Efficient Mean Estimation with Sub-Gaussian Rates

In this lecture, we study the problem of estimating the mean of a random vector X given a sample of N independent, identically distributed points in $\mathbb{R}^d$. We will introduce an estimation procedure called "median-of-means tournament" that can achieve sub-Gaussian rates only assuming that X has finite second moment. Then, we show how to find an approximate "median-of-means tournament" estimator efficiently with the help of SoS paradigm, that also guarantees sub-Gaussian rates.

1 Mean estimation

The problem we will study in this lecture takes a rather simple form—given N i.i.d. random vectors $X_1, \cdots, X_N$ such that $\mu = \mathbb{E}X_i$ and $\Sigma = \mathrm{Cov}(X_i) = \mathbb{E}\left[(X_i - \mu)(X_i - \mu)^\top\right]$, we want to find a "good estimator" $\hat{\mu}_N(X_1, \cdots, X_N)$ of the sample mean μ.

Indeed, the definition of "good estimator" depends on the loss function we try to minimize. For instance, if we consider the ℓ_2 loss $\mathbb{E}\left[\|\hat{\mu}_N - \mu\|^2\right]$,[1] it is well-known that the optimal linear estimator is exactly the empirical mean $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$.[2] In this lecture, we're instead interested in optimizing the *concentration* of $\hat{\mu}$ around the mean μ. That is, for any given $\delta > 0$, we wish to minimize the value r_δ that satisfies

$$\mathbb{P}\left(\|\hat{\mu}_N - \mu\| > r_\delta\right) \leqslant \delta. \tag{1}$$

For example, when X_i are collected from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, it follows that the N-sample empirical mean $\bar{X}$ satisfies (by [HW71]):

$$\mathbb{P}\left(\left\|\bar{X} - \mu\right\| > \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{2\|\Sigma\|\log(1/\delta)}{N}}\right) \leqslant \delta. \tag{2}$$

It is also shown in [Cat12] that

$$r_\delta = \Omega\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{N}}\right) \tag{3}$$

is a lower bound on the value of r_δ achievable by any estimator under some mild assumptions on the distribution of the X_i. For the mean estimation problem, does there exist an estimator which attains r_δ as in (3)? Lugosi and Mendelson [LM19] give an affirmative answer: the estimator achieving such r_δ is the median-of-means tournament estimator that we will introduce in the next section.

2 Median-of-means tournament estimator

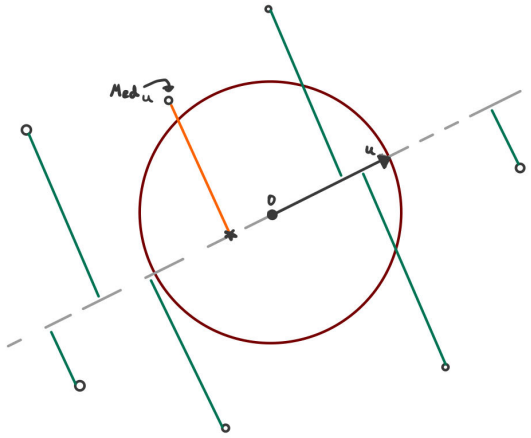
Before introducing the estimator, we might ask the question—can we still use the empirical mean estimator $\bar{X}$? If we have sub-Gaussianity for random samples X_i, one can still choose r_δ as in (3). Unfortunately,

[1] Throughout this lecture $\|\cdot\|$ denotes ℓ_2 norm for vectors and spectral norm for matrices.

[2] We point out that the empirical mean is in general not the optimal estimator without linearity assumption for mean squared error, cf. James-Stein estimator.

indicated by [Cat12], this is not true in general, especially when the distribution of X has heavy tail. The failure of empirical mean can be viewed as its sensitivity to outliers when the underlying distribution has heavy tail.

It is well-known that median is robust to outlier corruptions, and the estimator is based on the idea by extending the concept of median to $\mathbb{R}^d$. For any set of k vectors $v_1, \dots, v_k \in \mathbb{R}^d$ and any unit vector $u \in \mathbb{R}^d$, define their u-directional median $\mathrm{Med}_u(v_1, \dots, v_k) = v_i$ s.t. $\langle v_i, u\rangle = \mathrm{Median}(\langle v_1, u\rangle, \dots, \langle v_k, u\rangle)$ (ties over i broken arbitrarily).



Lugosi and Mendelson introduce a new notion of high-dimensional median which tries to minimize the maximum discrepancy between its projection on u and the u-directional median among all u.

Definition 2.1 (Tournament median, [LM19], as interpreted in [Hop20])**.** For any k vectors $v_1, \cdots, v_k \in \mathbb{R}^d$, their tournament median is defined by

$$\mathrm{Median}(v_1, \dots, v_k) = \arg\min_{x \in \mathbb{R}^d} \sup_{\|u\|=1} |\langle x - \mathrm{Med}_u(v_1, \dots, v_k), u\rangle|. \tag{4}$$

The tournament median can be equivalently defined in a different way, which will be important in our proof that it is a good estimator.

Lemma 2.2. *The tournament median is equivalently given by*

$$\mathrm{Median}(v_1, \dots, v_k) = \arg\min_{x \in \mathbb{R}^d} \sup_{y \in \mathbb{R}^d} r = \|x - y\| \quad \textit{s.t. } \|v_i - x\| \geqslant \|v_i - y\| \textit{ for at least } \frac{k}{2}\ v_i\textit{'s.} \tag{5}$$

Proof. We re-paramterize program (5) as $y = x - ru$ for $\|u\| = 1$. Then equivalent to the condition above is that for at least $k/2$ of v_i's,

$$\|v_i - x\|^2 \geqslant \|v_i - x + ru\|^2 \iff 0 \geqslant 2r\langle v_i - x, u\rangle + \|ru\|^2 \iff \langle x - v_i, u\rangle \geqslant \frac{r}{2}.$$

This implies that $|\langle x - \mathrm{Med}_u(v_1, \dots, v_k), u\rangle| \geqslant \frac{r}{2}$, so that (x, y) yield a pair (x, u) of value $\geqslant \frac{r}{2}$ for (4).

On the other hand, letting (x, u) a solution to (4) with value α, with $v_i = \mathrm{Med}_u(v_1, \dots, v_k)$, we have that $\alpha \leqslant |\langle x, u\rangle - \langle v_i, u\rangle|$. Because $v_i = \mathrm{Med}_u(v_1, \dots, v_k)$, this means that either $\alpha - \langle x, u\rangle \leqslant -\langle v_i, u\rangle \leqslant -\langle v_j, u\rangle$

for $\frac{k}{2}$ indices $j \in [k]$, or $\alpha + \langle x, u\rangle \leqslant \langle v_i, u\rangle \leqslant \langle v_j, u\rangle$ for $\frac{k}{2}$ indices $j \in [k]$. Choosing $y = x + 2\alpha u$ in the latter case , we have that for these j,

$$\|v_j - x\|^2 - \|v_j - y\|^2 = \|v_j - x\|^2 - \|v_j - x - 2\alpha u\|^2 = 4\alpha\langle v_j - x, u\rangle - 4\alpha^2 = 4\alpha(\langle v_j, u\rangle - \langle x, u\rangle - \alpha) \geqslant 0.$$

Hence, for at least $\frac{k}{2}$ indices $j \in [k]$ we have that $\|v_j - x\| \geqslant \|v_j - y\|$. An near-identical argument establishes the same for $y = x - 2\alpha u$ in the former case. Hence we have that an (x, u) pair with value $\alpha = \frac{r}{2}$ for program (4) witnesses a pair (x, y) with value $2\alpha = r$ for program (5). This concludes the proof. □

With this notion of high-dimensional median and its two characterizations in mind, we're ready to propose median-of-means tournament estimator for the population mean μ.

Definition 2.3 (Median-of-means tournament estimator, [LM19])**.** Given i.i.d. observations $X_1, \cdots, X_N$, partition the set $[N] := \{1, \cdots, N\}$ into k blocks $B_1, \cdots, B_k$, each of them has the size N/k[3]. Let $Z_i := \frac{1}{k}\sum_{j\in B_i} X_j$ be the mean of i-th batch B_i. The *median-of-means tournament estimator* is then defined by

$$\hat{\mu}(X_1, \cdots, X_N) := \text{Median}\left(Z_1, \cdots, Z_k\right). \tag{6}$$

The main result of Lugosi and Mendelson shows that for a properly chosen k, the median-of-means tournament estimator successfully achieves sub-Gaussian rate performance of mean estimation.

Theorem 2.4 ([LM19])**.** *Let $\delta \in (0, 1)$ and consider the mean estimator defined (6) with parameter $k = O(\log(1/\delta))$. Suppose $X_1, \cdots, X_N$ are random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ, then for all N*

$$\mathbb{P}\left(\|\hat{\mu}_N - \mu\| > O\left(\sqrt{\frac{\text{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{N}}\right)\right) \leqslant \delta, \tag{7}$$

which matches the lower bound (3).

The proof makes slick use of the optimality of $\hat{\mu}$ for (5). It proceeds in two steps—first, they show that the population mean μ has objective value at most r_δ for (5), and hence the median-of-means estimator $\hat{\mu}$ must achieve objective $\leqslant r_\delta$ as well. En route to doing so, they show that every point a with $\|\mu - a\| > r_\delta$ is a feasible pair for (5) with $(x = a, y = \mu)$ and value $r = \|\mu - a\|$, and so they conclude that since the optimizer $\hat{\mu}$ has value $\leqslant r_\delta$ for program (5), $\hat{\mu}$ must be at least r_δ-close to μ.

Sketch of proof. The proof is based on the following claim.

Claim 2.5. For any two vectors $x, y \in \mathbb{R}^d$ and fixed $p \in (0, 1)$, we say that x p-defeats y if

$$\frac{1}{m}\sum_{i\in B_j}\left(\|X_i - y\|^2 - \|X_i - x\|^2\right) > 0 \tag{8}$$

on more than kp blocks B_j. Then for any $\delta \in (0, 1)$, we can take $k = O(\log(1/\delta))$ and r_δ as in (3), such that with probability at least $1 - \delta$, the population mean μ $\frac{1}{2}$-defeats all $y \in \mathbb{R}^d$ such that $\|y - \mu\| \geqslant r_\delta$.

[3] If N is not divisible by k, we can round N to $k\lfloor N/k\rfloor$, and this affects the analysis negligibly.

The high level idea is to invoke Chebyshev's inequality for any fixed direction $v \in \mathbb{R}^d, \|v\| = r_\delta$ and then apply Hoeffding's inequality to k blocks $B_1, \cdots, B_k$. Taking the union bound on a properly chosen ϵ-net concludes the proof of this claim.

Now equipped with this claim, we're ready to prove Theorem 2.4. Note that

$$0 \leqslant \frac{1}{m} \sum_{i \in B_j} \left(\|X_i - y\|^2 - \|X_i - x\|^2 \right) = -2\langle Z_j, y \rangle + \|y\|^2 + 2\langle Z_j, x \rangle - \|x\|^2 = \|Z_j - y\|^2 - \|Z_j - x\|^2, \tag{9}$$

where in the first equality we used that Z_j is the empirical mean of X_i within block B_j. Then with probability $1 - \delta$, μ $\frac{1}{2}$-defeats b for all $\|b - \mu\| \geqslant r_\delta$. This implies two things: firstly, the program (5) is feasible with value r_δ. This is because we know that for the variable setting $x := \mu$ from (5), choosing $y = x = \mu$ is a feasible setting of variables with $r = 0$, and the fact that μ will $\frac{1}{2}$-defeat any y with $\|y - \mu\| \geqslant r_\delta$ gives an upper bound of $r \leqslant r_\delta$ when we choose $x := \mu$.

Secondly, we must have $\|\text{Median}(Z_1, \ldots, X_k) - \mu\| \leqslant r_\delta$, as otherwise choosing $y := \mu$ we would contradict the optimality of $x = \hat{\mu}$ the tournament median for program (5). Hence it must follow that

$$\|\text{Median}(Z_1, \cdots, Z_k) - \mu\| < r_\delta.$$

The proof is concluded. □

3 Efficiently computable algorithm by SoS

Although [LM19] gives a concise and beautiful construction that gives a mean estimator with sub-Gaussian rates mathematically, it still remains unaddressed—is this estimator efficiently computable? That is, can we find an algorithm running within polynomial time?

The answer is affirmative, first established by Hopkins who shows the following result.

Theorem 3.1 ([Hop20]). *There are universal constants C_0, C_1, C_2 such that for every positive integers N, d and $\delta > 2^{-n/C_2}$ there is an algorithm which runs in time $O(nd) + (d \log(1/\delta))^{C_0}$ such that for any i.i.d. random samples $X_1, \cdots, X_N$ from distribution with mean μ and covariance matrix Σ, the algorithm outputs a vector $\hat{\mu}_N$ such that*

$$\mathbb{P}\left(\|\hat{\mu}_N - \mu\| > C_1 \left(\sqrt{\frac{\text{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\| \log(1/\delta)}{N}} \right) \right) \leqslant \delta. \tag{10}$$

The algorithm that realizes Theorem 3.1 is a sophisticated implementation of the usual SoS recipe: the high level idea is to find a large sets of polynomial constraints that provides an approximate solution to something resembling Lugosi and Mendelson's median-of-means tournament estimator (6), then SoS-izing the proof that the estimator is close to the mean. The SoS proof has degree-8, so the naive running time is roughly $O(d^{24})$.

Speeding up SoS. Here, we will instead present a simpler algorithm and analysis due to Cherapanamjeri, Flammiron, and Bartlett [CFB19]. This algorithm is a dramatically "sped-up" version of Hopkins' algorithm: they realize that the information provided by the full degree-8 SoS SDP is overkill, and that a lightweight SDP (a degree-2 SoS program in $k + d$ variables and $2k + 1$ constraints) suffices to provide *gradient information* which allows them to implement a descent-based algorithm.

They use the following key concept of centrality introduced by Hopkins:

Definition 3.2 (Centrality)**.** Let $v_1, \cdots, v_k \in \mathbb{R}^d, r > 0$ and $p \in [0,1]$. We say that $x \in \mathbb{R}^d$ is (r,p)-central if for every unit $u \in \mathbb{R}^d$ there are at most pk vectors $v_1, \cdots, v_k$ such that $\langle v_i - x, u\rangle \geqslant r$. The minimum p such that x is (r,p)-central with respect to $v_1, \cdots, v_k$ is given by the optimum of the following quadratic program in variables $b_1, \cdots, b_k$ and $u_1, \cdots, u_d$.

$$\begin{aligned} \text{maximize } & \frac{1}{k}\sum_{i=1}^{k} b_i \\ \text{s.t. } & b_i = b_i^2, \quad \forall i \in [k] \\ & \|u\|^2 \leqslant 1, \\ & b_i\langle v_i - x, u\rangle \geqslant b_i r, \quad \forall i \in [k]. \end{aligned} \tag{MTE}$$

The variable b_i is meant to stand in for $\mathbf{1}[\langle v_i - x, u\rangle \geqslant r]$. For any fixed $p < 1/2$, we've already seen in Claim 2.5 that the population mean μ is (r_δ, p)-central with probability at least $1-\delta$ when $v_i = Z_i$ (the mean of the i-th block B_i in the median-of-means paradigm). Suppose we can find another vector $\hat{\mu}$ which is also (r_δ, p)-central, then since $p < 1/2$, for any $u, w \in \mathcal{S}^{d-1}$ there must be some $i \in [k]$ such that

$$\langle Z_i - \mu, u\rangle < r_\delta, \quad \text{and} \quad \langle Z_i - \hat{\mu}, w\rangle < r_\delta,$$

In particular, this allows us to take $u = (\mu - \hat{\mu})/\|\mu - \hat{\mu}\|$, which further implies

$$2r_\delta > \langle Z_i - \mu, -u\rangle + \langle Z_i - \hat{\mu}, u\rangle = \left\langle \mu - \hat{\mu}, \frac{\mu - \hat{\mu}}{\|\mu - \hat{\mu}\|} \right\rangle = \|\mu - \hat{\mu}\|,$$

meaning that $\hat{\mu}$ is a valid mean estimator with sub-Gaussian rates. We thus have a reduction to the following problem: can we find a (r_δ, p)-central $\hat{\mu}$ effectively?

At this point, one issue is that the polynomial optimization problem (MTE) (which we could try to solve approximately using SoS), only helps to verify whether a fixed vector x is (r_δ, p)-central. In [Hop20] the problem is dealt with by designing a larger and more sophisticated SoS program which optimizes over x as well. In the next section, we will see how to instead work with the simple subproblem (MTE) within an iterative gradient-descent style algorithm.

4 Speeding up mean estimation

The main insight of Cherpanamjeri et al. [CFB19] is that starting with a (possibly bad) estimate x for a central point, the centrality optimization problem gives a solution vector u that is correlated with the direction $x - \mu$. Informally, given an estimate of the central point x_t, we can repeatedly solve (MTE) with $x = x_t$ and use the solution r_t to estimate the distance $\|\mu - x_t\|$. With this good estimate r_t of distance, we can solve (MTE) fixing $r = r_t$ and use the solution u_t as a descent direction guaranteed to be correlated with the direction $\Delta = (\mu - x_t)/\|x_t - \mu\|$, so we update $x_{t+1} = x_t + \varepsilon u_t$ for a well-chose ε. Repeating this two step process gives us an algorithm that iteratively improves its estimate of μ.

Lemma 4.1. *Define* $r^* = 300\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{N}}\right)$*, and let* r_t *be the maximum value of* r *such that* (MTE) *with* $x = x_t$ *has value at least* 0.9*. Assume*

$$\|x_t - \mu\| \geqslant 4r^*$$

and let $\Delta = (\mu - x_t)/\|x - \mu\|$*. Then*

$$\langle u_t, \Delta\rangle \geqslant \frac{1}{2}$$

where u_t *is the solution direction* u *from (MTE) with* $x = x_t$ *and* $r = r_t$*.*

Before proving the lemma, note that given this lemma we could prove that the descent algorithm suggested above will work: we make progress towards μ so long as our distance is at least $4r^*$. Formally, we would need to bound the number of iterations, etc. However, we will omit these details here.

Proof of Lemma 4.1. By the definition of r_t, we have $\langle Z_i - x_t, u_t\rangle \geqslant r_t$ for $0.9k$ of the Z_i. We then use the following claim (which we will not prove), which roughly asserts that most of the bucket means are not too far from μ:

Claim 4.2. For the bucket means $Z_1, \dots, Z_k$, we have:

$$\forall u \in \mathbb{R}^d, \|u\| = 1 \implies |\{i : \langle Z_i - \mu, u\rangle \geqslant r^*\}| \leqslant 0.05k.$$

So by the pigeonhole principle there must be an index j where $\langle Z_j - x_t, u_t\rangle \geqslant r_t$ and also $\langle Z_j - \mu, u_t\rangle < r^*$. From this we have

$$r_t \leqslant \langle Z_j - x_t, u_t\rangle = \langle Z_j - \mu, u_t\rangle + \langle \mu - x_t, u_t\rangle \leqslant r^* + \|\mu - x_t\| \cdot \langle \Delta, u_t\rangle \implies \langle \Delta, u_t\rangle \geqslant \frac{r_t - r^*}{\|x_t - \mu\|}.$$

Finally, we will show that $r_t + r^* \geqslant \|x_t - \mu\|$, which is enough to complete the proof because $\|x_t - \mu\| \geqslant 4r^*$ by assumption, so $r_t - r^* \geqslant \frac{1}{2}\|x_t - \mu\|$. Note that if $\|x_t - \mu\| \leqslant r^*$ we are done, so we assume the opposite. In that case, applying Claim 4.2 with $u = \Delta$, we have that for $.95k$ of the Z_i,

$$\langle Z_i - x_t, \Delta\rangle = \langle Z_i - \mu, \Delta\rangle + \langle \mu - x_t, \Delta\rangle \geqslant \|x_t - \mu\| - r^* > 0.$$

Thus by definition of r_t we must have $r_t \geqslant \|x_t - \mu\| - r^*$. This concludes the proof. □

4.1 Efficient gradient computation via SoS relaxation

Now that we have seen how we can accurately estimate our distance from μ and find descent directions guaranteed to move us closer to μ, we will relax (MTE) and show that qualitatively the same results hold. At each step we will choose r_t to be the maximum value of r such that the *SoS relaxation* of (MTE) has value at least 0.9, and then we will choose u_t to be an eigenvector sampled from $\tilde{\mathbf{E}}[uu^\top]$ proportional to its eigenvalue. To show that this works, we need an SoS version of Lemma 4.1.

Lemma 4.3. *Let r^* be as above, let r_t be the maximum value of r such that the degree-2 SoS relaxation of (MTE) with $x = x_t$ has value at least 0.9, and assume that $r_t \geqslant 100r^*$. Also let $\Delta = (\mu - x_t)/\|\mu - x_t\|$. Then it must be the case that*

$$\|x_t - \mu\| \leqslant 1.02r_t,$$

And also, with high probability, in $\tilde{O}(d^3 + k)$ time we can use randomized rounding on $\tilde{\mathbf{E}}[uu^\top]$ to obtain a vector u_t which satisfies $\langle u_t, \Delta\rangle \geqslant \frac{1}{2}$.

Proof. Ultimately, we will choose u_t to be proportional to an eigenvector of $\tilde{\mathbf{E}}[uu^\top]$. Notice that $\tilde{\mathbf{E}}[uu^\top]$ is a trace-1 positive semidefinite matrix (from the constraint $\|u\|^2 = 1$), so the eigenvalues naturally give a distribution over the eigenvectors of $\tilde{\mathbf{E}}[uu^\top]$. For now, let U be a random unit eigenvector sampled from this distribution.

Recall that in the proof of Lemma 4.1, we made crucial use of two facts: firstly, that $\langle u_t, Z_i - x\rangle = \Omega(r_t)$ for many $i \in [k]$, and secondly that $\|x_t - \mu\| = O(r_t)$. We will need to establish both for u_t and r_t derived from the SoS relaxation.

In pursuit of this, we prove the following easy claim, which gives a lower bound on the expected squared correlation $\mathbf{E}_U\langle U, \mu - x_t\rangle^2$ from any subset of indices:

Claim 4.4. For any $S \subset [k]$, for $U \sim \tilde{\mathbf{E}}[uu^\top]$ according to the eigenvalue distribution,

$$\frac{r_t^2}{2k \cdot |S|} \cdot \left(\sum_{i \in S} \tilde{\mathbf{E}}[b_i] \right)^2 - \mathbf{E}_{i \sim S} \mathbf{E}_U \langle U, Z_i - \mu \rangle^2 \leqslant \mathbf{E}_U \langle U, x_t - \mu \rangle^2.$$

Proof. The proof follows from a sequence of degree-2 SoS inequalities. We have that

$$\left(\sum_{i \in S} b_i r_t \right)^2 \leqslant \left(\sum_{i \in S} b_i \langle u, Z_i - x_t \rangle \right)^2 \tag{11}$$

$$\leqslant \left(\sum_{i \in S} b_i^2 \right) \left(\sum_{i \in S} \langle u, Z_i - x_t \rangle^2 \right) \tag{12}$$

$$\leqslant k \cdot \sum_{i \in S} \langle u, Z_i - x_t \rangle^2 \;=\; k \cdot \sum_{i \in S} \left(\langle u, Z_i - \mu \rangle + \langle u, \mu - x_t \rangle \right)^2 \tag{13}$$

$$\leqslant 2k \cdot \sum_{i \in S} \left(\langle u, Z_i - \mu \rangle^2 + \langle u, \mu - x_t \rangle^2 \right) \tag{14}$$

where in (11) we have used the constraint $b_i r_t \leqslant b_i \langle u, Z_i - x_t \rangle$, in (12) we used Cauchy-Schwarz, in (13) we have used the constraints $b_i^2 = b_i$, and in the final inequality we have used that $a^2 + b^2 \geqslant \frac{1}{2} ab$ (a degree-2 SoS inequality) and Cauchy-Schwarz.

Applying the pseudoexpectation operator on both sides,

$$r_t^2 \tilde{\mathbf{E}} \left[\sum_{i \in S} b_i \right]^2 \leqslant 2k \cdot \left\langle \tilde{\mathbf{E}}[uu^\top], |S|(\mu - x_t)(\mu - x_t)^2 + \sum_{i \in S} (Z_i - \mu)(Z_i - \mu)^\top \right\rangle$$
$$= 2k|S| \cdot \left(\mathbf{E}_{i \sim S} \mathbf{E}_U \langle U, Z_i - \mu \rangle^2 + \mathbf{E}_U \langle U, \mu - x_t \rangle^2 \right).$$

where the final equality uses that $\tilde{\mathbf{E}}[uu^\top] = \sum_{\ell=1}^d p_\ell U_\ell U_\ell^\top$, for p_ℓ, U_ℓ the eigenvectors of $\tilde{\mathbf{E}}[uu^\top]$. This is our desired conclusion. □

In order to use the above to show that with reasonable probability the quantity $\langle U, x_t - \mu \rangle$ is $\Omega(r_t)$, we will apply it with $S \subset [k]$ the set of indices for which $|\langle Z_i - \mu, U \rangle| \leqslant r^*$. Notice that it is OK that we are choosing S to depend on U. Notice as well that by Claim 4.2, $|S| \geqslant 0.95k$. Since by assumption $\tilde{\mathbf{E}}[\sum_i b_i] \geqslant 0.9k$ and $\tilde{\mathbf{E}}[b_i] \in [0, 1]$ for all i, it must be the case that $\tilde{\mathbf{E}}[\sum_{i \in S} b_i] \geqslant 0.85k$. Substituting into Claim 4.4, we have that

$$0.37 \cdot r_t^2 \leqslant r_t^2 \frac{(0.85)^2}{2 \cdot 0.95} - (r^*)^2 \leqslant \mathbf{E}_U \langle U, \mu - x_t \rangle^2, \tag{15}$$

where on the left-hand side we have used that $r^* < \frac{1}{100} r_t$.

Now, we have that the expected squared correlation between U and $\mu - x_t$ is $\Omega(r_t^2)$. We must now argue that this correlation is significant (compared to $\|\mu - x_t\|$). For this, we will need to relate $\|\mu - x_t\|$ and r_t, for which we will need the following claim:

Claim 4.5. For each unit vector $w \in \mathbb{R}^d$, for at least $0.1k$ of $i \in [k]$, we have $\langle Z_i - x_t, w \rangle \leqslant 1.01 r_t$.

Proof. Since r_t is defined to be the maximum value of r such that the SoS relaxation of (MTE) with $x := x_t$ has value at least 0.9, the following distribution is *not* feasible as a pseudodistribution: with $r := 1.01 \cdot r_t$, set $u := w$, and set $b_i := 1$ for all i. Note that the only infeasible constraints are the constraints $b_i \langle Z_i - x, u \rangle \geqslant$

$b_i r$, so that it must be the case that for some subset of $T \leqslant k$ of the $i \in [k]$, $b_i\langle Z_i - x_t, \Delta\rangle \leqslant b_i 1.01 r_t$. Now, we update our pseudodistribution, setting $b_j = 0$ for each $j \in T$. Since the value of this updated pseudodistribution is $\frac{k-T}{k} \leqslant 0.9$ (by the maximality of r_t and the fact that we set $r = 1.01 r_t$), we must have $T \geqslant 0.1k$. □

Applying Claim 4.5 with $w = \Delta$ along with Claim 4.2 and the pigeonhole principle, we conclude that there must exist some $i \in [k]$ with

$$1.01 \cdot r_t \geqslant \langle Z_i - x_t, \Delta\rangle = \langle Z_i - \mu, \Delta\rangle + \langle x_t - \mu, \Delta\rangle \geqslant \|x_t - \mu\| - r^*, \quad \Longrightarrow \|x_t - \mu\| \leqslant 1.02 \cdot r_t,$$

where we have used that by assumption $r^* \leqslant \frac{1}{100} r_t$. Substituting this bound in to (15), we conclude that

$$\frac{1}{3}\|\mu - x_t\|^2 \leqslant 0.37 r_t^2 \leqslant \mathbf{E}_U \langle U, \mu - x_t\rangle^2.$$

Further, since the random variable $|\langle U, \mu - x_t\rangle|$ takes value at most $\|\mu - x_t\| \leqslant 1.02 r_t$, by an averaging argument, $\mathbf{Pr}[|\langle U, \mu - x_t\rangle| \geqslant \frac{1}{\sqrt{3}} r_t] \geqslant \frac{1}{2}$. So plugging in our bound on $\|x_t - \mu_t\|$, we conclude that with probability at least $\frac{1}{2}$ over the choice of U, $|\langle U, \Delta\rangle| \geqslant \frac{1}{1.02 \cdot \sqrt{3}}$.

One final detail remains, which is, once we have sampled U, how can we test whether this event has occured and either $-U$ or U is truly correlated with Δ? We will do this by testing whether, for at least $0.8k$ indices $i \in [k]$,

$$\langle Z_i - x_t, U\rangle \geqslant \frac{1}{\sqrt{3}} r_t - r^*,$$

in which case we set $u_t = U$ (otherwise we repeat the test with $-U$). To see that this test passes if and only if $\langle U, \Delta\rangle \geqslant \frac{1}{\sqrt{3}} \pm 0.05$, notice that by Claim 4.2, for $0.95k$ good indices $i \in [k]$, we must have $|\langle Z_i - \mu, U\rangle| < r$, and so for these indices,

$$\langle Z_i - x_t, U\rangle = \langle Z_i - \mu, U\rangle + \langle \mu - x_t, U\rangle = \langle \mu - x_t, U\rangle \pm r^*.$$

So if the test passes then these good indices i witness that $\langle \mu - x_t, U\rangle \geqslant \frac{1}{\sqrt{3}} r_t - 2r^* \geqslant \frac{1}{2}\|\mu - x_t\|$. Similarly, if $\langle U, \mu - x_t\rangle \geqslant \frac{1}{\sqrt{3}} r_t$ then the test will pass because there are more than $0.8k$ good indices. This gives us our conclusion.

Finally, if $\pm U$ do not have sufficient correlation with Δ, we resample a fresh U', until we find a U' satisfying one of the above conditions. Since each U is good with probability $\frac{1}{2}$, with high probability, we will not need to resample more than $\tilde{O}(1)$ times. Computing the eigenvalue decomposition of $\tilde{\mathbf{E}}[uu^\top]$ takes $O(d^3)$ time, each sampling step takes $O(d)$ time (to sample an eigenvector), and each testing step takes $O(kd)$ time (to compute k dot products of d-dimensional vectors). This concludes our proof. □

Running time. We briefly examine the runtime. Note that this relaxation naively has $O(k^2 + d^2)$ variables, but we can reduce dimensions by projecting onto a subspace containing the bucket means to have only $O(k^2)$ variables at the cost of $O(k^2 d)$ preprocessing. Then we have an SDP with $O(k^2)$ variables and $O(k)$ constraints, which standard methods can solve in $O(k^{3.5})$ time. Letting $\tilde{O}$ denote order up to logarithmic terms, a call to the distance estimation procedure requires only $\tilde{O}(1)$ calls to (MTE) to do binary search for r. Thus the total cost of estimating the distance to the mean is $\tilde{O}(k^{3.5})$. Similarly, the cost of estimating the gradient is also $\tilde{O}(k^{3.5})$. Since we only run $\tilde{O}(1)$ iterations, the total cost of the proposed algorithm is $\tilde{O}(k^{3.5} + k^2 d)$. This is still far from linear, but a vast improvement over the d^{24}-time algorithm.

5 Conclusion

We have introduced the problem of mean estimation and showed why the empirical mean is insufficient for heavy-tailed distributions when the goal is to be within a small neighborhood of the true mean. We then introduced an efficient median of means estimator that achieves sub-Gaussian rates under the mild assumptions of having two moments. Lastly, we showed how this could be framed as an SoS problem and an efficient polynomial-time descent algorithm could compute such an estimator.

Bibliographic remarks. Cherapanamjeri et al. give a different analysis of the SDP relaxation of (MTE), in traditional semidefinite programming analysis rather than taking an SoS perspective as in Lemma 4.3.[4] The omitted proof of Claim 2.5 is the proof of Lemma 1 in [LM19], and the omitted proof of Claim 4.2 is the proof of Corollary 5 in [CFB19].

Contact. Comments are welcome at tselil@stanford.edu.

References

[Cat12] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012. 1, 2

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806, 2019. 4, 5, 9

[Hop20] Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2):1193–1213, 2020. 2, 4, 5

[HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971. 1

[LM19] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019. 1, 2, 3, 4, 9

[4] And consequently, the presentation of Lemma 4.3 has not yet been refereed.