

408 A Preliminaries

409 **Lemma A.1** (High Dimensional Mean-Variance Decomposition). *For any set of points $A \subset \mathbb{R}^d$ and*
 410 *any $c \in \mathbb{R}^d$, we have $\sum_{p \in A} \|p - c\|^2 = \sum_{p \in A} \|p - \mu\|^2 + n \cdot \|\mu - c\|^2$.*

411 *Proof.* We have

$$\begin{aligned} \sum_{p \in A} \|p - c\|^2 &= \sum_{p \in A} \|p - \mu + \mu - c\|^2 \\ &= \sum_{p \in A} (\|p - \mu\|^2 + \|\mu - c\|^2 - 2(p - \mu)^\top (\mu - c)) \\ &= \sum_{p \in A} \|p - \mu\|^2 + n \cdot \|\mu - c\|^2, \end{aligned}$$

412 where the last equality follows from $\sum_{p \in A} (p - \mu) = 0$. □

413 B Mean Estimation via Order Statistics

414 In this section, we leverage order statistics of the candidate means so as to allow for a quicker
 415 aggregation of a suitable candidate mean.

Algorithm 3 MINSUMSELECT(P, i)

Input: Set of points $p_1, \dots, p_{|P|}$, recursion depth i
if $i = 0$ **or** $|P| = 1$ **then**
 $W \leftarrow P$
else
 Split P arbitrarily into $\sqrt{|P|}$ -sized clusters $\{P_1, \dots, P_{\sqrt{|P|}}\}$
 $W \leftarrow \emptyset$
 for each P_j **do**
 $W \leftarrow W \cup \{\text{MINSUMSELECT}(P_j, i - 1)\}$
Output COMPUTEWINNER(W)

Algorithm 4 COMPUTEWINNER(P)

for each $p_j \in P$ **do**
 Let $p'_j \in P$ be the $\frac{7}{10}|P|$ -closest point to p_j
 Compute $D_j := \sum_{p \in P, \|p - p_j\| \leq \|p'_j - p_j\|} \|p - p_j\|$
Output $\arg \min_{p_j \in P} D_j$

416 The main result is the following:

417 **Theorem B.1.** *Algorithm 1 run with Algorithm 3 as an aggregation routine outputs a $(1 + \varepsilon)$*
 418 *approximation with probability $1 - \delta$, using a sample of size $O(100^i \varepsilon^{-1} \log \delta^{-1})$, and running in*
 419 *time*

$$O\left(\left(100^i \varepsilon^{-1} + \log^{2^{-i}} \delta^{-1}\right) d \log \delta^{-1}\right),$$

420 *for any non-negative integer i .*

421 The key observation is that a good mean can always be identified via COMPUTEWINNER as an
 422 estimate $\hat{\mu}_j$ minimizing the sum of distances of the $\frac{7}{10}$ th means closest to $\hat{\mu}_j$. In a nutshell, any mean
 423 minimizing such a sum must be close to sufficiently many successful estimates. Unfortunately, a
 424 naive implementation of this idea takes time $O(\log^2 \delta^{-1} \cdot d)$, as proven in the following lemma.

425 **Lemma B.2.** *COMPUTEWINNER(P) takes time $O(|P|^2 \cdot d)$.*

426 *Proof.* We compute all pairwise distances between the points in P , which takes time $O(|P|^2 \cdot d)$.
 427 To compute D_j , we first have to find the $\frac{7}{10}|P|$ -closest point to p_j , which takes time $O(|P|)$ with a
 428 sufficiently good rank select procedure, see [Cormen et al., 2009, Chapter 9.3]. Thereafter, summing
 429 up all the distances takes time $O(|P|)$ per $p_j \in P$. \square

430 Nevertheless, these scores yield an improved running time by arbitrarily partitioning the estimates
 431 into $\sqrt{|P|}$ groups, finding a good estimate in every group via the truncated sum statistic each in time
 432 $|P| \cdot d$, for a total running time of $|P|^{3/2} \cdot d$, and then selecting the best estimate via another truncated
 433 sum statistic on all estimates returned by the groups in time $|P|$. The final algorithm consists of
 434 applying this idea recursively.

435 To prove that MINSUMSELECT outputs a good solution, we require a parameterized notion of
 436 a successful empirical mean. We say that a mean μ_j is γ -good, if $\|\mu_j - \mu\| \leq \gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}}$. A
 437 straightforward application of the Chernoff bound guarantees us that all but a very small fraction of
 438 the points are γ -good. Assuming this, the following lemma determines the quality of the computed
 439 solution.

440 **Lemma B.3.** *Let P be a set of means such that at least $\left(1 - \left(\frac{3}{10}\right)^{i+1}\right)$ of the means are γ -good.*
 441 *Then MINSUMSELECT(P, i) returns a mean that is $5^{i+1}\gamma$ -good.*

442 We prove this lemma by induction. We will use the following lemma in both the base case and the
 443 inductive step.

444 **Lemma B.4.** *Given a set of means P , suppose that at least $\frac{7}{10}|P|$ are γ -good. Then, the estimate
 445 returned by COMPUTEWINNER(P) is 5γ -good.*

446 *Proof.* First, let $\hat{\mu}' \in P$ be any γ -good estimator. We know that the $\frac{7}{10}|P|$ -closest estimators to $\hat{\mu}'$
 447 have distance at most $2\gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}}$, which likewise implies that $\min_{\mu_j \in P} D_j \leq \frac{7}{10}|P| \cdot 2\gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}}$. Now
 448 suppose that $\hat{\mu}_j$ is the estimator returned by the algorithm. Let $G(\hat{\mu}_j)$ be the set of γ -good estimators
 449 among the $\frac{7}{10}|P|$ closest to $\hat{\mu}_j$. By assumption, we have $|G(\hat{\mu}_j)| \geq \frac{4}{10}|P|$, which implies

$$\sum_{\hat{\mu}_k \in G(\hat{\mu}_j)} \|\hat{\mu}_j - \hat{\mu}_k\| \leq \frac{7}{10}|P| \cdot 2\gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}} = \frac{7}{5}\gamma|P|\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}},$$

450 which gives

$$\min_{\hat{\mu}_k \in G(\hat{\mu}_j)} \|\hat{\mu}_j - \hat{\mu}_k\| \leq \frac{7}{2}\gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}}.$$

451 Therefore,

$$\|\hat{\mu}_j - \mu\| \leq \min_{\hat{\mu}_k \in G(\hat{\mu}_j)} \|\hat{\mu}_j - \hat{\mu}_k\| + \|\hat{\mu}_k - \mu\| \leq \frac{7}{2}\gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}} + \gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}} \leq 5\gamma\sqrt{\frac{\varepsilon_{\text{OPT}}}{n}}. \quad \square$$

452 *Proof of Lemma B.3.* We proceed with the induction starting from $i = 0$.

453

454 **Base Case.** For the base case $i = 0$, MINSUMSELECT($P, 0$) only calls COMPUTEWINNER. Thus,
 455 this case holds due to Lemma B.4.

456

Inductive Step. Let $P_1, \dots, P_{\sqrt{|P|}}$ be the clusters of P computed when first calling
 MINSUMSELECT(P, i). By assumption, we have at most $\left(\frac{3}{10}\right)^{i+1}|P|$ means that are not γ -good.
 This implies that the number of clusters with more than $\left(\frac{3}{10}\right)^i \sqrt{|P|}$ means that are not γ -good is at
 most

$$\frac{\left(\frac{3}{10}\right)^{i+1}|P|}{\left(\frac{3}{10}\right)^i \sqrt{|P|}} = \frac{3}{10} \sqrt{|P|}.$$

457 Denote this set by $B(P)$ and let $G(P)$ be the remaining clusters. For each $P_j \in B(P) \cup G(P)$, let
 458 $\hat{\mu}_j$ returned by MINSUMSELECT($P_j, i - 1$). If $P_j \in G(P)$, we may use the inductive hypothesis

459 which states that the mean $\hat{\mu}_j$ returned by $\text{MINSUMSELECT}(P_j, i-1)$ is $5^i\gamma$ -good. Since at least
 460 $\frac{7}{10}\sqrt{|P|}$ of the thus computed means are $5^i\gamma$ -good, we may use Lemma B.4 which shows that the
 461 final mean returned by $\text{COMPUTEWINNER}(\cup_{P_j \in B(P) \cup G(P)} \{\hat{\mu}_j\})$ is $5^{i+1}\gamma$ -good, which concludes
 462 the proof. \square

463 To conclude the proof, we require two more arguments. First, we must show that, for an appropriate
 464 choice of a and b , at least a $\left(1 - \left(\frac{3}{10}\right)^{i+1}\right)$ fraction of the means are 1-good with probability at least
 465 $1 - \delta$, allowing us to use Lemma B.3. Second, we will argue the running time. The first is a simple
 466 application of the Chernoff bound.

467 **Lemma B.5.** *For $a \geq 2 \cdot 25^{i+1} \cdot \left(\frac{10}{3}\right)^{i+1}$ and $b = 3$, at least $\left(1 - \left(\frac{3}{10}\right)^{i+1}\right)$ of the means are*
 468 *$5^{-(i+1)}$ -good with probability $1 - \delta$.*

469 *Proof.* Let $X_i = \begin{cases} 1 & \text{if } \mu_i \text{ is not } 5^{-(i+1)}\text{-good} \\ 0 & \text{if } \mu_i \text{ is } 5^{-(i+1)}\text{-good} \end{cases}$. We have $\mathbb{E}[\|\hat{\mu}_i - \mu\|^2] = \frac{1}{a} \cdot \frac{\varepsilon_{\text{OPT}}}{n}$ due to Lemma
 470 2.1. This implies due to Markov's inequality that $\mathbb{P}[X_i = 1] \leq \frac{25^{i+1}}{a}$ which, by our choice of a , is
 471 less than $\frac{1}{2} \cdot \left(\frac{3}{10}\right)^{i+1}$. Thus, by the Chernoff bound

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^{b \log \delta^{-1}} X_i \geq \left(\frac{3}{10}\right)^{i+1} b \log \delta^{-1}\right] &\leq \mathbb{P}\left[\sum_{i=1}^{b \log \delta^{-1}} X_i \geq 2 \cdot \frac{25^{i+1}}{a} \cdot b \log \delta^{-1}\right] \\ &\leq \exp\left(-\frac{25^{i+1}}{3a} \cdot b \log \delta^{-1}\right), \end{aligned}$$

472 which is at most δ by our choice of b . \square

The approximation guarantee now follows from Lemma B.3 and Lemma B.5. What remains to be shown is the running time. Since we split a collection of t means into \sqrt{t} clusters, the running time consists of the time required to recursively run the algorithm on the \sqrt{t} clusters each of size \sqrt{t} and consolidating via COMPUTEWINNER , which takes $O(t \cdot d)$ time. Thus, starting with an instance of size $|P_0| \in O(\log \delta^{-1})$, we can solve for the recursion

$$T(|P|) = \begin{cases} \sqrt{|P|} \cdot T(\sqrt{|P|}) + |P| \cdot d & \text{if } |P| \geq |P_0|^{2^{-i}} \\ |P|^2 \cdot d & \text{if } |P| \leq |P_0|^{2^{-i}} \end{cases},$$

473 which yields the desired running time $O(i \cdot d \log \delta^{-1} + (\log \delta^{-1})^{1+2^{-i}} \cdot d) = O((\log \delta^{-1})^{1+2^{-i}} \cdot d)$.

474 We next formalize this and complete the proof of Theorem B.1.

475 *Proof of Theorem B.1.* Throughout this proof, assume $a \geq 2 \cdot 25^{i+1} \left(\frac{10}{3}\right)^{i+1}$ and $b \geq 3$.

476 We first argue correctness, then running time. Let P denote the entire set of means passed to the
 477 MINSUMSELECT algorithm. By Lemma B.3 and Lemma B.5 and with our choices of a and b ,
 478 $\text{MINSUMSELECT}(P, i)$ returns a 1-good mean, that is a $(1 + \varepsilon)$ -approximate mean with probability
 479 at least $1 - \delta$.

What remains to be shown is the running time. Denote by $|P|_0 = b \log \delta^{-1}$ the initial set of sample means. $\text{MINSUMSELECT}(P, i)$ computes $\sqrt{|P|}$ children, each of which recursively calls MINSUMSELECT . Consolidating via COMPUTEWINNER takes time $O(\sqrt{|P|}^2 \cdot d) = O(|P| \cdot d)$ due to Lemma B.2. Thus the overall recursion takes time

$$T(|P|) = \begin{cases} \sqrt{|P|} \cdot T(\sqrt{|P|}) + |P| \cdot d & \text{if } |P| \geq |P|_0^{2^{-i}} \\ |P|^2 \cdot d & \text{if } |P| \leq |P|_0^{2^{-i}} \end{cases},$$

480 which solves to a running time of $O(i \cdot |P|_0 \cdot d + |P|_0^{1+2^{-i}} \cdot d) = O((b \log \delta^{-1})^{1+2^{-i}} \cdot d)$.
 481 The time to compute the initial set of means is $O(a\varepsilon^{-1} \cdot b d \log \delta^{-1})$, which with our choice
 482 of a and b , and some overestimation of the constants, leads to an overall running time of
 483 $O\left(\left(100^i \varepsilon^{-1} + (\log \delta^{-1})^{2^{-i}}\right) \cdot d \log \delta^{-1}\right)$. \square

Remark B.6. For any constant choice of recursion depth, the sample complexity only increases by constants. We did not attempt to optimize the constants, but the exponential dependency on i is not avoidable. If we were to prioritize the running time over the sample complexity, we can set a recursion depth of $i = \log \log \log \delta^{-1}$, which achieves a running time and sample complexity of $O(\varepsilon^{-1} \log \delta^{-1} \text{poly}(\log \log \delta^{-1}) \cdot d)$.

Furthermore, if the recursion depth is shallow (i.e., for small values of i), the recursion can be improved via a better choice of cluster size. Specifically, in the case $i = 1$, that is with just one set of children, Theorem B.1 yields a running time of $O((\varepsilon^{-1} + \sqrt{\log \delta^{-1}}) \cdot d \log \delta^{-1})$. If we instead choose $(\log \delta^{-1})^{2/3}$ many clusters, each consisting of $\sqrt[3]{\log \delta^{-1}}$ many estimators, we obtain a running time of $O((\varepsilon^{-1} + \sqrt[3]{\log \delta^{-1}}) \cdot d \log \delta^{-1})$. Similar improvements for other values i are also possible, but these improvements become increasingly irrelevant compared to the bounds given in Theorem B.1 as i gets larger.

C Learning the Mean of a High-Dimensional Distribution

Theorem 3.2. Let X_1, X_2, \dots, X_N be independent samples from distribution \mathcal{D} with variance μ and covariance matrix Σ , then the coordinate-wise median-of-means estimator ν_{CWM} , with probability at least $1 - \delta$, satisfies

$$\|\nu_{\text{CWM}} - \mu\| \leq 40 \sqrt{\frac{\text{Tr}(\Sigma) \log(\delta^{-1})}{N}}$$

Proof. The N samples are partitioned into $8 \log \delta^{-1}$ subsamples of size $\frac{N}{8 \log \delta^{-1}}$ each. The algorithm computes the empirical mean of each subsample and returns the coordinate-wise median of the set of empirical means.

We prove the analog of Lemma 2.3 for this case, showing that a large fraction of sample means lie close to the true mean. For an empirical mean of $\frac{N}{8 \log \delta^{-1}}$ samples, we have $\mathbb{E}[\|\hat{\mu} - \mu\|^2] = \frac{8 \text{Tr}(\Sigma) \log \delta^{-1}}{N}$.

We call an empirical mean *good* if $\|\hat{\mu} - \mu\| \leq r$ where $r = 13 \sqrt{\frac{\text{Tr}(\Sigma) \log \delta^{-1}}{N}}$ and let G denotes the set of *good* means. Using $\text{Tr}(\Sigma) = \mathbb{E}[\|X - \mu\|^2]$ and the Markov's inequality, we have $\mathbb{P}[\hat{\mu} \in G] \geq 1 - \frac{8}{13^2} \geq 0.95$. Applying the Chernoff bound, we get that with probability at least $1 - \delta$, we have $|G| \geq \frac{7}{10} \log \delta^{-1}$.

Consider ν_{CWM} , the coordinate-wise median of all the sample means. For each coordinate k , let L_k and R_k be the sets of sample means such that $\hat{\mu}_{i,k} \leq \nu_{\text{CWM},k}$ and $\hat{\mu}_{i,k} \geq \nu_{\text{CWM},k}$ respectively. We know that L_k, R_k have cardinality at least $\frac{b \log \delta^{-1}}{2}$. Depending on whether $\nu_{\text{CWM},k} > \mu_k$ or not, at least for one of L_k, R_k , we have that $|\hat{\mu}_{i,k} - \mu_k| \geq |\nu_{\text{CWM},k} - \mu_k|$.

We know that at least $\frac{7}{10}$ fraction of the means are *good* (i.e., $\|\hat{\mu}_i - \mu\| \leq r$) with probability at least $1 - \delta$. Hence, we infer that at least $\frac{7}{10} - \frac{1}{2} = \frac{1}{5}$ of the sample means are *good* and satisfy $|\hat{\mu}_{i,k} - \mu_k| \geq |\nu_{\text{CWM},k} - \mu_k|$. On average,

$$\frac{1}{|G|} \sum_{i \in G} |\hat{\mu}_{i,k} - \mu_k|^2 \geq \frac{|\nu_{\text{CWM},k} - \mu_k|^2}{5}.$$

Summing over all coordinates k gives

$$\sum_{k \in [d]} \frac{1}{|G|} \sum_{i \in G} |\hat{\mu}_{i,k} - \mu_k|^2 \geq \sum_{k \in [d]} \frac{|\nu_{\text{CWM},k} - \mu_k|^2}{5} = \frac{\|\nu_{\text{CWM}} - \mu\|^2}{5}.$$

Interchanging the sums on the left-hand side, we get

$$\|\nu_{\text{CWM}} - \mu\|^2 \leq \frac{5}{|G|} \sum_{i \in G} \sum_{k \in [d]} |\hat{\mu}_{i,k} - \mu_k|^2 = \frac{5}{|G|} \sum_{i \in G} \|\hat{\mu}_i - \mu\|^2 \leq 5r^2.$$

The theorem follows from the definition of $r = 13 \sqrt{\frac{\text{Tr}(\Sigma) \log \delta^{-1}}{N}}$. \square

519 D Generalization Bounds

520 We place our results in the context of generalization bounds for clustering problems. In this setting,
 521 we are given an arbitrary but fixed distribution \mathcal{D} supported on the unit Euclidean ball B_2^d . The cost
 522 of a point c with respect to \mathcal{D} is defined as $\text{cost}_{\mathcal{D}}(c) = \int_{p \in B_2^d} \|p - c\|^2 \cdot \mathbb{P}[p] dp$. The risk is defined
 523 as $\mathcal{R} := \underset{c}{\text{argmin}} \text{cost}_{\mathcal{D}}(c)$. Given independent samples S drawn from \mathcal{D} , we wish to compute an
 524 estimate \hat{c} with cost $\hat{\mathcal{R}} = \text{cost}_{\mathcal{D}}(\hat{c})$ such that the excess risk $\hat{\mathcal{R}} - \mathcal{R}$ is minimized. The cost of a
 525 distribution \mathcal{D} is in a limiting sense the average cost of a point set with all points living in B_2^d . Since
 526 the average cost is at most the squared radius (i.e. 1), obtaining a $(1 + \varepsilon)$ approximate solution also
 527 yields a solution that has excess risk of at most $\frac{\varepsilon \text{OPT}}{n} \leq \varepsilon$, which we then rewrite in terms of the
 528 sample size $|S|$ as $O(\frac{\log \delta^{-1}}{|S|})$. This discussion is summarized in the following corollary.

Corollary D.1. *Given a set of independent samples S drawn from some underlying arbitrary but fixed distribution supported on B_2^d , the geometric median-of-means estimator has an excess risk for the least squared distances objective of*

$$\hat{\mathcal{R}} - \mathcal{R} \leq \gamma \cdot \frac{\log \delta^{-1}}{|S|}$$

529 with probability $1 - \delta$ for some absolute constant $\gamma > 0$.

530 Note that the learning rate given by Corollary D.1 stands in contrast to learning rates that are
 531 achievable for other center-based problems such as indeed the geometric median with objective

$$\text{cost}_{\mathcal{D}}(c) = \int_{p \in B_2^d} \|p - c\| \cdot \mathbb{P}[p] dp,$$

532 or for the k -means problem with objective

$$\text{cost}_{\mathcal{D}}(C) = \int_{p \in B_2^d} \min_{c \in C} \|p - c\|^2 \cdot \mathbb{P}[p] dp$$

533 for a k -center set C . As mentioned in the related work section, all of these objectives require learning
 534 rates of at least $\sqrt{\frac{1}{|S|}}$, ignoring problem specific parameters.

535 E Lower Bounds

536 We complement our algorithmic results with matching lower bounds.

537 E.1 A High Probability Lower Bound for Mean-Estimation

538 **Theorem E.1.** *Any sublinear algorithm that outputs a $(1 + \varepsilon)$ -approximate Euclidean mean with
 539 probability at least $1 - \delta$ must sample at least $\Omega(\varepsilon^{-1} \log \delta^{-1})$ many points.*

540 The idea is to generate two instances that a sublinear algorithm for approximating means has to
 541 distinguish between and then bound the number of samples required to distinguish between such
 542 distributions. The first instance is virtually identical to the one used by Cohen-Addad et al. [2021].
 543 It places n points at 0 and εn points at 1. Thus, the optimal mean is $\frac{\varepsilon}{1 + \varepsilon}$. A routine calculation via
 544 Lemma 2.2 shows that the optimal cost is $\frac{\varepsilon n}{1 + \varepsilon}$. The second instance places all points at 0. Since 0 is
 545 not a sufficiently good approximate mean for the first instance, a sublinear algorithm can distinguish
 546 between the two instances, which yields a lower bound on the sample complexity.

547 The two instances lie within the unit Euclidean ball. By normalizing the number of points, it also
 548 yields a distribution supported on the unit Euclidean ball, which implies that the generalization
 549 bounds given in Corollary D.1 are sharp.

550 *Proof.* We give two instances. The first instance places n points at 0 and εn points at 1. Thus, the
 551 optimal mean is placed at $\frac{\varepsilon}{1 + \varepsilon}$. A routine calculation via Lemma 2.2 shows that the optimal cost is
 552 $\text{OPT} = \frac{\varepsilon n}{1 + \varepsilon}$. The second instance places all points at 0.

First we argue that a sublinear algorithm can distinguish between these two instances. Any approximate mean for the second instance must output 0. If we output 0 for the first instance, we incur a cost of $\varepsilon n = \frac{\varepsilon n(1+\varepsilon)}{1+\varepsilon} = \text{OPT} + \varepsilon \cdot \text{OPT}$, hence any algorithm improving over a $(1 + \varepsilon)$ approximation for the former cannot output 0. Thus, a necessary condition to distinguish between the two instances is for the algorithm to sample at least one point at 1.

Suppose we sample m points. Our goal is to show that for $m \in \Omega(\varepsilon^{-1} \log \delta^{-1})$, the probability that all of the sampled points are drawn from 0 is less than δ for the first instance. Let X_i denote the indicator variable of the i th sampled point. We have

$$\begin{aligned} \delta &\geq \mathbb{P}[\forall i, X_i = 0] = \mathbb{P}[X_1 = 0]^m = \left(\frac{1}{1+\varepsilon}\right)^m \\ &= \exp\left(m \ln \frac{1}{1+\varepsilon}\right) = \exp(-m \ln(1+\varepsilon)) \\ &> \exp(-m\varepsilon), \end{aligned}$$

where the final inequality follows from the Mercator series $\ln(1+\varepsilon) = \sum_{i=1}^{\infty} \left(\frac{\varepsilon^i}{i}\right) \cdot (-1)^{i+1}$. Thus, for any $m < \varepsilon^{-1} \cdot \log \delta^{-1}$, we do not achieve the desired failure probability. Conversely, $m \in \Omega(\varepsilon^{-1} \log \delta^{-1})$ for an algorithm to succeed. \square

E.2 The Empirical Mean is not a Good Estimator

We briefly show that the arguably most natural algorithm that outputs the empirical mean of a subsampled point set has a substantial increase in sample complexity in the high success probability regime and therefore also a substantial increase in running time compared to the algorithms presented in this paper. The result is likely folklore, but included for completeness.

Theorem E.2. *For all $\varepsilon, \delta \in (0, 1)$, there exists an instance such that $\Omega(\varepsilon^{-1} \delta^{-1})$ independently sampled points are required for the empirical mean to be a $(1 + \varepsilon)$ -approximate Euclidean mean with probability at least $1 - \delta$.*

The proof is based on reducing the problem of finding a good approximate mean with high probability to giving an example distribution for which the Chebychev inequality is tight.

Proof. Consider a sample $|S|$, with $|S|$ being specified later. We generate an instance as follows. We place a $\frac{1}{2|S|^2\varepsilon}$ -fraction of the points each at $-|S|\sqrt{\varepsilon}$ and $|S|\sqrt{\varepsilon}$ and the remaining $1 - \frac{1}{|S|^2\varepsilon}$ fraction at 0. Then the optimal solution places the mean at 0 and has average cost $(|S|\sqrt{\varepsilon})^2 \cdot \frac{1}{|S|^2\varepsilon} = 1$. By Lemma 2.2, this implies that the empirical mean $\hat{\mu} = \frac{1}{|S|} \sum_{p \in S} p$ is a better than $(1 + \varepsilon)$ -approximate mean if and only if $|\hat{\mu}| < \sqrt{\varepsilon}$. We set the failure probability, that is the probability $\mathbb{P}[|\hat{\mu}| \geq \sqrt{\varepsilon}] = \delta$. Then $\mathbb{P}[|\hat{\mu}| \geq \sqrt{\varepsilon}]$ is at least the probability that we sample exactly one point from either $-|S|\sqrt{\varepsilon}$ or $|S|\sqrt{\varepsilon}$ and the remaining $|S| - 1$ points from 0. Using Bernoulli's inequality and the density function of the binomial distribution, we therefore have

$$\delta = \mathbb{P}[|\hat{\mu}| \geq \sqrt{\varepsilon}] \geq |S| \cdot \frac{1}{|S|^2\varepsilon} \cdot \left(1 - \frac{1}{|S|^2\varepsilon}\right)^{|S|-1} \geq \frac{1}{|S|\varepsilon} \cdot \left(1 - \frac{1}{|S|\varepsilon}\right).$$

Solving $\frac{1}{|S|\varepsilon} \cdot \left(1 - \frac{1}{|S|\varepsilon}\right) \leq \delta$ for $|S|$ then yields $|S| \in \Omega(\varepsilon^{-1} \delta^{-1})$, as desired. \square

F Comparison of Geometric Median and the Coordinate-wise median

In this section, we give two instances which demonstrate that neither of the coordinate-wise median or the geometric median is strictly better than the other for the problem of mean-estimation.

Proposition F.1. *Let $\delta > 0$ be fixed. There exist instances such that, with probability at $1 - O(\delta)$,*

- (1) *The coordinate-wise median of $K = \frac{\log \delta^{-1}}{2(\frac{1}{2} - e^{-1})^2}$ sample means of ε^{-1} sampled points coincides with the true mean, but the geometric median of the same K sample means does not;*

582 (2) The geometric median of $\log \delta^{-1}$ sample means of ε^{-1} sampled points is a better approxi-
583 mation to the true mean than the coordinate-wise median is.

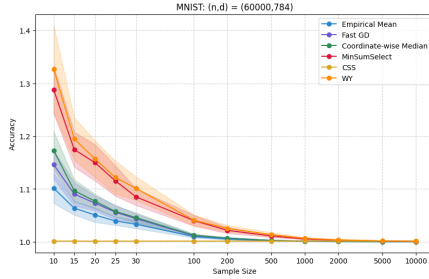
584 *Proof.* For part (1), we construct an instance for which the coordinate-wise median is a better
585 estimator than the geometric median. Consider the uniform distribution on the d -dimensional sim-
586 plex where $d = O(\varepsilon^{-2} \delta^{-2} \log^2 \delta^{-1})$. The value of d ensures that with probability at least $1 - \delta$,
587 each of the $\varepsilon^{-1} \log \delta^{-1}$ vertices sampled are distinct. We show that in this case the coordinate-
588 wise median is a better aggregation procedure than the geometric median. The coordinate me-
589 dian ν_{CWM} at the origin, while the true mean is $\mu = (1/d, 1/d, \dots, 1/d)$. The geometric me-
590 dian can be shown to lie at the empirical mean of all the samples due to symmetry, which is
591 $\mu_{\text{GM}} = (\varepsilon / \log(\delta^{-1}), \varepsilon / \log(\delta^{-1}), \dots, 0, \dots, 0)$ where there are exactly $\log(\delta^{-1}) / \varepsilon$ non-zero coord-
592 inates. Comparing the distance, we see that $\|\mu - \nu_{\text{CWM}}\| \leq \|\mu - \mu_{\text{GM}}\|$. To conclude, we note that
593 with probability at least $1 - \delta$, the coordinate-wise median is a better estimator than the geometric
594 median.

595 For part (2), we wish to prove that there exists an instance in which the geometric median is better than
596 the coordinate-wise median. Consider the uniform distribution on the d -dimensional simplex, where
597 $d = c\varepsilon^{-1}$. We choose c to be a constant large enough so that in every subsample, the probability of a
598 vertex sampled is $\frac{1}{3}$ (we only require it to be bounded away from $\frac{1}{2}$). As in the last example, we have
599 that the actual mean is $(1/d, 1/d, \dots, 1/d)$. We observe that the coordinate-wise median ν_{CWM} is the
600 origin with high probability. While as $\delta \rightarrow 0$, the geometric median tends towards true mean μ . \square

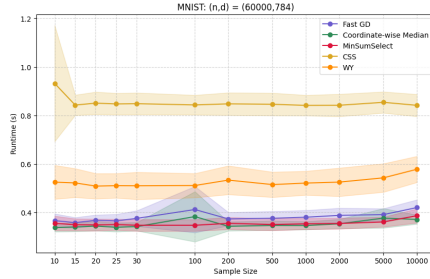
601 G Further Experimental Evaluation

Dataset	Shape	Mean	Std-Dev	Min	Max
MNIST	(60,000; 784)	0.1306	0.3081	0.0000	1.0000
Fashion-MNIST	(60,000; 784)	0.2860	0.3530	0.0000	1.0000
CoverType	(581,012; 54)	0.4567	0.4981	0.0000	1.0000

Table 1: Summary statistics for datasets.

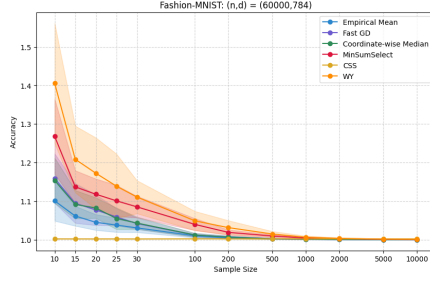


(a) Accuracy vs. sample size (in log-scale)

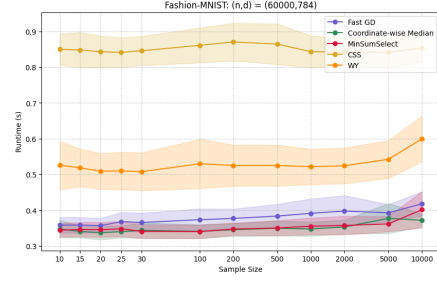


(b) Runtime vs. sample size

Figure 3: MNIST Dataset: Accuracy and runtime against sample size.

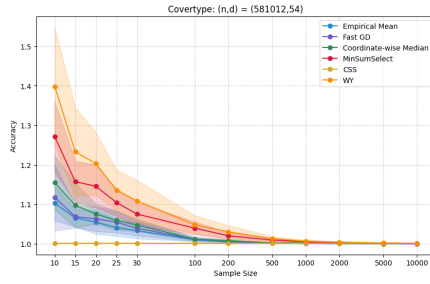


(a) Accuracy vs. sample size (in log-scale)

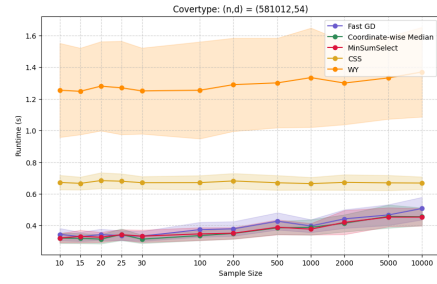


(b) Runtime vs. sample size

Figure 4: Fashion-MNIST Dataset: Accuracy and runtime against sample size.



(a) Accuracy vs. sample size (in log-scale)



(b) Runtime vs. sample size

Figure 5: CoverType Dataset: Accuracy and runtime against sample size.