_{i=1}^K s_i^* x_i + t^* \left(\sum_{i=K+1}^n w_i x_i \right) - \delta^* \\ &\geq 0\end{aligned}$$

where the last inequality uses constraint (i) of $(\mathcal{LP})$.

For condition (c), since $t^* \geq (K+2)^{-(K+2)/2} \cdot (1/W_T)$, constraint (iii) of $(\mathcal{LP})$ gives

$$\sum_{i=1}^K v_i = \sum_{i=1}^K s_i^* = 1 - t^* W_T \leq 1 - (K+2)^{-(K+2)/2}.$$

Using the fact that $\sum_{i=K+1}^n v_i = t^* W_T \geq (K+2)^{-(K+2)/2}$, we conclude that

$$\sum_{i=1}^k v_i \leq (K+2)^{(K+2)/2} \cdot \sum_{i=K+1}^n v_i.$$

This completes the proof of Theorem 2.

4 Case 1: v^* is an L-junta

In this section we prove the following theorem.

THEOREM 4. *There is a (deterministic) algorithm* **Find-Optimal-Junta** *with the following performance guarantee: The algorithm takes as input a vector of probabilities $p = (p_1, \ldots, p_L)$ satisfying (A1) and (A2), a threshold value $0 < \tau < 1$, and a parameter $0 \leq W \leq 1$. It runs in $\mathrm{poly}(n, 2^{\mathrm{poly}(L)}, \mathrm{bit}(\tau))$ time and outputs a head vector $w' \in \mathbb{R}^L_{\geq 0}$ such that $\sum_{i=1}^L w_i' \leq W$. Moreover, the vector w' maximizes $\mathbf{Pr}[w \cdot X^{(H)} \geq \tau]$ over all $w \in \mathbb{R}^L_{\geq 0}$ that have $\sum_{i=1}^L w_i \leq W$.*

Note that Theorem 4 is somewhat more general than we need in order to establish the desired result in Case 1; this is because **Find-Optimal-Junta** will also be used as a component of the algorithm for Case 2. As a direct corollary of Theorem 4 we get that **Find-Optimal-Junta** finds an optimal solution in Case 1:

COROLLARY 4.1. *If v^* is an L-junta, then* **Find-Optimal-Junta**$((p_1, \ldots, p_L), \theta, 1)$ *outputs a vector $w' = (w_1', \ldots, w_L')$ such that $(w', \mathbf{0}_{n-L}) \in \mathbb{R}^n_{\geq 0}$ is an optimal solution, i.e., $\mathrm{Obj}((w', \mathbf{0}_{n-L})) = \mathsf{OPT}$.*

Algorithm Find-Optimal-Junta:

Input: vector of probabilities $(p_1, \ldots, p_L)$; threshold $0 < \tau < 1$; parameter $W > 0$

Output: vector $w' \in \mathbb{R}^L_{\geq 0}$ that maximizes $\mathbf{Pr}[w \cdot X^{(H)} \geq \tau]$ over all $w \in \mathbb{R}^L_{\geq 0}$ that have $\sum_{i=1}^L w_i \leq W$

1. Let $\mathcal{S}$ be the set of all $2^{\Theta(L^2)}$ sets $S \subseteq \{0,1\}^L$ such that $S = \{x \in \{0,1\}^L : u \cdot x \geq c\}$ for some $u \in \mathbb{R}^L, c \in \mathbb{R}$.

2. For each $S \in \mathcal{S}$, check whether the following linear program over variables $w_1, \ldots, w_L$ is feasible and if so let $w^{(S)} \in \mathbb{R}^L$ be a feasible solution:

$$\text{For all } x \in S, w \cdot x \geq \tau; \quad w_1, \cdots, w_L \geq 0;$$

$$w_1 + \cdots + w_L \leq W.$$

3. For each $w^{(S)}$ obtained in the previous step, compute $\mathbf{Pr}[w^{(S)} \cdot X^{(H)} \geq \tau]$ and output the vector $w^{(S)}$ for which this is the largest.

This case is rather simple. Procedure **Find-Optimal-Junta** outputs a vector $w' = (w_1', \ldots, w_L')$ that maximizes the desired probability over all non-negative vectors whose coordinates sum to at most W. This is done in a straightforward way, using linear programming and an exhaustive enumeration of all linear threshold functions that depend only on the first L variables.

We now proceed with the proof of Theorem 4. We first give the simple running time analysis. It is well known (see e.g., [Cho61]) that, as claimed in Step 1 of **Find-Optimal-Junta**, there are $2^{\Theta(L^2)}$ distinct Boolean functions over $\{0,1\}^L$ that can be represented as halfspaces $u \cdot x \geq c$. It is also well known (see [MTT61]) that for every $S \in \mathcal{S}$, there is a vector $u = (u_1, \ldots, u_L)$ and a threshold c such that $S = \{x \in \{0,1\}^L : u \cdot x \geq c\}$ where each u_i and c is an integer of absolute value at most $2^{\Theta(L \log L)}$. Thus it is possible to enumerate over all elements $S \in \mathcal{S}$ in $2^{\Theta(L^2 \log L)}$ time. Since for each fixed S the linear program in Step 2 has $O(2^L)$ constraints over L variables, the claimed running time bound follows.

The correctness argument is equally simple. There must be some $S \in \mathcal{S}$ which is precisely the set of those $x \in \{0,1\}^L$ that maximizes $\mathbf{Pr}_{X \sim \mu_{p_1} \times \cdots \times \mu_{p_L}}[w \cdot X \geq \tau]$ over all $w \in \mathbb{R}^L_{\geq 0}$ that have $\sum_{i=1}^L w_i \leq W$. Step 2 will identify a feasible solution for this S, and hence the vector $w' = (w_1', \ldots, w_L')$ that **Find-Optimal-Junta** outputs will achieve this maximum probability. This concludes the proof of Theorem 4.

5 Case 2: v^* is type $L+1$

Recall that in Case 2 the optimal solution v^* is not an L-junta, so it satisfies $\sum_{i=1}^{L} v_i^* \leq (L+2)^{(L+2)/2} \cdot \sum_{i=L+1}^{n} v_i^*$, and $c(v^*, \epsilon) > L$. For this case we prove the following theorem:

THEOREM 5. *There is a (deterministic) algorithm* **Find-Near-Opt-Large-CI** *with the following performance guarantee: The algorithm takes as input a vector of probabilities $p = (p_1, \ldots, p_n)$ satisfying (A1) and (A2) and a threshold value $0 < \theta < 1$. It runs in $\mathrm{poly}(n, 2^{\mathrm{poly}(L)}, \mathrm{bit}(\theta))$ time and outputs a set of $N \leq \mathrm{poly}(n, 2^{\mathrm{poly}(L)})$ many feasible solutions. If v^* is of type $L+1$ then one of the feasible solutions w' that it outputs satisfies $\mathrm{Obj}(w') \geq \mathsf{OPT} - \epsilon/2$.*

5.1 Useful probabilistic tools and notation. Anti-concentration.

We say that a real-valued random variable Z is *ϵ-anti-concentrated at radius δ* if for every interval of radius δ, Z lands in that interval with probability at most ϵ, i.e.,

$$\text{for all } t \in \mathbb{R}, \quad \mathbf{Pr}[|Z - t| \leq \delta] \leq \epsilon.$$

We will use the following simple result, which says that anti-concentration of a linear form under a product distribution can only improve by adding more independent coordinates:

LEMMA 5.1. *Fix $(q_1, \ldots, q_n) \in [0,1]^n$ and let $\bigotimes_{i=1}^{n} \mu_{q_i}$ denote the corresponding product distribution over $\{0,1\}^n$. Fix any weight-vector $w^{(k)} \in \mathbb{R}^k$ and suppose that the random variable $w^{(k)} \cdot X^{(k)}$, where $X^{(k)} \sim \bigotimes_{i=1}^{k} \mu_{q_i}$, is ϵ-anti-concentrated at radius δ. Then for any $w^{(n-k)} \in \mathbb{R}^{n-k}$, the random variable $w \cdot X$, where $w = (w^{(k)}, w^{(n-k)})$ and $X \sim \bigotimes_{i=1}^{n} \mu_{q_i}$ is also ϵ-anti-concentrated at radius δ.*

Notation. Much of our analysis in this section will deal separately with the coordinates $1, \ldots, L$ and the coordinates $L+1, \ldots, n$; hence the following terminology and notation will be convenient. For an n-dimensional vector $w \in \mathbb{R}^n$, in this section we refer to $(w_1, \ldots, w_L)$ as the "head" of w and we write $w^{(H)}$ to denote this vector; similarly we write $w^{(T)}$ to denote the "tail" $(w_{L+1}, \ldots, w_n)$ of w. We sometimes refer to a vector in $\mathbb{R}^L$ as a "head vector" and to a vector in $\mathbb{R}^{n-L}$ as a "tail vector." In a random variable $w^{(H)} \cdot X^{(H)}$ the randomness is over the draw of $X^{(H)} \sim \bigotimes_{i=1}^{L} \mu_{p_i}$, and similarly for a random variable $w^{(T)} \cdot X^{(T)}$ the randomness is over the draw of $X^{(T)} \sim \bigotimes_{i=L+1}^{n} \mu_{p_i}$.

5.2 The algorithm and its analysis.

Case 2 is more involved than Case 1. We first explain some of the analysis that motivates our approach (Lemmas 5.2 and 5.3 below) and then explain how the algorithm works (see Steps 1 and 2 of **Find-Near-Opt-Large-CI**).

Let us say that a vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ has a *κ-granular tail* if the following condition holds (throughout the rest of Section 5, $\kappa = \mathrm{poly}(1/n, 1/2^{\mathrm{poly}(L)})$; we will specify its value more precisely later):

- **[$w = (w_1, \ldots, w_n)$ has a κ-granular tail]:** For $L+1 \leq i \leq n$, each coordinate w_i is an integer multiple of κ.

The first stage of our analysis is to show (assuming that v^* is type $L+1$) that there is a feasible solution such that both the head and tail have some useful properties: the tail weights are granular and the tail random variable is sharply concentrated around its mean, while the head gives a high-quality solution to a problem with a related threshold (see condition (3) below):

LEMMA 5.2. *Suppose v^* is type $L+1$. Then there is a feasible solution $w' = (w'_1, \ldots, w'_n) \in \mathbb{R}^n_{\geq 0}$ such that $w'_1 \geq \cdots \geq w'_n \geq 0$ which satisfies the following:*

1. *The vector w' has a κ-granular tail. Hence for $M \stackrel{def}{=} \mathrm{poly}(1/\kappa)$, there are non-negative integers $A', B', C' \leq M$ such that $\sum_{i=L+1}^{n} (w'_i)^2 = A'\kappa^2$, $\sum_{i=L+1}^{n} w'_i p_i = B'\kappa(\epsilon/(4n))$, and $\sum_{i=L+1}^{n} w'_i = C'\kappa$.*

2. *Let μ' denote $\mathbf{E}[w'^{(T)} \cdot X^{(T)}]$, i.e., $\mu' = B'\kappa(\epsilon/(4n))$. The random variable $w'^{(T)} \cdot X^{(T)}$ is strongly concentrated around its mean:*
(5.2)
$\mathbf{Pr}[|w'^{(T)} \cdot X^{(T)} - \mu'| \geq \sqrt{A' \cdot \ln(200/\epsilon)} \cdot \kappa] \leq \epsilon/100.$

3. *The head random variable $w'^{(H)} \cdot X^{(H)}$ satisfies*

$$\sum_{i=1}^{L} w'_i \leq 1 - C'\kappa \quad \textit{and} \tag{5.3}$$

$$\mathbf{Pr}[w'^{(H)} \cdot X^{(H)} \geq \theta - \mu' + \sqrt{A' \cdot \ln(200/\epsilon)} \cdot \kappa] \geq \mathsf{OPT} - \epsilon/40.$$

Next, our analysis shows that for *any* vector w'' with a κ-granular tail which matches the A', B', C' values from above, the value of its overall solution is essentially determined by the value that its head random variable $w''^{(H)} \cdot X^{(H)}$ achieves for the related-threshold problem. More precisely, let us say that a triple of non-negative integers (A, B, C) with $A, B, C \leq M$ is a *conceivable* triple. We say that a conceivable triple (A, B, C) is *achievable* if there exists a vector $(u_{L+1}, \ldots, u_n) \in \mathbb{R}^{n-L}_{\geq 0}$ whose coordinates are non-negative integer multiples of κ such that $\sum_{i=L+1}^{n} (u_i)^2 = A\kappa^2$, $\sum_{i=L+1}^{n} u_i p_i = B\kappa(\epsilon/(4n))$, and $\sum_{i=L+1}^{n} u_i = C\kappa$, and we say that such a vector $(u_{L+1}, \ldots, u_n)$ *achieves* the triple (A, B, C).

LEMMA 5.3. *As above suppose that v^* is type $L+1$. Let w', A', B', C' be as described in Lemma 5.2.*

Let $w'' = (w''_1, \ldots, w''_L, w''_{L+1}, \ldots, w''_n)$ be any vector with a κ-granular tail whose $n - L$ tail coordinates $(w''_{L+1}, \ldots, w''_n)$ achieve the triple (A', B', C'). Then like $w'^{(T)} \cdot X^{(T)}$, the random variable $w''^{(T)} \cdot X^{(T)}$ is strongly concentrated around its mean:

(5.4)
$$\mathbf{Pr}[|w''^{(T)} \cdot X^{(T)} - \mu'| \geq \sqrt{A' \cdot \ln(200/\epsilon)} \cdot \kappa] \leq \epsilon/100,$$

and hence

(5.5)
$$\mathbf{Pr}[w'' \cdot X \geq \theta] \geq$$
$$\mathbf{Pr}[w''^{(H)} \cdot X^{(H)} \geq \theta - \mu' + \sqrt{A' \cdot \ln(200/\epsilon)} \cdot \kappa] - \epsilon/100.$$

Intuitively, these two lemmas are useful because they allow us to "decouple" the problem of finding an n-dimensional solution vector w into two pieces, finding a head-vector and a tail-vector. For the tail, these lemmas say that it is enough to search over the (polynomially many) conceivable triples (A, B, C); if we can identify the achievable triples from within the conceivable triples, and for each achievable triple construct *any* κ-granular tail vector that achieves it, then this is essentially as good as finding the actual tail vector of w'. For the right triple (A', B', C') given by Lemma 5.2, all that remains is to come up with a vector of head coordinates that yields a high-value solution to the related-threshold problem (note that part (3) of Lemma 5.2 establishes that indeed such a head-vector must exist). This is highly reminiscent of Case 1, and indeed we can apply machinery (the **Find-Optimal-Junta** procedure) from that case for this purpose. These lemmas thus motivate the two main steps of the algorithm, Steps 1 and 2, which we describe below.

While there are only polynomially many conceivable triples, it is a nontrivial task to identify whether any given conceivable triple is achievable (note that there are exponentially many different vectors $(u_{L+1}, \ldots, u_n)$ that might achieve a given triple). However, this does turn out to be a feasible task; Algorithm **Construct-Achievable-Tails**, called in Step 1 of **Find-Near-Opt-Large-CI**, is an efficient algorithm (based on dynamic programming) which searches across all conceivable triples (A, B, C) and identifies those which are achievable. For each triple that is found to be achievable, **Construct-Achievable-Tails** constructs a κ-granular tail which achieves it. We have the following lemma:

LEMMA 5.4. *There is a (deterministic) algorithm* **Construct-Achievable-Tails** *that outputs a list consisting precisely of all the achievable (A, B, C) triples, and for each achievable triple it outputs a corresponding tail vector $(w''_{L+1}, \ldots, w''_n)$ that achieves it. The algorithm runs in time $\mathrm{poly}(n, 1/\kappa) = \mathrm{poly}(1/\kappa)$.*

Finally, for each achievable triple (A, B, C) and corresponding tail vector $(w''_{L+1}, \ldots, w''_n)$ that is generated by **Construct-Achievable-Tails**, the procedure **Find-Optimal-Junta** is used to find a setting of the head coordinates that yields a high-quality solution.

Algorithm Find-Near-Opt-Large-CI:

Input: probability vector $p = (p_1, \ldots, p_n)$ satisfying (A1) and (A2); parameter $0 < \theta < 1$
Output: if v^* is type $L+1$, a set $\mathcal{FEAS}$ of feasible solutions w such that one of them satisfies $\mathrm{Obj}(w) \geq \mathsf{OPT} - \epsilon/2$

1. Run Algorithm **Construct-Achievable-Tails** to obtain a list $\mathcal{T}$ of all achievable triples (A, B, C) and, for each one, a tail vector $u = (u_{L+1}, \ldots, u_n)$ that achieves it.
2. For each triple (A, B, C) in $\mathcal{T}$ and its associated tail vector $u = (u_{L+1}, \ldots, u_n)$:
   - Run **Find-Optimal-Junta**$((p_1, \ldots, p_L), \theta - B\kappa\epsilon/(4n) + \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A}, 1 - C\kappa)$ to obtain a head $(u_1, \ldots, u_L)$.
   - Add the concatenated vector $(u_1, \ldots, u_L, u_{L+1}, \ldots, u_n)$ to the set $\mathcal{FEAS}$ (initially empty) of feasible solutions that will be returned.
3. Return the set $\mathcal{FEAS}$ of feasible solutions constructed as described above.

We prove the aforementioned lemmas in the next subsection. We conclude this subsection by showing how Theorem 5 follows from these lemmas.

Proof of Theorem 5 given Lemmas 5.2, 5.3, and 5.4: The claimed running time bound is immediate from inspection of **Find-Near-Opt-Large-CI**, Lemma 5.4 (to bound the running time of **Construct-Achievable-Tails**) and Theorem 4 (to bound the running time of **Find-Optimal-Junta**).

To prove correctness, suppose that v^* is of type $L+1$. One of the achievable triples that is listed by **Construct-Achievable-Tails** will be the (A', B', C') triple that is achieved by the tail $(w'_{L+1}, \ldots, w'_n)$ of the vector $w' = (w'_1, \ldots, w'_n)$ whose existence is asserted by Lemma 5.2. By Lemma 5.4, **Construct-Achievable-Tails** outputs this (A', B', C') along with a corresponding tail vector $(w''_{L+1}, \ldots, w''_n)$ that achieves it; by Lemma 5.3, any combination $u = (u_1, \ldots, u_L, w''_{L+1}, \ldots, w''_n)$ of a head vector with this tail vector will have $\mathrm{Obj}(u) \geq \mathbf{Pr}[u^{(H)} \cdot X^{(H)} \geq \theta - \mu' + \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A'}] - \epsilon/100$. Lemma 5.2 ensures that there exists some head vector $w'^{(H)}$ that has $\sum_{i=1}^{L} w'_i \leq 1 - C'\kappa$ and $\mathbf{Pr}[w'^{(H)} \cdot X^{(H)} \geq \theta - \mu' + \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A'}] \geq \mathsf{OPT} - \epsilon/40$, so when **Find-Optimal-Junta** is called with in-

put parameters $((p_1,\ldots,p_L),\theta - B'\kappa(\epsilon/(4n)) + \kappa\cdot\sqrt{\ln(200/\epsilon)\cdot A'}, 1-C'\kappa)$, by Theorem 4 it will construct a head $u^{(H)} = (u_1,\ldots,u_L)$ with $u_1,\ldots,u_L \geq 0$, $u_1 + \cdots + u_L \leq 1 - C'\kappa$ which is such that $\mathbf{Pr}[u^{(H)}\cdot X^{(H)} \geq \theta - \mu' + \kappa\cdot\sqrt{\ln(200/\epsilon)\cdot A'}] \geq \mathsf{OPT} - \epsilon/40$, and hence the resulting overall vector $u = (u_1,\ldots,u_L,w''_{L+1},\ldots,w''_n)$ is a feasible solution which has $\mathbf{Pr}[u\cdot X \geq \theta] \geq \mathsf{OPT} - 7\epsilon/200$. This concludes the proof of Theorem 5 (modulo the proofs of Lemmas 5.2, 5.3, and 5.4).

5.3 Proof of Lemmas 5.2, 5.3, and 5.4.

5.3.1 Proof of Lemma 5.2. Recall from Equation (2.1) that $L = \min\{n, \Theta(1/(\epsilon^2\gamma^2)\cdot(1/\gamma)\cdot(\log 1/(\epsilon\gamma))\cdot(\log(1/\epsilon))\}$; since we are in Case 2, we have that $L = \Theta(1/(\epsilon^2\gamma^2)\cdot(1/\gamma)\cdot(\log 1/(\epsilon\gamma))\cdot(\log(1/\epsilon))$. Since the $\epsilon\gamma/200$-critical index of v^* is at least L, Lemma 5.5 of [DGJ$^+$10] gives us that there is a subsequence of weights $v^*_{i_1},\ldots,v^*_{i_s}$ with $i_s < L$ and $s \geq t/\gamma$, where $t \stackrel{\text{def}}{=} \ln(200^2/\epsilon^3\gamma)$, such that $v^*_{i_{j+1}} \leq v^*_{i_j}/3$ for all $j = 1,\ldots,s-1$. Given this, Claim 5.7 of [DGJ$^+$10] implies that for any two points $x \neq x' \in \{0,1\}^s$, we have

$$\left|\sum_{\ell=1}^{s} v^*_{i_\ell}x_{i_\ell} - \sum_{\ell=1}^{s} v^*_{i_\ell}x'_{i_\ell}\right| \geq \frac{v^*_{i_s}}{2}. \tag{5.6}$$

(We note that both Lemma 5.5 and Claim 5.7 are simple results with proofs of a few lines.) Equation (5.6) clearly implies that for every $\nu \in \mathbb{R}$ there is at most one $x \in \{0,1\}^s$ such that $\sum_{\ell=1}^s v^*_{i_\ell}x_{i_\ell} = \nu$; recalling the definition of γ, we further have that $\mathbf{Pr}_{(X_{i_1},\ldots,X_{i_s})\sim\otimes_{j=1}^s\mu_{p_{i_j}}}\left[\sum_{\ell=1}^s v^*_{i_\ell}X_{i_\ell} = \nu\right] \leq (1-\gamma)^s$ for every $\nu \in \mathbb{R}$. Together with (5.6), this gives that for every $\nu \in \mathbb{R}$ and every integer $k \geq 0$, we have

$$\begin{aligned}\mathbf{Pr}_{(X_{i_1},\ldots,X_{i_s})\sim\otimes_{j=1}^s\mu_{p_{i_j}}}&\left[|\sum_{\ell=1}^{s} v^*_{i_\ell}X_{i_\ell} - \nu| \leq kv^*_{i_s}/2\right]\\ &\leq (2k+1)(1-\gamma)^s \leq (2k+1)e^{-t}\\ &= (2k+1)\epsilon^3\gamma/200^2.\end{aligned}$$

By independence, using Lemma 5.1 we get that this anti-concentration extends to the linear form over all of the first L coordinates, and hence we get that for all $\nu \in \mathbb{R}$,

$$\mathbf{Pr}\left[\left|(v^*)^{(H)}X^{(H)} - \nu\right| \leq kv^*_{i_s}/2\right] \leq (2k+1)\epsilon^3\gamma/200^2. \tag{5.7}$$

Now, recall that we are in Case 2 and hence $\sum_{j>L} v^*_j \geq 1/((L+2)^{(L+2)/2}+1)$. Since $v^*_{i_s} \geq v_j$ for all $j > L$, we have that $v^*_{i_s} \geq 1/(n((L+2)^{(L+2)/2}+1))$. Hence (5.7) yields that for all $\nu \in \mathbb{R}$,

$$\begin{aligned}\mathbf{Pr}\left[\left|(v^*)^{(H)}\cdot X^{(H)} - \nu\right| \leq k/(2n((L+2)^{(L+2)/2}+1))\right]&\\ \leq (2k+1)\epsilon^3\gamma/200^2.&\end{aligned} \tag{5.8}$$

We now turn from analyzing the head of v^* to analyzing the tail. Recalling again that the $\epsilon\gamma/200$-critical index of v^* is greater than L, another application of Lemma 5.5 of [DGJ$^+$10] gives that $\sigma_L^2(v^*) \stackrel{\text{def}}{=} \sum_{j>L}(v^*_j)^2 \leq 200^2(v^*_{i_s})^2/(\epsilon^2\gamma^2)$. The expected value of $(v^*)^{(T)}\cdot X^{(T)}$ is $\mu = \sum_{j>L} v^*_j p_j$; an additive Hoeffding bound gives that for $r > 0$,

$$\mathbf{Pr}[|(v^*)^{(T)}\cdot X^{(T)} - \mu| \geq r\cdot\sigma_L(v^*)] \leq 2e^{-r^2}.$$

Fixing $r = \sqrt{\ln(200/\epsilon)}$, as a consequence of the above we get that

$$\begin{aligned}\mathbf{Pr}[(v^*)^{(T)}\cdot X^{(T)} \geq \mu + \sqrt{\ln(200/\epsilon)}\cdot\sigma_L(v^*)]&\\ \leq 2e^{-\ln(200/\epsilon)} = \epsilon/100.&\end{aligned}$$

Since $\mathsf{OPT} = \mathbf{Pr}[v^*\cdot X \geq \theta]$, we get that

$$\begin{aligned}\mathbf{Pr}[(v^*)^{(H)}\cdot X^{(H)} \geq \theta - \mu - \sqrt{\ln(200/\epsilon)}\cdot\sigma_L(v^*)]&\\ \geq \mathsf{OPT} - \epsilon/100.&\end{aligned}$$

Combining with (5.7), we get that

$$\begin{aligned}\mathbf{Pr}[(v^*)^{(H)}\cdot X^{(H)} \geq \theta - \mu + \sqrt{\ln(200/\epsilon)}\cdot\sigma_L(v^*)]&\\ \geq \mathsf{OPT} - \epsilon/50.&\end{aligned} \tag{5.9}$$

We are now ready to define the vector w'. Its head coordinates are the same as v^*, i.e., for $1 \leq i \leq L$ we have $w'_i = v^*_i$. We define the quantity

$$\kappa = 1/(n^2((L+2)^{(L+2)/2}+1)).$$

For $L+1 \leq i \leq n$, the tail coordinates w'_i of w' are obtained by rounding v^*_i down to the nearest integer multiple of κ. It is immediate from this definition that part (1) of the lemma holds, i.e., w' has a κ-granular tail and there are non-negative integers $A, B, C \leq M$ as specified in part (1). Since $\sum_{i=1}^n w'_i \leq \sum_{i=1}^n v^*_i = 1$, it must be the case that $\sum_{i=1}^L w'_i \leq 1 - C'\kappa$, giving the first part of Equation (5.3).

Write μ' to denote $\mathbf{E}[w'^{(T)}\cdot X^{(T)}] = \sum_{j>L} w'_j p_j = B'\kappa(\epsilon/(4n))$. Define $\sigma_L^2(w) \stackrel{\text{def}}{=} \sum_{j>L}(w'_j)^2$. By an application of the Hoeffding bound, we get that $(w')^{(T)}\cdot X^{(T)}$ is concentrated around its mean μ'. More precisely,

$$\begin{aligned}\mathbf{Pr}[|(w')^{(T)}\cdot X^{(T)} - \mu'| \geq \sqrt{\ln(200/\epsilon)}\cdot\sigma_L(w)]&\\ \leq 2e^{-\ln(200/\epsilon)} \leq \epsilon/100,&\end{aligned}$$

giving part (2) of Lemma 5.2. Note that $\sigma_L^2(w) \leq \sigma_L^2(v^*) \leq 200^2(v^*_{i_s})^2/(\epsilon^2\gamma^2)$.

It remains only to establish the second part of Equation (5.3). Equation (5.9) almost gives us this – it falls short only in having μ in place of μ' in the lower bound for $(w')^{(H)}\cdot X^{(H)}$ (recall that $(v^*)^{(H)}$ is identical to $(w')^{(H)}$). To get around this we use the anti-concentration property of the head that was established in (5.8) above. Since

$|\mu - \mu'| \leq n\kappa = 1/(n((L+2)^{(L+2)/2}+1))$, Equation (5.8) gives that

$$\mathbf{Pr}[(w')^{(H)} \cdot X^{(H)} \in [\theta - \mu + \sqrt{\ln(200/\epsilon)} \cdot \sigma_L(w),$$

$$\theta - \mu' + \sqrt{\ln(200/\epsilon)} \cdot \sigma_L(w)]] \leq \epsilon/200$$

and combining this with (5.9) gives

$$\mathbf{Pr}[(w')^{(H)} \cdot X^{(H)} \geq \theta - \mu' + \sqrt{\ln(200/\epsilon)} \cdot \sigma_L(w)]$$

$$\geq \mathsf{OPT} - 5\epsilon/200,$$

the desired second part of Equation (5.3). This concludes the proof of Lemma 5.2.

5.3.2 Proof of Lemma 5.3. Since by assumption the tail of w'' achieves the triple (A', B', C'), we have that the mean $\mathbf{E}[(w'')^{(T)} \cdot X^{(T)}]$ equals $B'\kappa(\epsilon/(4n))$ and thus is the same as μ', the mean of $(w')^{(T)} \cdot X^{(T)}]$. Since $\sum_{j>L}(w''_j)^2 = \sum_{j>L}(w'_j)^2$, just as was the case for w' we get that a Hoeffding bound gives the desired concentration bound,

$$\mathbf{Pr}[|w''^{(T)} \cdot X^{(T)} - \mu'| \geq \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A'}] \leq \epsilon/100.$$

Thus, we have established Equation (5.4).

Equation (5.4) implies that $w''^{(T)} \cdot X^{(T)} < \mu' - \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A'}$ with probability at most $\epsilon/100$. Since $w''^{(H)} \cdot X^{(H)} \geq \theta - \mu' + \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A'}$ and $w''^{(T)} \cdot X^{(T)} \geq \mu' - \kappa \cdot \sqrt{\ln(200/\epsilon) \cdot A'}$ together imply that $w'' \cdot X \geq \theta$, we thus get Equation (5.5), and the lemma is proved.

5.3.3 Proof of Lemma 5.4. The algorithm **Construct-Achievable-Tails** is based on dynamic programming. Let $w = (w_{L+1}, \ldots, w_n)$ be a tail weight vector such that each w_i is a non-negative integer multiple of κ. We define the quantities

$$A(w) = \sum_{i>L}(w_i)^2/\kappa^2; \quad B(w) = \sum_{i>L} w_i p_i/(\kappa\epsilon/(4n));$$

$$C(w) = \sum_{i>L} w_i/\kappa.$$

Recalling Assumption (A2), we see that each of $A(w), B(w), C(w)$ is a non-negative integer bounded by $\mathrm{poly}(1/\kappa)$.

For each conceivable triple (A, B, C) and for every $t \in \{L+1, \ldots, n\}$, we create a sub-problem in which the goal is to determine whether there is a choice of weights $w_{L+1}, \ldots, w_t$ (each of which is a non-negative integer multiple of κ, with all other weights $w_{t+1}, \ldots, w_n$ set to 0) such that $A(w) = A, B(w) = B$, and $C(w) = C$. Such a choice of weights $w_{L+1}, \ldots, w_t$ exists if and only if there is a nonnegative-integer-multiple-of-κ choice of w_t for which there is a nonnegative-integer-multiple-of-κ choice of weights $w_{L+1}, \ldots, w_{t-1}$ (with all subsequent weights set to 0) such that $A(w) = A - (w_t)^2/\kappa$, $B(w) = B - w_t p_t/(\kappa\epsilon/(4n))$, and $C(w) = C - w_t/\kappa$.

Thus, given the set of all triples that are achievable with only weights $w_{L+1}, \ldots, w_{t-1}$ allowed to be nonzero, it is straightforward to efficiently (in $\mathrm{poly}(1/\kappa)$ time) identify the set of all triples that are achievable with only weights $w_{L+1}, \ldots, w_t$ allowed to be nonzero. This is because for a given candidate (conceivable) triple (A, B, C), one can check over all possible values of w_t (that are integer multiples of κ and upper bounded by 1) whether the triple $(A - (w_t)^2/\kappa, B - w_t p_t/(\kappa\epsilon/(4n)), C - w_t/\kappa)$ is achievable with only weights $w_{L+1}, \ldots, w_{t-1}$ allowed to be nonzero. Since there are only $O(1/\kappa)$ choices of the weight w_t and the overall number of sub-problems in this dynamic program is bounded by $\mathrm{poly}(n, 1/\kappa) = \mathrm{poly}(1/\kappa)$, the overall entire dynamic program runs in $\mathrm{poly}(1/\kappa)$ time. This concludes the proof of Lemma 5.4.

6 Case 3: v^* is type K for some $1 \leq K \leq L$

Recall that in Case 3 the optimal solution v^* is not an L-junta, so it satisfies $\sum_{i=1}^{L} v_i^* \leq (L+2)^{(L+2)/2} \cdot \sum_{i=L+1}^{n} v_i^*$, and $c(v^*, \epsilon) = K$ for some $1 \leq K \leq L$. For this case we prove the following theorem:

THEOREM 6. *There is a randomized algorithm* **Find-Near-Opt-Small-CI** *with the following performance guarantee: The algorithm takes as input a vector of probabilities $p = (p_1, \ldots, p_n)$ satisfying (A1) and (A2), a threshold value $0 < \theta < 1$, a value $1 \leq K \leq L$, and a confidence parameter $0 < \delta < 1$. It runs in $\mathrm{poly}(n, 2^{\mathrm{poly}(L)}, \mathrm{bit}(\theta)) \cdot \log(1/\delta)$ time and outputs a set of $N \leq \mathrm{poly}(n, 2^{\mathrm{poly}(L)})$ many feasible solutions. If v^* is of type K then with probability $1 - \delta$ one of the feasible solutions w that it outputs satisfies $\mathrm{Obj}(w) \geq \mathsf{OPT} - \epsilon/2$.*

6.1 Useful probabilistic tools and notation.

Kolmogorov distance. For X, Y two real-valued random variables we say the *Kolmogorov distance* $d_{\mathrm{K}}(X, Y)$ between X and Y is $d_{\mathrm{K}}(X, Y) \stackrel{\text{def}}{=} \sup_{t \in \mathbf{R}} |\mathbf{Pr}[X \leq t] - \mathbf{Pr}[Y \leq t]|$.

Remark. If w is an optimal solution of problem (P) and the random variables $w \cdot X, w' \cdot X$ have Kolmogorov distance at most ϵ then $\mathrm{Obj}(w') \geq \mathsf{OPT} - \epsilon$.

We recall the following useful elementary fact about Kolmogorov distance:

FACT 6.1. *Let X, Y, Z be real-valued random variables such that X is independent of Y and independent of Z. Then we have that $d_{\mathrm{K}}(X + Y, X + Z) \leq d_{\mathrm{K}}(Y, Z)$.*

The *Dvoretsky-Kiefer-Wolfowitz (DKW) inequality* is a considerably more sophisticated fact about Kolmogorov distance that will also be useful. Given m independent samples $t_1, \ldots, t_m$ drawn from a real-valued random variable X, the *empirical distribution* $\widehat{X}_m$ is defined as

the real-valued random variable which is uniform over the multiset $\{t_1, \ldots, t_m\}$. The DKW inequality states that for $m = \Omega((1/\epsilon^2) \cdot \ln(1/\delta))$, with probability $1 - \delta$ the empirical distribution $\widehat{X}_m$ will be ϵ-close to p in Kolmogorov distance:

THEOREM 7. ([DKW56, MAS90]) *For all $\epsilon > 0$ and any real-valued random variable X, we have $\mathbf{Pr}[d_{\mathrm{K}}(p, \widehat{p}_m) > \epsilon] \leq 2e^{-2m\epsilon^2}$.*

We will also require a corollary of the Berry-Esséen theorem (see e.g., [Fel68]). We begin by recalling the theorem:

THEOREM 8. *(Berry-Esséen) Let $X_1, \ldots, X_n$ be a sequence of independent random variables satisfying $\mathbf{E}[X_i] = 0$ for all i, $\sum_i \mathbf{E}[X_i^2] = \sigma^2$, and $\sum_i \mathbf{E}[|X_i|^3] = \rho_3$. Let $S = X_1 + \cdots + X_n$ and let F denote the cumulative distribution function (cdf) of S. Then*

$$\sup_x |F(x) - \Phi_\sigma(x)| \leq C\rho_3/\sigma^3,$$

where Φ_σ is the cdf of a $N(0, \sigma^2)$ Gaussian random variable (with mean zero and variance σ^2), and C is a universal constant. [Shi86] has shown that one can take $C = .7915$.

COROLLARY 6.1. *Let $X = (X_1, \ldots, X_n) \sim \mathcal{D}_p$ and suppose that $\min_{i \in [n]}\{p_i, 1 - p_i\} \geq \gamma > 0$. Let $w \in \mathbb{R}^n$ be τ-regular. Let Z be the random variable $w \cdot X$ and define $\mu = \mathbf{E}[w \cdot X] = \sum_{i=1}^n w_i p_i$, $\sigma^2 = \mathrm{Var}[w \cdot X] = \sum_{i=1}^n w_i^2 \cdot p_i(1 - p_i)$. Then $d_{\mathrm{K}}(Z, N(\mu, \sigma^2)) \leq \eta$ where $\eta = \tau/\gamma$.*

Proof. Define the random variable $Y_i = w_i(X_i - p_i)$, so $\mathbf{E}[Y_i] = 0$. It suffices to show that the random variable $Y = \sum_{i=1}^n w_i Y_i$ has $d_{\mathrm{K}}(Y, N(0, \sigma^2))$. We have $\sum_i \mathbf{E}[Y_i^2] = \sigma^2 = \sum_{i=1}^n w_i^2 p_i(1 - p_i)$ and

$$\begin{aligned} \mathbf{E}[|y_i|^3] &= w_i^3 \left(p_i \cdot (1 - p_i)^3 + (1 - p_i) \cdot (p_i)^3\right) \\ &= w_i^3 p_i(1 - p_i) \cdot (p_i^2 + (1 - p_i)^2), \text{ so} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n \mathbf{E}[|Y_i|^3] &= \sum_{i=1}^n w_i^3 p_i(1 - p_i)(p_i^2 + (1 - p_i)^2) \\ &\leq \sum_{i=1}^n w_i^3 p_i(1 - p_i). \end{aligned}$$

The Berry-Esséen theorem thus gives

$$\begin{aligned} d_{\mathrm{K}}(Y, N(0, \sigma^2)) &\leq \frac{\sum_{i=1}^n w_i^3 p_i(1 - p_i)}{\left(\sum_{i=1}^n w_i^2 p_i(1 - p_i)\right)^{3/2}} \\ &\leq \max_{i=1}^n |w_i| \cdot \frac{\sum_{i=1}^n w_i^2 p_i(1 - p_i)}{\left(\sum_{i=1}^n w_i^2 p_i(1 - p_i)\right)^{3/2}} \\ &= \max_{i=1}^n |w_i| \cdot \frac{1}{\sigma}. \end{aligned}$$

Recalling that (by regularity) we have $\max_i w_i \leq \tau\sqrt{\sum_i w_i^2}$, and that by definition of γ and σ we have $\sigma \geq \gamma\sqrt{\sum_i w_i^2}$, we get that $\max_{i=1}^n |w_i| \cdot \frac{1}{\sigma} \leq \tau/\gamma$ as desired.

Finally, we recall the well-known fact that an $N(\mu, \sigma^2)$ Gaussian is ϵ-anti-concentrated at radius $\epsilon\sigma$ (this follows directly from the fact that the pdf of an $N(\mu, \sigma^2)$ Gaussian is given by $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$).

Notation. In this section our analysis will deal separately with the coordinates $1, \ldots, K$ and the coordinates $K + 1, \ldots, n$, so we use the following notational conventions. For an n-dimensional vector $w \in \mathbb{R}^n$, in this section we refer to $(w_1, \ldots, w_{K-1})$ as the "head" of w and we write $w^{(H)}$ to denote this vector; similarly we write $w^{(T)}$ to denote the "tail" $(w_K, \ldots, w_n)$ of w. We sometimes refer to a vector in $\mathbb{R}^{K-1}$ as a "head vector" and to a vector in $\mathbb{R}^{n-K+1}$ as a "tail vector." In a random variable $w^{(H)} \cdot X^{(H)}$ the randomness is over the draw of $X^{(H)} \sim \bigotimes_{i=1}^{K-1} \mu_{p_i}$, and similarly for the random variable $w^{(T)} \cdot X^{(T)}$ the randomness is over the draw of $X^{(T)} \sim \bigotimes_{i=K}^{n} \mu_{p_i}$.

We additionally modify some of the terminology from Section 5 dealing with granular vectors and achievable triples. Fix $\kappa = \mathrm{poly}(1/n, 1/2^{\mathrm{poly}(L)})$ (we give a more precise value of κ later). We say that a vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ has a *κ-granular tail* if each coordinate w_i, $K \leq i \leq n$, is an integer multiple of κ. It is easy to see that for any vector $w \in \mathbb{R}^n_{\leq 0}$ with $\sum_{i=1}^n w_i \leq 1$ that has a κ-granular tail, for $M \stackrel{\text{def}}{=} \mathrm{poly}(1/\kappa)$ there must exist non-negative integers $A, B, C \leq M$ such that $\mathbf{E}[w^{(T)} \cdot X^{(T)}] = \sum_{i=K}^n w_i p_i = A\kappa(\epsilon/(4n))$, $\mathrm{Var}[w^{(T)} \cdot X^{(T)}] = \sum_{i=K}^n w_i^2 p_i(1 - p_i) = B\kappa^2(\epsilon/(4n))^2$, and $\sum_{i=K}^n w_i = C'\kappa$. We say that a triple of non-negative integers (A, B, C) with $A, B, C \leq M$ is a *conceivable* triple. We say that a conceivable triple (A, B, C) is *ϵ'-regular achievable* if there exists an ϵ'-regular vector $u^{(T)} = (u_{K+1}, \ldots, u_n) \in \mathbb{R}^{n-K+1}_{\geq 0}$ whose coordinates are non-negative integer multiples of κ such that $\mathbf{E}[u^{(T)} \cdot X^{(T)}] = A\kappa(\epsilon/(4n))$, $\mathrm{Var}[u^{(T)} \cdot X^{(T)}] = B\kappa^2(\epsilon/(4n))^2$, and $\sum_{i=K}^n u_i = C\kappa$, and we say that such a vector $(u_{L+1}, \ldots, u_n)$ *achieves* the triple (A, B, C).

Algorithm Find-Near-Opt-Small-CI:

Input: probability vector $p = (p_1, \ldots, p_n)$ satisfying (A1) and (A2); parameter $0 < \theta < 1$; parameter $1 \leq K \leq L$; confidence parameter $0 < \delta < 1$
Output: if v^* is type K, a set $\mathcal{FEAS}$ of feasible solutions w such that one of them satisfies $\mathrm{Obj}(w) \geq \mathsf{OPT} - \epsilon/2$

1. Run Algorithm **Construct-Achievable-Regular-Tails**$(\epsilon\gamma/100)$ to obtain a list $\mathcal{T}$

of all triples (A, B, C) that are achieved by some $\epsilon\gamma/100$-regular tail vector and, and, for each one, an $\epsilon\gamma/100$-regular tail vector $u = (u_{L+1}, \ldots, u_n)$ that achieves it.

2. For each triple (A, B, C) in $\mathcal{T}$ and its associated tail vector $u = (u_K, \ldots, u_n)$,
   - Run **Find-Approximately-Best-Head** $(u_K, \ldots, u_n, \epsilon/200, \delta/(2|\mathcal{T}|))$ to obtain a head vector $(u_1, \ldots, u_{K-1})$
   - Add the concatenated vector $(u_1, \ldots, u_{K-1}, u_K, \ldots, u_n)$ to the set $\mathcal{FEAS}$ (initially empty) of feasible solutions that will be returned.
3. Return the set $\mathcal{FEAS}$ of feasible solutions constructed as described above.

6.2 The algorithm and an intuitive explanation of its performance. Similar to Case 2, the high level idea of this case is to decouple the problem of finding a good solution into two pieces, namely finding a good tail and finding a good head. However, in Case 2 the anti-concentration of the head random variable (see Equation (5.8)) played an essential role; in contrast, here in Case 3 the fact that the tail random variable is close to a Gaussian will play the key role. At a high level, the analysis for this case proceeds as follows.

First, using the facts that the vector $(v_K^*, \ldots, v_n^*)$ is $\epsilon\gamma/200$-regular and that $\sum_{i=1}^{L} v_i^* \leq (L+2)^{(L+2)/2} \cdot \sum_{i=L+1}^{n} v_i^*$, we get that the tail random variable $(v^*)^{(T)} \cdot X^{(T)}$ is $O(\epsilon)$-close to a Gaussian $N(\mu, \sigma^2)$ in Kolmogorov distance, where the variance σ^2 is "not too small" (see Lemma 6.1). Next, we argue that for any head vector $(w')^{(H)} = (w'_1, \ldots, w'_{K-1})$, there exists a tail vector $(w')^{(T)} = (w'_K, \ldots, w'_n)$, obtained by rounding the tail coordinates $v_K^*, \ldots, v_n^*$ down to some not-too-small granularity κ, which is "nice" (i.e., regular and with not-too-small variance) and which gives a solution of almost equal quality to what would be obtained by having the actual $(v_K^*, \ldots, v_n^*)$ as the tail weights (see Lemma 6.2). We then strengthen this by showing that for any head vector, *any* tail vector which is regular and has the right mean and variance similarly gives a solution of almost equal quality to what would be obtained by having the actual $(v_K^*, \ldots, v_n^*)$ as the tail weights (see Lemma 6.3). This motivates the **Construct-Achievable-Regular-Tails** procedure (called in Step 1); it uses dynamic programming to efficiently search across all conceivable triples and identify precisely those that are achieved by some $\epsilon\gamma/100$-regular κ-granular tail vector (and for each achievable triple, identify a tail vector $(u_K, \ldots, u_n)$ that achieves it).

Intuitively, at this point the algorithm has identified a polynomial-sized collection of tail vectors one of which "is good" (does almost as well as the optimal tail vector $(v_K^*, \ldots, v_n^*)$ if it were paired with the optimal head vector). It remains to show that it is possible to find a high-quality head vector and that combining such a head vector with this "good" tail vector yields an overall high-quality solution. We do this, and conclude the proof of Theorem 6, in Section 6.4.

6.3 Good tails exist and can be found efficiently: Proofs of Lemmas 6.1 – 6.3 and analysis of Construct-Achievable-Regular-Tails. Let

$$\mu \stackrel{\text{def}}{=} \mathbf{E}[(v^*)^{(T)} \cdot X^{(T)}] = \sum_{i=K}^{n} v_i^* p_i \quad \text{and} \tag{6.10}$$

$$\sigma^2 \stackrel{\text{def}}{=} \mathrm{Var}[(v^*)^{(T)} \cdot X^{(T)}] = \sum_{i=K}^{n} (v_i^*)^2 p_i (1 - p_i).$$

LEMMA 6.1. *Suppose v^* is type K. Then $d_{\mathrm{K}}((v^*)^{(T)} \cdot X^{(T)}, N(\mu, \sigma^2)) \leq \epsilon/200$, and $\sigma \geq \frac{\gamma}{((L+2)^{(L+2)/2}+1)n}$.*

Proof. Since v^* is type K, we have that $(v^*)^{(T)}$ is $\epsilon\gamma/200$-regular, and hence Corollary 6.1 gives that $d_{\mathrm{K}}((v^*)^{(T)} \cdot X^{(T)}, N(\mu, \sigma^2)) \leq \epsilon/200$.

For the lower bound on σ, we observe that since $K \leq L$, $\sum_{i=1}^{L} v_i^* \leq (L+2)^{(L+2)/2} \sum_{i=L+1}^{n} v_i^*$, and $\sum_{i=1}^{n} v_i^* = 1$, we have

$$v_K^* + \cdots + v_n^* \geq v_{L+1}^* + \cdots + v_n^* \geq \frac{1}{((L+2)^{(L+2)/2}+1)}.$$

Hence Cauchy-Schwarz implies that

$$\begin{aligned} \sqrt{\sum_{i=K}^{n} (v_i^*)^2} &\geq \frac{1}{((L+2)^{(L+2)/2}+1)(n-K)} \\ &\geq \frac{1}{((L+2)^{(L+2)/2}+1)n} \end{aligned}$$

so

$$\sigma = \sqrt{\sum_{i=K}^{n} (v_i^*)^2 p_i (1-p_i)} \geq \frac{\gamma}{((L+2)^{(L+2)/2}+1)n}.$$

We now define the value of κ to be

$$\kappa = \frac{\epsilon\gamma^2}{200((L+2)^{(L+2)/2}+1)^2 n^3}.$$

LEMMA 6.2. *As above suppose v^* is type K. Let $w' \in \mathbb{R}^n_{\geq 0}$ be a feasible solution which is such that for $K \leq i \leq n$, the value w'_i is obtained from v_i^* by rounding down to the nearest integer multiple of κ. Then*

1. *The vector $(w')^{(T)} = (w'_K, \ldots, w'_n)$ is $\epsilon\gamma/100$-regular;*

2. *The variance* $(\sigma')^2 \stackrel{def}{=} \mathrm{Var}[(w')^{(T)} \cdot X^{(T)}]$ *is at least* $\frac{1}{2}\sigma^2 \geq \frac{1}{2} \cdot \frac{\gamma^2}{((L+2)^{(L+2)/2}+1)^2 n^2}$*; and*

3. $\mathrm{Obj}(w') \geq \mathrm{Obj}(w'_1, \ldots, w'_{K-1}, v^*_K, \ldots, v^*_n) - \epsilon/40$.

Proof. We start by lower bounding $(\sigma')^2$ as follows. Since each w'_i, $K \leq i \leq n$, is less than v^*_i by at most κ, we have that $\sum_{i=K}^n (w'_i)^2$ is less than $\sum_{i=K}^n (v^*_i)^2$ by at most $2\kappa n$ and hence

$$\begin{aligned}
\sigma^2 - (\sigma')^2 &\leq 2\kappa n \cdot \max_{i=K}^{n} p_i(1-p_i) \\
&\leq \frac{\kappa n}{2} \\
&< \frac{1}{2} \cdot \frac{\gamma^2}{((L+2)^{(L+2)/2}+1)^2 n^2} \\
&\leq \frac{1}{2} \cdot \sigma^2
\end{aligned}$$

so $(\sigma')^2 \geq \frac{1}{2}\sigma^2$, giving (2). Part (1) follows easily from (2) and the fact that $w'_i \leq v^*_i$ for $K \leq i \leq n$.

For part (3) we use the fact that the tail $w'^{(T)} \cdot X^{(T)}$ is anti-concentrated (since, by regularity, it is close to a Gaussian). In more detail, fix an outcome $(y_1, \ldots, y_{K-1}) \in \{0,1\}^{K-1}$ for the head bits. Since $\sum_{i=K}^n w'_i y_i \geq \sum_{i=K}^n v^*_i y_i - \kappa n$ for all $(y_K, \ldots, y_n) \in \{0,1\}^{n-k+1}$, we have

$$\begin{aligned}
&\mathbf{Pr}\left[\sum_{j=1}^{K-1} w'_j y_j + (v^*_i)^{(T)} \cdot X^{(T)} \geq \theta\right] - \\
&\mathbf{Pr}\left[\sum_{j=1}^{K-1} w'_j y_j + (w')^{(T)} \cdot X^{(T)} \geq \theta\right] \leq \\
&\mathbf{Pr}\Big[(w')^{(T)} \cdot X^{(T)} \in [\theta - \sum_{j=1}^{K-1} w'_j y_j - \kappa n, \\
&\qquad \theta - \sum_{j=1}^{K-1} w'_j y_j]\Big]. \qquad (6.11)
\end{aligned}$$

Since by (1) we know that $(w')^{(T)}$ is $\epsilon\gamma/100$-regular, Corollary 6.1 gives us that

$$\begin{aligned}
&d_{\mathrm{K}}\Big((w')^{(T)} \cdot X^{(T)}, N\big(\mathbf{E}[(w')^{(T)} \cdot X^{(T)}], (\sigma')^2\big)\Big) \\
&\qquad \leq \epsilon/100.
\end{aligned}$$

Since $\kappa n/2 \leq \epsilon\sigma'/200$, as noted after Lemma 5.1 a random variable $Z \sim N(\mathbf{E}[(w')^{(T)} \cdot X^{(T)}], (\sigma')^2)$ has $\mathbf{Pr}[Z \in I] \leq \epsilon/200$ for any interval I of length κn. Hence (6.12) is at most $\frac{\epsilon}{100} + \frac{\epsilon}{100} + \frac{\epsilon}{200} = \frac{\epsilon}{40}$. Since this holds for each fixed $(y_1, \ldots, y_{K-1}) \in \{0,1\}^{K-1}$, we get (3).

LEMMA 6.3. *As above suppose v^* is type K. Fix $(w'')^{(T)} = (w''_K, \ldots, w''_n) \in \mathbb{R}^{n-K+1}_{\geq 0}$ to be* any *$\epsilon\gamma/100$-regular tail vector such that $\mu'' \stackrel{def}{=} \mathbf{E}[(w'')^{(T)} \cdot X^{(T)}]$ equals $\mu' \stackrel{def}{=} \mathbf{E}[(w')^{(T)} \cdot X^{(T)}]$, and $(\sigma'')^2 \stackrel{def}{=} \mathrm{Var}[(w'')^{(T)} \cdot X^{(T)}]$ equals $(\sigma')^2$ (see part (2) of Lemma 6.2). Then for any head vector $(w'')^{(H)} = (w''_1, \ldots, w''_{K-1})$, we have that $\mathrm{Obj}((w''_1, \ldots, w''_{K-1}, w''_K, \ldots, w''_n)) \geq \mathrm{Obj}(w''_1, \ldots, w''_{K-1}, v^*_K, \ldots, v^*_n) - \epsilon/40$.*

Proof. The proof is identical to part (3) of Lemma 6.2.

Having established the existence of a "good" tail (the vector $(w')^{(T)}$ from Lemma 6.2), we now argue that **Construct-Achievable-Regular-Tails** can efficiently construct a list containing some such good tail vector. Lemma 6.3 ensures that finding any such good tail vector is as good as finding the actual tail vector $(w')^{(T)}$ obtained from $(v^*)^{(T)}$ by rounding down as described in Lemma 6.2.

LEMMA 6.4. *There is a (deterministic) algorithm* **Construct-Achievable-Regular-Tails**(ϵ') *that, given input parameters ϵ' and K, outputs a list consisting precisely of all the ϵ'-regular achievable (A, B, C) triples, and for each achievable triple it outputs a corresponding tail vector $(w''_K, \ldots, w''_n)$ that achieves it. The algorithm runs in time* $\mathrm{poly}(n, 1/\kappa) = \mathrm{poly}(1/\kappa)$.

Proof. Similar to the earlier **Construct-Achievable-Tails** algorithm, the main idea is to use dynamic programming; however the details are somewhat different, chiefly because of the need to ensure regularity (and also because the numerical quantities involved are somewhat different from before).

Let $w = (w_K, \ldots, w_n)$ be a tail weight vector such that each w_i is a non-negative integer multiple of κ. We define the quantities

$$\begin{aligned}
A(w) &= \sum_{i=K}^{n} w_i p_i / (\kappa\epsilon/(4n)); \\
B(w) &= \sum_{i=K}^{n} w_i^2 p_i(1-p_i) / (\kappa^2(\epsilon/(4n))^2); \\
C(w) &= \sum_{i=K}^{n} w_i/\kappa; \quad D(w) = \sum_{i=K}^{n} w_i^2/\kappa^2; \\
E(w) &= \max_{i=K}^{n} w_i/\kappa.
\end{aligned}$$

Recalling Assumption (A2), we see that each of $A(w), B(w), C(w), D(w), E(w)$ is a non-negative integer. We say that a quintuple (A, B, C, D, E) is *conceivable* if all values are non-negative integers at most M.

For each conceivable quintuple (A, B, C, D, E) and for every $t \in \{K, \ldots, n\}$, we create a sub-problem in which the goal is to determine whether there is a choice of weights $w_K, \ldots, w_t$ (each of which is a non-negative integer multiple of κ, with all other weights $w_{t+1}, \ldots, w_n$ set to 0) such that $A(w) = A$, $B(w) = B$,

$C(w) = C$, $D(w) = D$ and $E(w) = E$. Such a choice of weights $w_K, \ldots, w_t$ exists if and only if there is a nonnegative-integer-multiple-of-κ choice of w_t for which there is a nonnegative-integer-multiple-of-κ choice of weights $w_K, \ldots, w_{t-1}$ (with all subsequent weights set to 0) such that $A(w) = A - w_t p_t/(\kappa\epsilon/(4n))$, $B(w) = B - w_t^2 p_t(1-p_t)/(\kappa^2(\epsilon/(4n))^2)$, $C(w) = C - w_t/\kappa$, $D(w) = D - w_t^2/\kappa^2$, and $E = \max\{E(w), w_t/\kappa\}$.

Thus, given the set of all quintuples that are achievable with only weights $w_K, \ldots, w_{t-1}$ allowed to be nonzero, it is straightforward to efficiently (in $\mathrm{poly}(1/\kappa)$ time) identify the set of all quintuples that are achievable with only weights $w_K, \ldots, w_t$ allowed to be nonzero. Since there are only $O(1/\kappa)$ choices of the weight w_t and the overall number of sub-problems in this dynamic program is bounded by $\mathrm{poly}(n, 1/\kappa) = \mathrm{poly}(1/\kappa)$, the overall entire dynamic program runs in $\mathrm{poly}(1/\kappa)$ time.

Once the set of all achievable quintuples has been obtained, it is straightforward for each quintuple (A, B, C, D, E) to determine whether or not it is ϵ'-regular (by computing $E/\sqrt{D}$ and comparing against ϵ'). Having identified the set of all ϵ'-regular quintuples, it is easy to output a list consisting of all the ϵ'-regular achievable (A, B, C) triples (and from the dynamic program it is easy to maintain a tail vector achieving the triple in the usual way). This concludes the proof of Lemma 6.4.

6.4 Finding a good head vector: The Find-Approximately-Best-Head procedure and the proof of Theorem 6. By Lemma 6.4 the **Construct-Achievable-Regular-Tails** procedure generates a tail vector $(w'')^{(T)}$ that matches the mean, variance and L_1-norm of the $(w')^{(T)}$ vector whose existence is asserted by Lemma 6.2. In the rest of this section we consider the execution of **Find-Approximately-Best-Head** when it is run on this tail vector $(w'')^{(T)}$ as input.

By the DKW inequality (Theorem 7), with high probability the random variable R has $d_{\mathrm{K}}(R, (w'')^{(T)} \cdot X^{(T)}) \leq \epsilon/200$; we henceforth assume that this is indeed the case. Fact 6.1 implies that $d_{\mathrm{K}}((v^*)^{(H)} \cdot X^{(H)} + R, (v^*)^{(H)} \cdot X^{(H)} + (w'')^{(T)} \cdot X^{(T)}) \leq \epsilon/200$. Since $\mathrm{Obj}(v_1^*, \ldots, v_{K-1}^*, w_K'', \ldots, w_n'') \geq \mathsf{OPT} - \epsilon/40$ by Lemma 6.3, we get that $\mathbf{Pr}[(v^*)^{(H)} \cdot X^{(H)} + R \geq \theta] \geq \mathsf{OPT} - 6\epsilon/200$.

By Lemma 6.5, the **Find-Best-Head** procedure returns a head vector $u^{(H)} = (u_1, \ldots, u_{K-1})$ such that $\mathbf{Pr}[u^{(H)} \cdot X^{(H)} + R \geq \theta] \geq \mathbf{Pr}[(v^*)^{(H)} + R \geq \theta]$, so $\mathbf{Pr}[u^{(H)} \cdot X^{(H)} + R \geq \theta] \geq \mathsf{OPT} - 6\epsilon/200$. Now recalling that $d_{\mathrm{K}}(R, (w'')^{(T)} \cdot X^{(T)}) \leq \epsilon/200$, applying Fact 6.1 again gives us that $d_{\mathrm{K}}(u^{(H)} \cdot X^{(H)} + R, u^{(H)} \cdot X^{(H)} + (w'')^{(T)} \cdot X^{(T)}) \leq \epsilon/200$. Hence it must be the case that $\mathbf{Pr}[u^{(H)} \cdot X^{(H)} + (w'')^{(T)} \cdot X^{(T)} \geq \theta] \geq \mathsf{OPT} - 7\epsilon/200$. Since $u_1 + \cdots + u_{k-1} + w_k'' + \cdots + w_n'' \leq 1$ by Lemma 6.5, this vector is a near-optimal feasible solution. This concludes the proof of Theorem 6, modulo the proof of Lemma 6.5.

Algorithm Find-Approximately-Best-Head:

Input: vector of tail weights $(u_K, \ldots, u_n)$ with $u_K + \cdots + u_n \leq 1$; parameters ϵ', δ'

Output: if v^* is type K, with probability $1 - \delta'$ a head vector such that $\mathbf{Pr}[u \cdot X \geq \theta] \geq \mathbf{Pr}[(u_1', \ldots, u_{K-1}', u_K, \ldots, u_n) \cdot X \geq \theta] - \epsilon'$ for all $(u_1', \ldots, u_{K-1}') \in \mathbb{R}_{\geq 0}^{K-1}$ such that $u_1' + \cdots + u_{K-1}' + u_k + \cdots + u_n \leq 1$

1. Sample $m = \Theta(\log(1/\delta')/(\epsilon')^2)$ points $t_1, \ldots, t_m$ from the random variable $(u_K, \ldots, u_n) \cdot X^{(T)}$. Let R be the random variable which is uniform over the multiset $\{t_1, \ldots, t_m\}$.
2. Run Algorithm **Find-Best-Head**$(t_1, \ldots, t_m, 1 - \sum_{j=K}^n u_j, K)$ and return the head vector $(u_1, \ldots, u_{K-1})$ that it returns.

Algorithm Find-Best-Head:

Input: points $t_1, \ldots, t_m$, weight value $0 \leq W \leq 1$, parameter K

Output: Returns the non-negative head vector $u^{(H)} = (u_1, \ldots, u_{K-1})$ that maximizes $\mathbf{Pr}[u^{(H)} \cdot X^{(H)} + R \geq \theta]$ subject to $u_1 + \cdots + u_{K-1} \leq W$, where R is the random variable that is uniform over multiset $\{t_1, \ldots, t_m\}$

1. Let $\mathcal{S}$ be the set of all $2^{\Theta(K^2)}$ sets $S \subseteq \{0,1\}^{K-1}$ such that $S = \{x \in \{0,1\}^{K-1} : u \cdot x \geq c\}$ for some $u \in \mathbb{R}^{K-1}, c \in \mathbb{R}$.
2. For each $S = (S_1, \ldots, S_m) \in \mathcal{S}^m$, check whether the following linear program over variables $w_1, \ldots, w_{K-1}$ is feasible and if so let $w^{(S)} \in \mathbb{R}^L$ be a feasible solution:
   (a) For each $i \in [m]$ and each $x \in S_i$, $w \cdot x + t_i \geq \theta$;
   (b) $w_1, \cdots, w_{K-1} \geq 0$;
   (c) $w_1 + \cdots + w_{K-1} \leq W$.
3. For each $w^{(S)}$ obtained in the previous step, compute $\mathbf{Pr}[w^{(S)} \cdot X^{(H)} + R \geq \theta]$ and output the vector $w^{(S)}$ for which this is the largest.

LEMMA 6.5. *The (deterministic) algorithm* **Find-Best-Head** *runs in time* $2^{\mathrm{poly}(m,K)}$ *and outputs a vector* $u^{(H)} = (u_1, \ldots, u_{K-1}) \in \mathbb{R}_{\geq 0}^{K-1}$ *with* $\|u^{(H)}\|_1 \leq W$ *which is such that for every* $(u')^{(H)} \in \mathbb{R}_{\geq 0}^{K-1}$ *with* $\|(u')^{(H)}\|_1 \leq W$, *we have* $\mathbf{Pr}[u^{(H)} \cdot X^{(H)} + R] \geq$

$\mathbf{Pr}[(u') \cdot X^{(H)} + R]$.

Proof. The claimed running time bound follows easily from the fact that $|\mathcal{S}| = 2^{\Theta(mK^2)}$ (note that the running time of the linear program and the time required to explicitly compute the probabilities in Step 3 are both dominated by the enumeration over all elements of $\mathcal{S}^m$.).

The correctness argument is similar to the proof of Theorem 4. As in that proof, $\mathcal{S}$ consists of all possible sets of satisfying assignments to a $(K-1)$-variable halfspace. The optimal head vector that maximizes $\mathbf{Pr}[u^{(H)} \cdot X^{(H)} + R \geq \theta]$ subject to $u_1 + \cdots + u_{K-1} \leq W$ must be such that there is some $S = (S_1, \ldots, S_m) \in \mathcal{S}^m$ such that for $1 \leq i \leq m$, S_i is precisely the set of those $x \in \{0,1\}^L$ for which $u^{(H)} \cdot x + t_i \geq \theta$. By searching over all $S = (S_1, \ldots, S_m) \in \mathcal{S}^m$ in Step 2, the algorithm will encounter this S and will construct a feasible head vector for it. Such a feasible head vector will be identified as maximizing the probability in Step 3, and hence **Find-Best-Head** will indeed output an optimal head vector as claimed. This concludes the proof of Theorem 4.

7 Putting it together: proof of Theorem 3

In this section we prove Theorem 3 using Theorems 4, 5 and 6.

The overall algorithm works as follows. First, it runs **Find-Optimal-Junta**$((p_1, \ldots, p_L), \theta, 1)$ to obtain a feasible solution w^{junta}. Next, for each $K = 1, \ldots, L$ it runs Algorithm **Find-Near-Opt-Small-CI**$((p_1, \ldots, p_n), \theta, K, \delta/(2L))$ to obtain a set $\mathcal{FEAS}^{(K)}$ of feasible solutions. Finally, it runs Algorithm **Find-Near-Opt-Large-CI**$((p_1, \ldots, p_n), \theta)$ to obtain a final set $\mathcal{FEAS}^{(L+1)}$ of feasible solutions. It is easy to see from Theorems 4, 5 and 6 that the running time of the overall algorithm is as claimed.

Let $\mathcal{ALL}$ denote the union of the sets $\{w^{\mathrm{junta}}\}$, $\mathcal{FEAS}^{(1)}$, $\ldots$, $\mathcal{FEAS}^{(L)}$ and $\mathcal{FEAS}^{(L+1)}$. Since v^* must fall in either Case 1, Case 2 or Case 3, Theorems 4, 5 and 6 together guarantee that $\mathcal{ALL}$ is a set of $\mathrm{poly}(n, 2^{\mathrm{poly}(L)})$ many feasible solutions that with probability at least $1 - \delta/2$ contains a feasible solution w with $\mathrm{Obj}(w) \geq \mathsf{OPT} - \epsilon/2$.

Next, we sample $m = \Theta((1/\epsilon)^2 \cdot (\log |\mathcal{ALL}|/\delta))$ points independently from $\mathcal{D}_p$. For each feasible solution $w \in \mathcal{ALL}$ we use these m points to obtain an empirical estimate $\widetilde{\mathrm{Obj}}(w)$ of $\mathrm{Obj}(w)$ (recall that $\mathrm{Obj}(w) = \mathbf{Pr}_{X \sim \mathcal{D}_p}[w \cdot X \geq \theta]$), i.e., we set $\widetilde{\mathrm{Obj}}(w)$ to be the fraction of the m points that satisfy $w \cdot X \geq \theta$. A straightforward Chernoff bound implies that with probability at least $1 - \delta/2$, for each w we have $|\widetilde{\mathrm{Obj}}(w) - \mathrm{Obj}(w)| \leq \epsilon/4$.

Finally, we output the vector $w^* \in \mathcal{ALL}$ that maximizes $\widetilde{\mathrm{Obj}}(w)$ (breaking ties arbitrarily), together with the value $\widetilde{\mathrm{Obj}}(w)$. With overall probability at least $1 - \delta$ this w^* has $\mathrm{Obj}(w^*) \geq \mathsf{OPT} - 3\epsilon/4$ and $|\widetilde{\mathrm{Obj}}(w) - \mathsf{OPT}| \leq \epsilon$ as desired. This proves Theorem 3.

References

[AFH+12] N. Alon, P. Frankl, H. Huang, V. Rödl, A. Rucinski, and B. Sudakov. Large Matchings in Uniform Hypergraphs and the Conjectures of Erdös and Samuels. *Journal of Combinatorial Theory, Series A*, 2012.

[Bea55] E. M. L. Beale. On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):pp. 173–184, 1955.

[BGK11] Anand Bhalgat, Ashish Goel, and Sanjeev Khanna. Improved approximation results for stochastic knapsack problems. In *SODA*, pages 1647–1665, 2011.

[BL97] John R. Birge and François Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer, 1997.

[BV04] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[CC62] A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, 1962.

[Cho61] C.K. Chow. On the characterization of threshold functions. In *Proceedings of the Symposium on Switching Circuit Theory and Logical Design (FOCS)*, pages 34–38, 1961.

[CK05] Chandra Chekuri and Sanjeev Khanna. A polynomial time approximation scheme for the multiple knapsack problem. *SIAM J. Comput.*, 35(3):713–728, 2005.

[Dan55] George B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3/4):pp. 197–206, 1955.

[DDFS12] A. De, I. Diakonikolas, V. Feldman, and R. Servedio. Near-optimal solutions for the Chow Parameters Problem and low-weight approximation of halfspaces. In *Proc. 44th ACM Symposium on Theory of Computing (STOC)*, pages 729–746, 2012.

[DGJ+10] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM J. on Comput.*, 39(8):3441–3462, 2010.

[DGV08] Brian C. Dean, Michel X. Goemans, and Jan Vondrák. Approximating the stochastic knapsack problem: The benefit of adaptivity. *Math. Oper. Res.*, 33(4):945–964, 2008.

[DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Mathematical Statistics*, 27(3):642–669, 1956.

[DPR05] A.G. Dimakis, V. Prabhakaran, and K. Ramchandran. Ubiquitous access to distributed data in large-scale sensor networks through decentralized erasure codes. In *International Symposium on Information Processing in Sensor Networks (IPSN)*, 2005.

[DS09] I. Diakonikolas and R. Servedio. Improved approximation of linear threshold functions. In *Proc. 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 161–172, 2009.

[Erd65] P. Erdös. A problem on independent r-tuples. In *Annales Universitatis Scientiarum Budapest*, 1965.

[Fal03] K. Fall. A delay-tolerant network architecture for challenged internets. In *Proc. of ACM SIGCOMM*, 2003.

[Fel68] W. Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, 1968.

[Has94] J. Hastad. On the size of weights for threshold gates. *SIAM Journal on Discrete Mathematics*, 7(3):484–492, 1994.

[JB03] A. Jiand and J. Bruck. Memory allocation in information storage networks. In *IEEE International Symposium on Information Theory (ISIT)*, 2003.

[JDPF05] S. Jain, M.J. Demmer, R.K. Patra, and K.R. Fall. Using redundancy to cope with failures in a delay tolerant network. In *Proc. of ACM SIGCOMM*, 2005.

[JFP04] S. Jain, K.R. Fall, and R.K. Patra. Routing in a delay tolerant network. In *Proc. of ACM SIGCOMM*, 2004.

[Kle06] R. Kleinberg. Personal Communication, October 2006.

[KRT97a] J. Kleinberg, Y. Rabani, and E. Tardos. Allocating bandwidth for bursty connections. In *ACM STOC*, 1997.

[KRT97b] Jon Kleinberg, Yuval Rabani, and Éva Tardos. Allocating bandwidth for bursty connections. In *STOC*, pages 664–673, 1997.

[LD11] Jian Li and Amol Deshpande. Maximizing expected utility for stochastic combinatorial optimization problems. In *FOCS*, pages 797–806, 2011.

[LDT09] D. Leong, A.G. Dimakis, and T.Ho. Distributed storage allocation problems. In *NetCod*, 2009.

[LDT10a] D. Leong, A.G. Dimakis, and T.Ho. Distributed storage allocation for high reliability. In *IEEE International Conference on Communications (ICC)*, 2010.

[LDT10b] D. Leong, A.G. Dimakis, and T.Ho. Symmetric allocations for distributed storage. In *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[LMSS02] M. Luby, M. Mitzenmacher, M.A. Shokrollahi, and D.A. Spielman. Efficient erasure correcting codes. In *IEEE Transactions on Information Theory*, 2002.

[Lub02] M. Luby. LT Codes. In *IEEE FOCS*, 2002.

[LY13] Jian Li and Wen Yuan. Stochastic combinatorial optimization via poisson approximation. In *STOC*, pages 971–980, 2013.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18(3):1269–1283, 1990.

[Mit04] M. Mitzenmacher. Digital fountains: A survey and look forward. In *Information Theory Workshop*, 2004.

[MORS10] K. Matulef, R. O'Donnell, R. Rubinfeld, and R. Servedio. Testing halfspaces. *SIAM J. on Comput.*, 39(5):2004–2047, 2010.

[MTT61] S. Muroga, I. Toda, and S. Takasu. Theory of majority switching elements. *J. Franklin Institute*, 271:376–418, 1961.

[Mur71] S. Muroga. *Threshold logic and its applications*. Wiley-Interscience, New York, 1971.

[MZ10] Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proc. 42nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 427–436, 2010.

[Nik10] Evdokia Nikolova. Approximation algorithms for reliable stochastic combinatorial optimization. In *APPROX-RANDOM*, pages 338–351, 2010.

[OS11] R. O'Donnell and R. Servedio. The Chow Parameters Problem. *SIAM J. on Comput.*, 40(1):165–199, 2011.

[Rag88] P. Raghavan. Learning in threshold networks. In *First Workshop on Computational Learning Theory*, pages 19–27, 1988.

[Ser07] R. Servedio. Every linear threshold function has a low-weight approximator. *Comput. Complexity*, 16(2):180–209, 2007.

[Shi86] I.S. Shiganov. Refinement of the upper bound of the constant in the central limit theorem. *Journal of Soviet Mathematics*, pages 2545–2550, 1986.

[Sho06] M.A. Shokrollahi. Raptor codes. In *IEEE Transactions on Information Theory*, 2006.

[SRFS10] M. Sardari, R. Restrepo, F. Fekri, and E. Soljanin. Memory allocation in distributed storage networks. In *IEEE International Symposium on Information Theory (ISIT)*, 2010.

[Swa11] Chaitanya Swamy. Risk-averse stochastic optimization: Probabilistically-constrained models and algorithms for black-box distributions. In *SODA*, pages 1627–1646, 2011.

[TV09] T.Tao and V. H. Vu. Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. *Annals of Mathematics*, 169:595–632, 2009.