Learning Entangled Single-Sample Gaussians

Flavio Chierichetti* Sapienza University Rome, Italy

Anirban Dasgupta Yahoo! Research Sunnyvale, CA

Ravi Kumar Google Mountain View, CA

Silvio Lattanzi Google Research New York, NY

Abstract

We introduce a new model of Gaussian mixtures, motivated by the setting where the data points correspond to ratings on a set of items provided by users who have widely varying expertise, and each user can rate an item at most once. In this mixture model, each item i has a true quality μ_i, each user has a variance (lack of expertise) σ_j^2, and the rating of a user j on an item i consists of a *single* sample independently drawn from the Normal distribution $N(\mu_i, \sigma_j^2)$. The aim is to learn the unknown item qualities μ_i's as precisely as possible. We study the single item case and obtain efficient algorithms for the problem, complemented by near-matching lower bounds; we also obtain preliminary results for the multiple items case.

1 Introduction

Consider the following setting. We have a set of items where each item has a true quality that we wish to determine. We also have a set of users with varying levels of expertise, where the expertise of each user is typically unknown. These users can provide their rating of the quality of items, where each item is rated by several or possibly all users. Given such ratings, can we hope to learn the quality of the items? Questions of this type are at the heart of crowdsourcing designed to solve computationally challenging problems: for example, the items can be images of car license plates, the quality can be the actual plate number, and each user's assessment is their individual perception of the number when shown the image. Crowdsourcing is already a billion-dollar business (especially with incarnations such as the Amazon Mechanical Turk crowdsourcing platform) and has entered the research mainstream [6, 10, 15, 19, 22–24, 26–28].

There have been some attempts to model this estimation problem in a formal framework with provable guarantees. Ghosh, Kale, and McAfee [12] modeled it in the following manner: each item i has a binary quality q_i, each user j has a probability p_j of error, and a user j rates an item i as q_i with probability $1-p_j$ and as $1-q_i$ with probability p_j. They used a spectral algorithm to provably reconstruct the item qualities when the user–item graph is either complete or random; their algorithm was extended to the arbitrary graph case by Dalvi et al. [8]. A similar model was also considered by Karger, Oh, and Shah [18], who proposed a belief-propagation based algorithm that worked when the user–item graph is sparse and the item qualities are random. While there have been several heuristics for crowdsourcing problems, progress in rigorous analyses has been relatively modest. For example, nothing is known about extending the above models to the non-binary case, which appears to be non-obvious.

The primary motivation for our work stems from such crowdsourcing settings. In our case, each item i has a *quality* μ_i and each user j has (a lack of) *expertise* σ_j. When the user j rates the item i, the rating is independently drawn from the normal distribution $N(\mu_i, \sigma_j^2)$. Notice that besides being non-binary, modeling in this way naturally captures the intuition that what makes an expert is the consistent ability to closely estimate the true quality of items. The question then is, given such a matrix of ratings, can we estimate the unknown μ_i's?

This question can also be thought of as a version of learning a mixture of Gaussians. Unlike crowdsourcing, the literature on learning mixtures (in particular, learning Gaussian mixtures) is rich and illustrious. Starting from the work of Dasgupta [9], there has been steady progress in algorithms for learning mixtures of Gaussians [1, 2, 17, 25], including the very recent algorithms based on the method of moments [3, 4, 16, 21]. Our model, however, cannot directly benefit from these results for three main reasons:

(i) We only obtain at most *one* sample from each Gaussian. This restriction is to capture the

*Supported by a Google Faculty Research Award and by the MULTIPLEX project (EU-FET-317532).

crowdsourcing situation where it does not make sense for a user to rate an item more than once. On the other hand, algorithms for Gaussian mixture models require several examples from the same Gaussian.

(ii) In Gaussian mixture learning, the number of Gaussians is assumed to be *small*, since the complexity is at least singly exponential in this number; we do not want to impose this restriction.

(iii) Even though there are mn Gaussians, where m is the number of items and n is the number of users, there are only $m + n$ underlying parameters to learn, i.e., the Gaussians even though independent are parameter-wise *entangled*.

One of the issues that will concern us in developing algorithms with provable bounds is the dependence of the error on the sequence of variances of the users. In a crowdsourced setting, it is natural to have a wide mixture of expertise, implying that there could be a very large skew in the values σ_j. Robust estimators for means and other central moments have been studied before [14, 20], but their aim is mainly to deal with non-normal data, as well as explicit outliers. There has also been work trying to characterize the "relative efficiency" of mean estimators compared to the variance of the optimal estimator, given by the Cramer–Rao lower bound [7]; however we are unaware of any result bounding the efficiency of estimators for Gaussian mixture models.

Main results. We begin by studying the case where the user variances $\sigma_1 \leq \cdots \leq \sigma_n$ are known. Here, we show that the minimum (expected) additive loss on the reconstruction of the unknown item qualities is $\Theta(1/\sqrt{\sum_{i=1}^{n} \sigma_i^{-2}})$ (Theorem 3.1); this is tight. We then move to the more interesting unknown variances case. First, we observe that it is generally not possible to guess who is the user with the smallest variance; this highlights the difficulty of working with just a single sample from each Gaussian. Thus, we cannot bound our loss in terms of the smallest variance σ_1; instead we bound it in terms of $\sigma_2, \ldots, \sigma_n$.

We then focus on the single-item case ($m = 1$). We present three simple and natural algorithms, namely, Arithmetic Mean, k-Median, k-Shortest Gap, and analyze their performance. None of these algorithms dominates the other two in terms of the expected loss. Our main algorithmic contribution is that the k-Median and the k-Shortest Gap algorithms can be carefully combined to produce a loss of $\tilde{O}(\sqrt{n}\sigma_{\log n})$ (Theorem 4.2). We then show that in the worst-case scenario this is near-optimal: in fact there exists instances where a loss of $n^{1/2-\epsilon}\sigma_k$ (Lemma 4.5), for k up to poly(n), is necessary even when there is a *polynomial* number of users each having the same smallest variance σ_1. To prove this near-optimality result we analyze the anti-concentration properties of a rather complex random variable, using Berry–Esseen-style bounds. The intuition behind those surprising, and slightly counterintuitive, results (one may expect the error to decrease with the number of samples) is that by adding several high variance raters, an adversary can hide the signal of good raters in the noise generated by the bad raters.

Finally, we consider the multi-item case ($m > 1$). We show that the loss (on the guessed quality of each specific item) can decrease dramatically from the single-item case. We consider both a complete-graph setting (where each user rates every item), and a Erdös–Renyi-like setting (where user i rates item j with probability α). We show that, in these cases, if we disregard the contribution of the user with smallest variance, we can achieve the optimal performance of the known-variance case (Theorem 5.3).

2 Preliminaries

Let $N(\mu, \sigma^2)$ denote the Normal distribution with mean μ and standard deviation σ. For normally distributed random variables, we have the following three properties:

$$aN(\mu, \sigma^2) + b = N(a\mu + b, a^2\sigma^2), \tag{2.1}$$

for any real a, b, and

$$N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2), \tag{2.2}$$

if the two normal variables in the LHS are independent, and

$$E[|N(\mu, \sigma^2) - \mu|] = \sigma\sqrt{\frac{2}{\pi}}. \tag{2.3}$$

For a vector u, let $\|u\|$ denote its 2-norm and let $x \sim D$ denote that the random variable x has the distribution D. Let $\mu = \langle \mu_1, \ldots, \mu_m \rangle$ and $\sigma = \langle \sigma_1, \ldots, \sigma_n \rangle$. We assume throughout that the maximum ratio between the σ's is bounded by a polynomial, that is: $\max_{1 \leq i < j \leq n} \sigma_i/\sigma_j = O(\text{poly}(n))$.

Consider an $m \times n$ matrix where each entry is either empty or contains $x_{ij} \sim N(\mu_i, \sigma_j^2)$, independently generated. In the *entangled Gaussians problem*, both μ and σ are unknown and given a matrix as above, the goal is to output $\hat{\mu}_i$ for $1 \leq i \leq m$ in such a way that the *loss* $E[|\mu_i - \hat{\mu}_i|]$ is minimized. Here, the expectation is taken over the generation of the matrix. We often consider the special case of $m = 1$; in this case, let μ denote the unknown mean.

Recall that the error function is related to the cumulative distribution function of the Normal distribution as

$$\Pr[N(0,1) \le x] = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).$$

From now on, "w.h.p" will mean that the statement holds with probability at least $1 - 1/\text{poly}(n)$, where n will be clear from the context.

3 Warm-up cases

In this section we consider two important special cases of the problem for $m = 1$: the first is when the variances are actually known to the algorithm and the second is when the variance values are not known but something about their skew is known. The intuition obtained in these settings will be useful in the general case. For $j = 1, \ldots, n$, let $x_j \sim N(\mu, \sigma_j^2)$ be independently generated. Recall that we are given the x_j's and the goal is to output $\hat{\mu}$ to minimize $E[|\mu - \hat{\mu}|]$.

For the case when the algorithm knows the σ_j's, we can get optimal bounds as we show below.[1] Let $P(\sigma) = \left(\sum_{j=1}^n \frac{1}{\sigma_j^2}\right)^{-1/2}$.

THEOREM 3.1. *Define* $\hat{\mu} = \left(\sum_{j=1}^n \frac{x_j}{\sigma_j^2}\right) P^2(\sigma)$. *The loss satisfies* $E\left[|\hat{\mu} - \mu|\right] = \Theta(P(\sigma))$. *Moreover, this loss is tight for any algorithm.*

Proof. From (2.1) and (2.2), the random variable $\hat{\mu}$ has the distribution

$$\frac{\sum_{j=1}^n \frac{N(\mu,\sigma_j^2)}{\sigma_j^2}}{\sum_{j=1}^n \frac{1}{\sigma_j^2}} = N\left(\mu, \frac{1}{\sum_{j=1}^n \frac{1}{\sigma_j^2}}\right) = N\left(\mu, P^2(\sigma)\right);$$

and the loss bound then follows from (2.3). Note that the expression for $\hat{\mu}$ appropriately down-weights samples from the high-variance Gaussians; in fact, $\hat{\mu}$ is the solution to the associated maximum likelihood problem.

For the lower bound, let $P = P(\sigma)$ and let μ be a random variable uniform in $\{\pm\epsilon P\}$, where $\epsilon > 0$ is a constant. Observe that to incur a loss smaller than ϵP (which is obtained by returning $\hat{\mu} = 0$), an algorithm has to correctly infer μ. Let $\mathcal{L}_-$ be the likelihood if $\mu = -\epsilon P$ and $\mathcal{L}_+$ be the likelihood if $\mu = \epsilon P$:

$$\mathcal{L}_+ = \prod_{j=1}^n \left(\frac{1}{\sigma_j\sqrt{2\pi}} \cdot e^{-\frac{(x_j - \epsilon P)^2}{2\sigma_j^2}}\right),$$

and

$$\mathcal{L}_- = \prod_{j=1}^n \left(\frac{1}{\sigma_j\sqrt{2\pi}} \cdot e^{-\frac{(x_j + \epsilon P)^2}{2\sigma_j^2}}\right).$$

We now compute the distribution of $\ln(\mathcal{L}_-/\mathcal{L}_+)$ as

$$\sum_{j=1}^n \frac{(x_j + \epsilon P)^2 - (x_j - \epsilon P)^2}{2\sigma_j^2} = \sum_{j=1}^n \frac{4x_j\epsilon P}{2\sigma_j^2} = 2\epsilon P \sum_{j=1}^n \frac{x_j}{\sigma_j^2}.$$

Since each x_j is either $N(\epsilon P, \sigma_j)$ or $N(-\epsilon P, \sigma_j)$ with equal probability, from (2.1) and (2.2), and using the fact that $P^{-2}(\sigma) = \sum_j \frac{1}{\sigma_j^2}$, it follows that $\ln(\mathcal{L}_-/\mathcal{L}_+)$ is a uniform mixture of $N(2\epsilon^2, 4\epsilon^2)$ and $N(-2\epsilon^2, 4\epsilon^2)$. Thus, regardless of the original choice of μ,

$$\Pr\left[\mathcal{L}_+ > \mathcal{L}_-\right] = \frac{1}{2} \pm \Theta(\epsilon).$$

It follows that (even if it knows the mapping between the σ_j's and the x_j's), the loss of the algorithm has to be at least cP, for some constant $c > 0$. ■

We then consider the case where the variance skew is known and illustrate two algorithms for this problem. Even though these two algorithms will be suboptimal, the insights derived from these will be useful to obtain a near-optimal algorithm. First we present a natural algorithm that just outputs the arithmetic mean of the given samples.

DEFINITION 1. (ARITHMETIC MEAN ALGORITHM) *Output* $\hat{\mu} = \frac{1}{n}\sum_{j=1}^n x_j$.

We next show an easy bound on the loss of this algorithm.

LEMMA 3.1. *The Arithmetic Mean algorithm has loss* $\Theta(\|\sigma\|/n)$.

Proof. From (2.1) and (2.2), we have $\hat{\mu} \sim N(\mu, \|\sigma\|^2/n^2)$. The proof follows from (2.3).

Clearly, the loss depends on the skew of the σ_j's. Next, we present another natural algorithm that outputs an arbitrary point in the minimal interval of the samples.

DEFINITION 2. (SHORTEST GAP ALGORITHM) *Let* $\{j, j'\} = \arg\min_{j \ne j'} |x_j - x_{j'}|$ *and output* $\hat{\mu}$ *to be any point in the interval* $[\min(x_j, x_{j'}), \max(x_j, x_{j'})]$.

We now state a bound on the loss of this algorithm, provided the variances satisfy some properties. Without loss of generality, let $\sigma_1 \le \cdots \le \sigma_n$.

[1] This is probably folklore in Statistics, see e.g. [13]. We provide the proof for completeness.

LEMMA 3.2. *The Shortest Gap algorithm has loss* $\Theta(\sigma_2)$ *if, for all* $j \geq 3$*, we have* $\sigma_j = \omega\left(\sigma_2 \cdot j^2 \log^{1+\epsilon} j\right)$.

Proof. The basic idea behind the proof is that under such a skew, the closest pair will consist of the points having the smallest two variances, none of the other points will have someone close w.h.p.

Since x_j's are normally distributed and since $\sigma_1 \leq \sigma_2$, we have $\Pr[\max(|x_1-\mu|, |x_2-\mu|) = \omega(\sigma_2)] = o(1)$. Next, consider any $x_j, j \geq 3$. We compute the probability p_j that there exists a $j' < j$ such that $|x_j - x_{j'}| < |x_1 - x_2|$. We first compute the probability $p_{jj'}$ defined to be the probability that $|x_j - x_{j'}| < |x_1 - x_2|$ for a fixed $j' < j$.

$$\begin{aligned}
p_{jj'} &= \Pr[|x_j - x_{j'}| < |x_1 - x_2|] \\
&= \Pr\left[|N(0, \sigma_j^2 + \sigma_{j'}^2)| < |N(0, \sigma_1^2 + \sigma_2^2)|\right] \\
&= \Pr\left[|N(0,1)| < \left|N\left(0, \frac{\sigma_1^2 + \sigma_2^2}{\sigma_j^2 + \sigma_{j'}^2}\right)\right|\right] \\
&\quad \text{(on dividing both variables by } (\sigma_j^2 + \sigma_{j'}^2)^{1/2}\text{)}, \\
&\leq \Pr\left[|N(0,1)| < \left|N\left(0, \frac{\sigma_2^2}{2\sigma_j^2}\right)\right|\right] \\
&\quad \text{(since } \frac{\sigma_2^2}{2\sigma_j^2} \leq \frac{\sigma_1^2 + \sigma_2^2}{\sigma_j^2 + \sigma_{j'}^2}\text{)}, \\
&\leq \Theta\left(\frac{\sigma_2}{\sigma_j}\right)\Theta(1) = \Theta\left(\frac{\sigma_2}{\sigma_j}\right),
\end{aligned}$$

where the last inequality follows since $\frac{\sigma_2}{\sigma_j} < 1$. Hence we can now bound,

$$p_j = \sum_{j'<j} p_{jj'} \leq (j-1)\Theta\left(\frac{\sigma_2}{\sigma_j}\right) = o\left(\frac{1}{j \log^{1+\epsilon} j}\right).$$

By the union bound, the probability that the minimum distance between two samples is less than $|x_1 - x_2|$ is at most

$$\sum_{j=3}^{n} p_j \leq \sum_{j=1}^{\infty} o\left(\frac{1}{j \log^{1+\epsilon} j}\right) = o(1).$$

Thus, the shortest gap will be the one induced by x_1 and x_2 with probability $1 - o(1)$. Therefore, the loss of the shortest gap algorithm is $O(\sigma_2)$. ■

Unfortunately, it is easy to see that in general none of the two algorithms dominates the other. Suppose $\sigma_1 = \cdots = \sigma_n = \sigma$. Then, the Arithmetic Mean algorithm will have loss $\Theta(\sigma/\sqrt{n})$, whereas, the Shortest Gap algorithm will have a much larger loss of $\Theta(\sigma)$, since the shortest interval has a roughly equal probability of happening in any sub-interval of $[-\sigma, \sigma]$. In contrast, consider an instance where $\sigma_1 < \sigma_2$ and $\sigma_j = \omega(\sigma_2) \cdot j^2 \log^{1+\epsilon} j$ for some small $\epsilon > 0$. The loss of the Arithmetic Mean algorithm on this instance will be $\omega\left(\frac{\sigma_2}{n}\sqrt{\sum_{j=1}^{n} j^4 \log^{2+2\epsilon} j}\right) = \omega\left(\sigma_2 \cdot n^{\frac{3}{2}} \log^{1+\epsilon} n\right)$, whereas, the Shortest Gap algorithm will have a loss of just $O(\sigma_2)$. Thus, these algorithms have differing performance depending on the skew of the σ_j's; an optimal algorithm will have to take this skew into consideration.

4 Single item

In this section we consider the single-item case ($m = 1$) and present a near-optimal algorithm. For simplicity we first study a stylized version where the set of variances is known to the algorithm but the algorithm does not know the mapping between the variances and the actual samples. Then we will see how the guarantees of this algorithm can extended also to the case where the variances are unknown. Finally, we show the optimality of this algorithm by establishing a lower bound.

4.1 Algorithms We present an algorithm that is a combination of two simple algorithms that are each intuitive. The first of these is the k-Median algorithm that selects the "central" k points and then creates an estimate. The second is the k-Shortest Gap algorithm, that, instead of selecting the k central points, selects the smallest interval that contains k points and reports a estimate from this interval. We will show some nice properties for each of these algorithms and show how to judiciously combine them to get an estimate with provable theoretical guarantees for arbitrary variances.

DEFINITION 3. (k-MEDIAN ALGORITHM) *Given a set* $\{x_1, \ldots, x_n\}$ *of samples, output a subset* S *such that* $\forall x_i \in S$, x_i *is bigger than at least* $n\left(1 - \frac{k}{2n}\right)$ *other samples and there are at least* $n\left(1 - \frac{k}{2n}\right)$ *other samples that are bigger than* x_i.

This algorithm can be easily implemented in $O(n \log n)$ time and has the following nice property.

LEMMA 4.1. *Let* S *be the set returned by the* $4\sqrt{cn \log n}$*-Median algorithm on* n *samples. Let* I_S *be the interval defined by* S*. Then with probability at least* $1 - n^{-\Theta(c)}$*,* I_S *contains* μ*, and (b) for any* $0 \leq k \leq \sqrt{cn \log n}$*,* S *contains the closest* k *samples to the right and to the left of* μ.

Proof. Note that every sample is equally likely to be on the right or on the left of μ. Thus for any polynomial in n, by the Chernoff bound we can conclude that with probability at least $1 - n^{-\Theta(c)}$, the difference between the number of samples on the left of μ and on the right of μ is bounded by $\sqrt{cn \log n}$. Thus, with at least the same probability, S will include at least $\sqrt{cn \log n}$ elements from the right and from the left of μ. ■

The second ingredient is the k-Shortest Gap algorithm, a generalization of Definition 2.

DEFINITION 4. (k-SHORTEST GAP ALGORITHM) *Given a set $\{x_1, \ldots, x_n\}$ of samples, output $\{i_1, \ldots, i_k\} \subset [n]$ that minimizes $\max_{j,\ell \in [k]} |x_{i_j} - x_{i_\ell}|$.*

This algorithm can also be implemented in $O(n \log n)$ time, by first sorting the elements and then by looking for the shortest gap. Now we are ready to combine the two algorithms.

Algorithm *(known variances, unknown assignment)*

Input: n samples of distinct Gaussians with the same mean μ but different variances, the variances are known, but the assignment of the variances to the samples is not known.

Output: An estimate $\hat{\mu}$ for the mean μ.

Compute a set S using the $4\sqrt{cn \log n}$-Median algorithm.

Let s be the size of the interval spanned by the samples in S.

if $s < \sqrt{n} \cdot \sigma_{\log n} \cdot \log^4 n$ **then**

  Output any point in S.

else

  Let T be the output of the $\log n$-Shortest Gap algorithm on S.

  Output any point in T.

Note that the previous algorithm has running time $O(n \log^2 n)$. The main intuition behind the combined algorithm is that the k-Median points help us disregard the samples from the high variance Gaussians. The k-Shortest Gap algorithm then returns a good estimate since the density function is mostly concentrated around the mean. Before stating the theoretical guarantees of the above algorithm, we show few useful properties of each of the two component algorithms. First, we show that the k-Shortest Gap inside the median set S is small.

LEMMA 4.2. *Let S be the set returned by the $4\sqrt{cn \log n}$−Median algorithm on n samples with standard deviations $\sigma_1 \leq \cdots \leq \sigma_n$. Let I_S be the interval defined by S. Then with probability at least $1 - n^{-\Theta(c)}$, for any $0 \leq k \leq \log n$, I_S contains an interval that has k samples and is of size at most $2c\sigma_k \log n$.*

Proof. First note that with high probability the samples $x_1, \ldots, x_k$ corresponding to $\sigma_1, \ldots, \sigma_k$ are in an interval of size at most $2\sigma_k \log n$. Indeed, for each of those samples, the probability of not lying in the interval $[\mu - c\sigma_k \log n, \mu + c\sigma_k \log n]$ is $1 - \text{erf}\left(\frac{c \log n}{\sqrt{2}}\right) < \frac{e^{-(c \log n/\sqrt{2})^2}}{\frac{c \log n}{\sqrt{2}}\sqrt{\pi}} = \Theta\left(\frac{n^{-c^2 \log n/2}}{c \log n}\right)$. Thus for each $k \leq \log n$, by the union bound we have that w.h.p., the samples $x_1, \ldots, x_k$ are at most at distance $2c\sigma_k \log n$.

Furthermore by selecting the $4\sqrt{cn \log n}$ median elements we have by Lemma 4.1 that with probability at least $1 - n^{-\Theta(c)}$, μ and k points on the left and on the right of μ are inside the interval defined by S. So we have that the distance between $x_1, \ldots, x_k$ is an upper bound on the distance between the closest k points in S; thus the claim follows. ■

Now we prove a proposition showing that the probability that k samples with high standard deviation are close is low. This shows that the k-Shortest Gap cannot consist of high standard deviation samples.

LEMMA 4.3. *Let $x_1, \ldots, x_{4\sqrt{cn \log n}}$ be samples with standard deviation at least $n^{1/2(1+1/(k-1))}\sigma_k \log^3 n$. Then with probability $1 - (\log n)^{-\Theta(k)}$ there is no interval of size $4c\sigma_k \log n$ that contains k samples even if we conditioned all the samples to be in any given interval of size at most $n^{1/2(1+1/(k-1))}\sigma_k \log^4 n$.*

Proof. The general idea of the proof is to show that the probability that any interval of size $4c\sigma_k \log n$ centered around any sample contains at least other $k - 1$ samples is $\Theta(n^{-1/2} \log^{-3/2} n)$. Thus by the union bound on the number of samples, we will get the claim. Let I be the interval $[\mu - n^{1/2(1+1/(k-1))}\sigma_k \log^4 n, \mu + n^{1/2(1+1/(k-1))}\sigma_k \log^4 n]$.

First note that for any interval of size $4c\sigma_k \log n$, the probability of containing $k-1$ samples is smaller than the probability q that the interval of size $4c\sigma_k \log n$ centered in μ contains the same $k-1$ samples. This property holds even if the samples are all conditioned to lie in some superset of the interval I. Furthermore note that in this setting the probability q is maximized when all the samples have standard deviation equal to $n^{1/2(1+1/(k-1))}\sigma_k \log^3 n$. Thus to upper bound the probability that an interval of size $4c\sigma_k \log n$ contains $k - 1$ samples, we simply need

to compute the probability that the interval of size $4c\sigma_k \log n$ centered around μ contains $k-1$ samples, when we have $4\sqrt{cn \log n}$ samples that all have standard deviations equal to $n^{1/2(1+1/(k-1))}\sigma_k \log^3 n$, and are conditioned to be in some superset of I.

Note the probability that a sample of standard deviation $n^{1/2(1+1/(k-1))}\sigma_k \log^3 n$ is in I is at least $1/2 - o(1)$. The unconditioned probability for a single sample to be in $[\mu - 2c\sigma_k \log n, \mu + 2c\sigma_k \log n]$ is $\text{erf}\left(\frac{\sqrt{2}c}{n^{1/2(1+1/(k-1))}\log^2 n}\right) \in \Theta\left(\frac{c}{n^{1/2(1+1/(k-1))}\log^2 n}\right)$. Thus, conditioned on being in some superset of I, the probability for a single sample to be in $[\mu - 2c\sigma_k \log n, \mu + 2c\sigma_k \log n]$ is at most $\Theta\left(\frac{c}{n^{1/2(1+1/(k-1))}\log^2 n}\right)$.

Thus the probability that $k-1$ out of $4\sqrt{cn \log n}$ samples are in the interval is smaller than

$$\binom{4\sqrt{cn\log n}}{k-1}\Theta\left(\frac{c}{n^{1/2(1+1/(k-1))}\log^2 n}\right)^{k-1} = 2^{\Theta(c)} n^{-1/2} \log^{-3/2(k-1)} n.$$

By taking a union bound over the $4\sqrt{cn \log n}$ intervals centered at each of the points, we get the claim. ■

Finally, we use the above results to prove bounds on the above algorithm that combines the k-Shortest-Gap and k-Median algorithms.

THEOREM 4.1. *The estimate $\hat{\mu}$ returned by the above algorithm has a loss $E[|\mu - \hat{\mu}|] = \tilde{O}(\sqrt{n} \cdot \sigma_{\log n})$.*

Proof. Let I be $I = [\mu - cn^{1/2}\sigma_{\log n}\log^4 n, \mu + cn^{1/2}\sigma_{\log n}\log^4 n]$, S contains the $4\sqrt{cn\log n}$-Medians, and T be the interval containing the $\log n$-Shortest Gap points contained in S.

We divide our analysis in two parts. First we prove that if in the set T there is at least one sample with standard deviation smaller than $n^{1/2}\sigma_{\log n}\log^3 n$, then we have that with probability at least $1 - n^{-\Theta(c)}$, the returned point has a loss of at most $O(cn^{1/2}\sigma_{\log n}\log^4 n)$. Then, we show that, with probability at least $1 - n^{-\Theta(c)}$, either S is contained in I, or T contains some points with standard deviation less than $n^{1/2}\sigma_{\log n}\log^3 n$. We will show that in both cases the claim follows.

First note the probability that a sample with standard deviation smaller than $n^{1/2}\sigma_{\log n}\log^3 n$ lies outside the interval I is upper bounded by $1 - \text{erf}\left(\frac{c\log n}{\sqrt{2}}\right) < \sqrt{\frac{2}{\pi}}\frac{n^{-c^2\log n/2}}{\log n} \in o\left(n^{-c^2}\right)$, and thus by the union bound, no point with standard deviation smaller than $n^{1/2}\sigma_{\log n}\log^3 n$ lies outside the interval I with probability at least $1 - n^{-\Theta(c)}$. Now, by combining this with Lemma 4.2, we get that with the same probability, if a sample with standard deviation smaller than $n^{1/2}\sigma_{\log n}\log^3 n$ is in T, then the loss is at most $\tilde{O}(n^{1/2}\sigma_{\log n})$.

To finish the proof, we will show that if the set S is not contained in I then with probability at least $1 - n^{-\Theta(c)}$, the $\log n$-Shortest Gap contains some point with standard deviation smaller than $n^{1/2}\sigma_{\log n}\log^3 n$. By the Chernoff bound, S contains μ. If by assumption S is not entirely contained in I, it must have size at least $n^{1/2}\sigma_{\log n}\log^4 n$. Lemma 4.3 then shows that, with probability at least $1 - n^{-\Theta(c)}$, the $\log n$-Shortest Gap cannot exclusively consist of points with standard deviation larger than or equal $n^{1/2}\sigma_{\log n}\log^3 n$. Hence T must contain some points with standard deviation less than $n^{1/2}\sigma_{\log n}\log^3 n$ and the claim follows.

Now, the probability that one or more of the (constantly many) above events does not happen is at most $n^{-\Theta(c)}$. Since the algorithm never returns a point at expected distance more that $O(\sigma_n)$ from μ, the expected loss of the algorithm, if some of those events does not happen, is at most $\sigma_n \cdot n^{-\Theta(c)} \leq \sigma_1 \cdot \text{poly}(n) \cdot n^{-\Theta(c)}$. Since we can choose c to be any positive constant, we have that the loss of the algorithm is at most $o(\sigma_1)$ if some of the bad events happen. ■

We now turn our attention to the the case in which the variances are unknown, where we show that a very similar approach leads to a similar bound. The main intuition behind the algorithm is that even if we do not know the variances, we can still use the length of the k-Shortest Gap interval as a proxy for it. In fact by Lemma 4.2 we have that the length of the k-Shortest Gap is always smaller than $\sigma_k \log n$. Furthermore in the proof of Lemma 4.1, we use the standard deviation only to bound the size of the k-Shortest Gap. So in this setting we can use essentially the same algorithm after substituting the length of the k-Shortest Gap interval in place of σ_k.

Algorithm *(unknown variances)*
Input: n samples of distinct Gaussians with the same mean μ but different variances.
Output: An estimate $\hat{\mu}$ for the mean μ.
Compute the set $S = 3\sqrt{n \log n}$-Median algorithm.
Let s be the size of the interval spanned by the samples in S.
For $2 \leq k \leq \log n$ let l_k be the length of the k-Shortest Gap interval over S.
if $s < n^{1/2} l_{\log n} \log^4 n$ **then**
  Output any point in S.

else

Let T be the output of the $\log n$-Shortest Gap algorithm on S.

Output any point in T.

The following theorem bounds the loss of the above algorithm.

THEOREM 4.2. *In the setting that variances are all unknown the above algorithm that uses $4\sqrt{cn\log n}$-Median followed by $\log n$-Shortest Gap outputs $\hat{\mu}$ such that $E[|\mu-\hat{\mu}|]=\tilde{O}(n^{1/2}\sigma_{\log n})$.*

Proof. The proof is identical to the proof of Lemma 4.1. The only difference is that we use the fact that inside S the shortest k interval has size l_k instead of $4c\sigma_k\log n$. The proof then follows by applying Lemma 4.2. ■

4.2 Lower bounds We finally turn our attention to the lower bounds where we prove that the presented algorithm for the unknown variance case is almost tight from a worst-case perspective. First, we observe that it is easy to come up with instances such that the reconstruction loss in the unknown variance case is strictly worse than in the known variance case. For instance, if we only have two ratings from two different users, $x_1 \sim N(\mu,\sigma_1)$, $x_2 \sim N(\mu,\sigma_2)$, with $\sigma_1 \ll \sigma_2$, then the known variance algorithm, by Theorem 3.1, gives us a loss of σ_1. However, in the unknown variance case, it is not possible to identify which of the sample has the smaller variance, and hence the best one could do is the obtain a $O(\sigma_2)$ expected loss (i.e., by returning any of the two points, or their average). Thus, in general, the ratio between the loss in the unknown and known variance case can be as large as $\frac{\sigma_2}{\sigma_1}$, which can be unbounded.

The above lower bound, though, simply follows from the fact that the known variance algorithm might effectively use only one sample, while this sample might not be identifiable without actually knowing the variances, leading to an unbounded ratio. In the rest of the section, we present a more general lower bound, that shows the gap between the unknown and known variance cases in a more general setting, e.g., when none of the variances is much smaller than the others (e.g. when every variance value is used by a multitude of samples).

First though, recall that the algorithm that we presented for the unknown variance case had a loss upper bounded by $\tilde{O}(n^{1/2+o(1)}\sigma_{\log n})$. On the other hand, the loss in the known variance case is never larger than σ_i, for each $i\in[1,n]$. Therefore, there is (at least) a $\tilde{\Omega}(n^{1/2+o(1)})$ gap between the loss of our unknown variance and the (optimal) known variance algorithm. We show in this section that such a gap is, in general, necessary (even if there are many samples with the same minimum variance).

We consider the following family of instances described for our lower bound.

EXAMPLE 1. *Choose the number of samples $n>0$, and choose $p\in\left(0,\frac{1}{\sqrt{2n}}\right)$. For each $i=1,\ldots,n$, let σ_i be chosen iid according to the following distribution: with probability p, let it be equal to p^2n and with probability $1-p$, let it be equal to $1-p$.*

The loss of the known-variance algorithm on this instance can be bounded by Theorem 3.1 as follows.

LEMMA 4.4. *Assuming that $p=\Omega\left(\frac{\log n}{n}\right)$, that $p=O\left(n^{-2/3}\right)$, and that the standard deviations σ_i's are known, the loss $E[|\mu-\hat{\mu}|]$ obtained by using the estimate $\hat{\mu}$ in Theorem 3.1 on Example 1 is $\Theta\left(p^{3/2}n^{1/2}\right)$.*

Proof. With probability $1-n^{-\Theta(1)}$ we have that the number of Gaussians with standard deviation p^2n is at least cpn, for some constant $c>0$. Therefore, the loss L suffered by the estimate in Theorem 3.1 satisfies: $L=\Theta\left(\sqrt{\frac{1}{\sum_{i=1}^n\sigma_i^{-2}}}\right)=\Theta\left(\sqrt{\frac{1}{p^{-3}n^{-1}+n}}\right)$. If $p=O\left(n^{-2/3}\right)$, we have $p^{-3}n^{-1}\geq\Omega(n)$, and therefore $L=\Theta\left(p^{3/2}n^{1/2}\right)$. ■

We next show a lower bound for the case when the variances are not known. The basic intuition for the lower bound follows the idea of the lower bound in the known variance case in Theorem 3.1. We will set $L=\Theta\left(n^{-1/2}\right)$ and set the mean of the above instance to be L or $-L$ with equal probability. We will compute the ratio of the likelihoods and show that in both these cases, the other choice of mean has a higher likelihood with constant probability. This will imply a lower bound of L on the loss of the unknown variance case.

LEMMA 4.5. *Assuming that $p=\Omega\left(\frac{\log n}{n}\right)$, that $p=o(n^{-1/2})$, and that the standard deviations σ_i's are unknown, the loss $E[|\mu-\hat{\mu}|]$ suffered by any algorithm on Example 1 is at least $\Omega\left(\frac{1}{\sqrt{n}}\right)$.*

Proof. Let μ be chosen uniformly at random in $\{-L,L\}$, for $L=\frac{\epsilon}{\sqrt{n}}$, for an unspecified constant $\epsilon>0$. Let x_i denote the set of samples drawn from this instance. Given the samples x_i, the likelihood of

the mean being L is given as

$$\mathcal{L}_+ = (2\pi)^{-n/2} \cdot \prod_{i=1}^{n} \left(\frac{p}{p^2 n} \cdot e^{-(p^2 n)^{-2}(x_i - L)^2/2} + \frac{1-p}{1-p} \cdot e^{-(1-p)^{-2}(x_i - L)^2/2} \right).$$

Similarly, the likelihood of the mean being $-L$ is

$$\mathcal{L}_- = (2\pi)^{-n/2} \cdot \prod_{i=1}^{n} \left(\frac{p}{p^2 n} \cdot e^{-(p^2 n)^{-2}(x_i + L)^2/2} + \frac{1-p}{1-p} \cdot e^{-(1-p)^{-2}(x_i + L)^2/2} \right).$$

We first simplify the ratio $\mathcal{L}_+/\mathcal{L}_-$ and partition into two sums, each of which we will tackle separately. It is easy to derive that

$$\ln \frac{\mathcal{L}_+}{\mathcal{L}_-} = \frac{2L}{(1-p)^2} \cdot \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \ln \frac{1 + \frac{1}{pn} e^{-\left(p^{-4} n^{-2} - (1-p)^{-2}\right)(x_i - L)^2/2}}{1 + \frac{1}{pn} e^{-(p^{-4} n^{-2} - (1-p)^{-2})(x_i + L)^2/2}}.$$

Next, we handle each of the two parts separately. The final aim is to show that $\ln \frac{\mathcal{L}_+}{\mathcal{L}_-}$ can be positive or negative with constant probability.

Define $W = \frac{2L}{(1-p)^2} \cdot \sum_{i=1}^{n} x_i$. The variables x_i are independent, and are Gaussians with mean $\mu \in \{L, -L\}$ and variance less than 1. Therefore, $\sum_i x_i$ is a Gaussian with mean $n\mu$ and variance n, and hence W is a Gaussian with mean at most $2L\mu n$ and variance at most $4L^2 n$. Hence, for $0 < \epsilon < 1$, with probability $1 - \Theta(\exp(-1/\epsilon^2))$, we have $|W| \le 2L^2 n + \frac{2L\sqrt{n}}{\epsilon} \le 3$, where the last inequality is obtained by plugging in the value of L.

Now let us bound the latter sum denoted X.

$$X = \sum_{i=1}^{n} \ln \frac{1 + \frac{1}{pn} e^{-\left(p^{-4} n^{-2} - (1-p)^{-2}\right)(x_i - L)^2/2}}{1 + \frac{1}{pn} e^{-(p^{-4} n^{-2} - (1-p)^{-2})(x_i + L)^2/2}}.$$

We will show that the random variable X will (roughly) behave like a Gaussian with a standard deviation equal to a large constant; therefore X will have a large absolute value with constant probability, and its sign will be chosen (roughly) uniformly at random. We will use Berry–Esseen Theorem [5, 11] to prove this Gaussian-like behavior. Unfortunately, though, we cannot apply the Theorem directly since the terms of X are overly complicated and it is quite hard to compute their second and third moments. We will therefore start by simplifying X: we first show that if we remove the terms whose x_i's have a variance close to 1 we only change X by a constant that is small in absolute value. Then we show that by losing another constant of small absolute value, X can be approximated by the sum of differences between two exponential functions of x_i. We will then apply Berry–Essen Theorem to show that this sum of differences of exponentials is close to a Gaussian with sufficiently large variance that will be needed to tame the various additive constant approximations that we make along the way.

Each term of the sum of X can be upper bounded by $\ln\left(1 + \frac{1}{pn}\right) \le \frac{1}{pn}$, and lower bounded by $\ln\left(1 + \frac{1}{pn}\right)^{-1} \ge -\frac{1}{pn}$. Let $S \subseteq [n]$ be the set of indices such that $\sigma_i = 1 - p$. Using the Chernoff bound, we have $|[n] \setminus S| \le 2pn$ with probability $1 - n^{-\Theta(1)}$. Let X' be the contribution to X of the terms of its sum whose indices that belong to S, and X'' the contribution of the terms with indices not in S. Then, $X = X' + X''$. Moreover, $-2 \le X'' \le 2$ with high probability. Next, we bound X'.

Let $N_i = \frac{1}{pn} e^{-\left(p^{-4} n^{-2} - (1-p)^{-2}\right)(x_i - L)^2/2}$, and $D_i = \frac{1}{pn} e^{-\left(p^{-4} n^{-2} - (1-p)^{-2}\right)(x_i + L)^2/2}$; observe that $0 \le N_i, D_i \le \frac{1}{np}$. Moreover, $X' = \sum_{i \in S} \left(\ln(1 + N_i) - \ln(1 + D_i)\right)$. First observe that, if $q < Q < 1$, then $q \ge \ln(1 + q) \ge q - qQ/2 = (1 - Q/2) \cdot q$. Observe that both N_i, $D_i \le \frac{1}{np} = Q$. Therefore,

$$\sum_{i \in S} \left(\left(1 - \frac{1}{2np}\right) N_i - D_i \right) \le X'$$

$$X' \le \sum_{i \in S} \left(N_i - \left(1 - \frac{1}{2np}\right) D_i \right).$$

Putting these together,

$$-\frac{\sum_{i \in S} N_i}{2np} \le X' - \sum_{i \in S} (N_i - D_i) \le \frac{\sum_{i \in S} D_i}{2np}. \tag{4.4}$$

Now, for $i \in S$, by explicitly integrating,

$$E[N_i | \mu = L] = E[D_i | \mu = -L] = \frac{p}{1-p}, \tag{4.5}$$

and,

$$E[N_i | \mu = -L] = E[D_i | \mu = L] = \frac{p}{1-p} \cdot e^{2(1-p)^{-4} L^2 (np^2 + p - 1)(np^2 - p + 1)} = (1 \pm O(L^2)) \cdot \frac{p}{1-p}. \tag{4.6}$$

Thus, $E[\sum_{i\in S} N_i] \leq 2p|S|$, and similarly for $\sum_{i\in S} D_i$. Furthermore, using the Chernoff bounds on $\sum_i N_i$ and $\sum_i D_i$, we can show that with probability $1-n^{-\Theta(1)}$, $\sum_{i\in S} N_i \leq 4p|S|$ and $\sum_{i\in S} D_i \leq 4p|S|$. Using these bounds in inequality (4.4) we get, with probability $1-n^{-\Theta(1)}$,

$$|X' - \sum_{i\in S}^{n} (N_i - D_i)\,| \leq \frac{1}{2np} \cdot 4p|S| \leq 2p. \tag{4.7}$$

Next, we will show that $\sum_{i\in S}(N_i - D_i)$ is anti-concentrated. In order to do so, we first bound the first and second moments of $N_i - D_i$. Using equation (4.5) and inequality (4.7), we have that $E[N_i - D_i|\mu = L] = O(pL^2)$. Also, since $E[N_i - D_i|\mu = -L] = -E[N_i - D_i|\mu = L]$, we have that $E[N_i - D_i|\mu] = \mu \cdot O(pL)$.

The second moment of $N_i - D_i$ can be explicitly computed by integrating as follows:

$$E[(N_i - D_i)^2|\mu = -L] = \frac{1 + e^{-4\frac{\left(1-2p+p^2-p^4n^2\right)L^2}{(2-4p+2p^2-p^4n^2)(1-p)^2}} - 2\,e^{-2\frac{\left(1-2p+p^2-p^4n^2\right)L^2}{p^4n^2(2-4p+2p^2-p^4n^2)}}}{n\sqrt{2-4p+2p^2-p^4n^2}}.$$

We plug in $L = O(n^{-1/2})$. Then, if $p = O\left(n^{-3/4}\right) = O\left(\sqrt{\frac{L}{n}}\right)$, we get that $E[(N_i - D_i)^2|\mu = -L] = O\left(n^{-1}\right)$. If, instead, $p = \omega\left(n^{-3/4}\right)$ we have $E[(N_i - D_i)^2|\mu = -L] = O\left(\frac{L^2}{n^3p^4}\right) = O\left(n^{-4}p^{-4}\right)$.

Analogously, $E[(N_i - D_i)^2|\mu = L] = O\left(n^{-1}\right)$, if $p = O\left(n^{-3/4}\right)$, and $E[(N_i - D_i)^2|\mu = L] = O\left(n^{-4}p^{-4}\right)$, if $p = \omega\left(n^{-3/4}\right)$.

Therefore, no matter of whether we condition on μ being L or $-L$, we get

$$E[(N_i - D_i)^2|\mu] = O\left(\min\left(n^{-1}, n^{-4}p^{-4}\right)\right).$$

Define the deviation $Z_i = (N_i - D_i) - E[N_i - D_i]$, implicitly under the conditioning $\mu = |L|$. In order to apply Berry–Essen Theorem on Z_i, we have to bound the first three moments of Z_i. Obviously, $E[Z_i] = 0$. Let the second moment be given by $s^2 = E[Z_i^2]$. Thus, $s^2 = E[(N_i - D_i)^2] - E^2[N_i - D_i] = E[(N_i - D_i)^2] - O(L^4p^2) = (1 \pm o(1))E[(N_i - D_i)^2]$, since $p < o\left(n^{1/3}\right)$.

Finally, define $r = E[|Z_i|^3]$. Hence,

$$\begin{aligned} r &\leq E[Z_i^2] \cdot \max|Z_i| \\ &\leq O\left(\min\left(n^{-1}, n^{-4}p^{-4}\right)\right) \cdot (\max N_i + \max D_i) \\ &= O\left(\min\left(n^{-2}p^{-1}, n^{-5}p^{-5}\right)\right). \end{aligned}$$

By the Berry–Essen Theorem, we have that the CDF $F(x)$ of $Z = \frac{1}{s\sqrt{n}}\sum_{i\in S} Z_i$ satisfies $|F(x) - \Phi(x)| < O\left(\frac{r}{s^3\sqrt{n}}\right)$, where $\Phi(x)$ is the CDF of a standard normal random variable. If $p = O(n^{-3/4})$, this additive error simplifies to $O\left(\frac{n^{-2}p^{-1}}{n^{-3/2}\sqrt{n}}\right) = O\left(\frac{1}{np}\right) = o(1)$, by using $p = \omega\left(\frac{1}{n}\right)$. If, instead, $p = \omega(n^{-3/4})$, the additive error simplifies to $O\left(\frac{n^{-5}p^{-5}}{n^{-6}p^{-6}\sqrt{n}}\right) = O\left(p\sqrt{n}\right) = o(1)$, by using $p = o(n^{-1/2})$.

Since $s\sqrt{n} = \Theta(1)$, $\Pr[\sum_i Z_i > 100] = \Pr[Z > \Theta(1)] \geq c$ for some constant c. The same holds for $\Pr[\sum_i Z_i < -100]$. Also, $|E[\sum_{i\in S} N_i - D_i]| \leq O(npL^2) \leq 1$. Thus, $\Pr[\sum_{i\in S} N_i - D_i > 99] \geq \Pr[\sum_i Z_i > 100] \geq c$. Using inequality (4.7), we get that both $\Pr[X' > 90] \geq c$ and $\Pr[X' < -90] \geq c$. Since $|X - X'| = |X''| \leq 2$, and and since $\log\frac{\mathcal{L}_+}{\mathcal{L}_-} = W + X$, with $|W|$ being at most 3 with probability $1 - e^{-\Theta(\epsilon^{-2})}$, we have that $\log\frac{\mathcal{L}_+}{\mathcal{L}_-}$ can be positive with probability $c - e^{-\Theta(\epsilon^{-2})}$ and negative with probability $c - e^{-\Theta(\epsilon^{-2})}$, regardless of whether μ is L or $-L$. Hence, regardless of the initial choice of μ, with probability $c - e^{-\Theta(\epsilon^{-2})}$, the $\mu = L$ hypothesis is more likely, and similarly for $\mu = -L$.

It follows that the expected loss is at least $\left(c - e^{-\Theta(\epsilon^{-2})}\right) \cdot L = \Omega(L)$. This gives us the statement of the Lemma. ■

The following corollary that bounds the gap of the unknown and known variance case can then be obtained by choosing $p = \Theta\left(\frac{\log n}{n}\right)$.

COROLLARY 4.1. *There exist instances where loss $E[|\mu - \hat{\mu}|]$ of any algorithm that does not know the variances is a factor $\Omega\left(\sqrt{\frac{n}{\log^3 n}}\right)$ worse than the loss of the (optimal) known variance estimate in Theorem 3.1.*

Figure 1 shows how the gap between the two settings varies as a function of p.

5 Multiple items

In this section we present algorithms that work in the multiple items case. The question is whether we can utilize the presence of multiple items to better estimate the variances of the users, and hence the means of the individual items. In this section we assume we are always in the unknown variance setting.

For each user $i \in [n]$, let $I(i)$ be the set of items rated by the user and for each item $j \in [m]$, let $U(j)$ be the set of users rating this item. Let G denote the user–item rating matrix, i.e., if user i has rated

item j, then $G_{ij} = x_{ij} \sim N(\mu_j, \sigma_i^2)$ and is empty otherwise. For two users i and ℓ, let $c_{i\ell} = |I(i) \cap I(\ell)|$ be the number of items rated by both i and ℓ, and let $c_{\min} = \min_{i \neq \ell} c_{i\ell}$. We consider two cases: G is complete and G is random.

For the complete graph case, we present an algorithm based on ideas underlying the k-Shortest Gap algorithm for $m = 1$ case; this algorithm will leverage the fact that there many ratings from a single user. The main intuition behind this algorithm is to find a distance C, and a set of raters who have all of their ratings at a distance at most C from all the true item means. For each item, one of the ratings from these raters will be returned as an estimate.

Algorithm(G complete, $m = O(\log n)$)
Input: m items with mean $\{\mu_j\}$, and n samples $x_{ij} \sim N(\mu_j, \sigma_i^2)$ for each item j.
Output: An estimate for each μ_j, $j \in [m]$.
Let $k = \log n$.
Initialize C_k with the size of maximum k-Shortest Gap over all items.
Let $T_k = \emptyset$.
while $T_k = \emptyset$ **do**
  $C_k = 2C_k$.
  Look for a set of k raters such that for each item the maximum distance between two ratings given by those raters is C_k, if such a set exists then add it to T_k.
Return the ratings of a rater in T_k

Note that the above algorithm w.h.p. has running time $\tilde{O}(n^2 \log \frac{D}{C_{\log n}})$, where D is the maximum distance between the ratings of two objects.

We state the following theorem about the performance of the algorithm above.

THEOREM 5.1. *The above algorithm has a loss of at most* $O(n^{1/m}\sigma_{\log n} \operatorname{polylog} n)$.

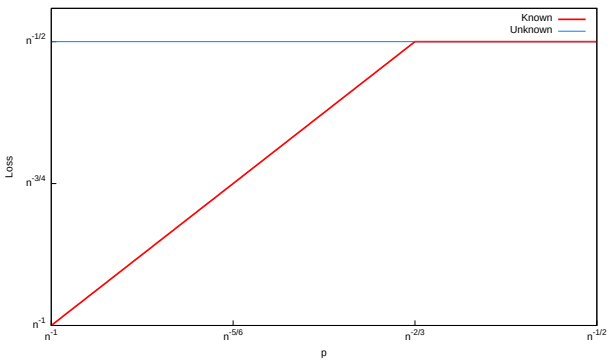

Figure 1: The optimal known-variance and unknown-variance losses of Example 1, as p varies. Both axes are in a log-scale.

Proof. Let $k = \log n$, and let C_k the smallest value for which $T_k \neq \emptyset$. We then know that for each item j the size of k-Shortest Gap is at most C_k. Now we proceed as in Lemma 4.1.

We call a rater with standard deviation more than $(n \log n)^{(1+1/(k-1))1/m} C_k$ as a *bad* rater, and other to be *good* raters. First, we show that if a good rater is in T_k, then the statement holds. Next, we prove that w.h.p. T_k has at least one good rater.

The probability that a rating of a good rater gives rating that is outside $[\mu - (n \log n)^{(1+1/(k-1))1/m} C_k \log^2 n, \mu + (n \log n)^{(1+1/(k-1))1/m} C_k \log^2 n]$ is upper bounded by $1 - \operatorname{erf}\left(\frac{\log^2 n}{\sqrt{2}}\right) < \sqrt{\frac{2}{\pi}} \frac{n^{-\log^3 n/2}}{\log n} \in \frac{1}{(mn)^{\Theta(\log^3 n)}}$. Hence, by the union bound, with probability at least $1 - n^{-\Theta(\log n)}$ all the ratings of good raters are inside $[\mu - (n \log n)^{(1+1/(k-1))1/m} C_k \log^2 n, \mu + (n \log n)^{(1+1/(k-1))1/m} C_k \log^2 n]$. Hence, if T_k contains one of the good raters then we are done. Now to conclude our argument, that is, to show that T_k must contain a good rater, we need to show that with probability at least $1 - n^{-\Theta(\log n)}$ there is no set of m intervals of size C_k that covers the ratings of k bad for all the items.

We proceed by first computing the probability that any set of m intervals, each of size $2C_k$, contain all the ratings of a bad rater. It is possible to see that the intervals that are more likely to contains those ratings are the intervals of size $2C_k$ centered around the means μ_j for each of the j items. For each item j, the probability that the rating by a single bad rater is in the interval $[\mu_j - C_k, \mu_j + C_k]$ is $\operatorname{erf}\left(\frac{1}{(n \log n)^{(1+1/(k-1))1/m}}\right) \in \Theta\left(\frac{1}{(n \log n)^{(1+1/(k-1))1/m}}\right)$. So the probability that all the ratings by this bad rater are contained in the set of size $2C_k$ intervals is $\Theta\left(\left(\frac{1}{(n \log n)^{(1+1/(k-1))1/m}}\right)^m\right) = \Theta\left(\left(\frac{1}{n \log n}\right)^{1+1/(k-1)}\right)$.

Thus the probability that the intervals of size $2C_k$ centered around the ratings of a bad rater contains the ratings of other $k-1$ bad raters is upper bounded by $\binom{n}{k-1}\Theta\left(\left(\frac{1}{n \log n}\right)^{1+1/(k-1)}\right)^{k-1} \in O(\frac{1}{n^{\log n}})$. Hence, by using the union bound on the number of bad raters, we get that with probability at least $1 - n^{-\log n}$ there is no set of m intervals of size $2C_k$ that covers the ratings of k bad raters for all the items. Hence, T_k must contain a good rater.

The proof is complete by arguing as in the single item case. ■

We next study two cases: G is complete with $m = \Omega(\log n)$, and G is a random graph. For the case the G is complete and $m = \Omega(\log n)$, we can almost match the optimal loss in the known variance case.

5.1 G complete, $m = \Omega(\log n)$. We now present an algorithm with a performance guarantee that almost matches the known-variance case. The intuition is that, since two users j and k overlap on $\Omega(\log n)$ items, it is possible to estimate the sum of their variance $\sigma_j^2 + \sigma_k^2$ reasonably well. Also, for most users j, by looking at the set of values $\sigma_j^2 + \sigma_k^2$ we can derive an estimate of σ_j^2 that is only a constant factor off.

For users $j, k \in [n]$, define

$$Z_{jk} = \frac{1}{c_{jk}} \sum_{i \in I(j) \cap I(k)} (x_{ij} - x_{ik})^2.$$

Since x_{ij} is a Gaussian, each $x_{ij} - x_{ik} \sim N(0, \sigma_j^2 + \sigma_k^2)$. The random variable Z_{jk} is thus distributed as a χ^2 distribution with expectation $\sigma_j^2 + \sigma_k^2$ and with degree of freedom c_{jk}. Let $F_{jk}(z)$ denote the CDF of the random variable Z_{jk}. Define the shorthand $v_{jk} = \sigma_j^2 + \sigma_k^2$. Hence, for $0 < z < 1$, the CDF satisfies $F_{jk}(v_{jk}z) \leq (ze^{1-z})^{c_{jk}/2}$ and for $z > 1$, the CDF satisfies $F_{jk}(v_{jk}z) \geq 1 - (ze^{1-z})^{c_{jk}/2}$.

By taking $z = 1/2$ in the first and $z = 2$ in the second, and plugging in $c_{jk} = c\log(n)$, for $c > 8$, we have that with probability $1 - n^{-\Theta(c)}$, for all pairs $(j,k) \in [n] \times [n]$, $Z_{jk} \in [v_{jk}/2, 2v_{jk}]$. Define $\hat{\sigma}_j^2 = \min_{k \neq j} Z_{jk}$ as the minimum over all pairs (j,k). Finally, for each item i define $\hat{\mu}_i = \frac{\sum_{j \in U(i)} x_{ij}/\hat{\sigma}_j^2}{\sum_{j \in U(i)} 1/\hat{\sigma}_j^2}$.

LEMMA 5.1. *Let user 1 be the user with minimum variance σ_1^2. For $j \neq 1$, with probability $1 - n^{-\Theta(c)}$, $\sigma_j^2/2 \leq \hat{\sigma}_j^2 \leq 4\sigma_j^2$. For $j = 1$, $\hat{\sigma}_1^2 \geq \sigma_2^2/2$.*

Proof. The proof for $j > 1$ simply follows from the fact that for each $j \neq 1$, $Z_{j1} \leq 2v_{j1} \leq 2(\sigma_j^2 + \sigma_1^2) \leq 4\sigma_j^2$. Hence $\hat{\sigma}_j^2 \leq 4\sigma_j^2$. Similarly, for each j, for each k, $Z_{jk} \geq v_{jk}/2 \geq \sigma_j^2/2$. Hence the first part of the Lemma follows. For $j = 1$, we can only guarantee that $\hat{\sigma}_1^2 \geq v_{12}/2$, which implies $\hat{\sigma}_1^2 \geq \sigma_2^2/2$. ■

The above lemma can directly be used to prove Theorem 5.2; the proof is analogous to that of Theorem 3.1 and hence is omitted.

THEOREM 5.2. *Consider any item i. The loss of item i satisfies $E\|\hat{\mu}_i - \mu_i\| \leq O\left(\sqrt{\frac{1}{\sum_{j=2}^n \frac{1}{\sigma_j^2}}}\right)$.*

5.2 G random, sparse. Next we consider the case the assignment G is random, $G = G(n, m, \alpha), m = \Omega(\log n)$ and $\alpha = \Omega\left(\sqrt{\frac{\log n}{m}}\right)$, i.e. G is created with a specific probability $\alpha = \Omega\left(\sqrt{\frac{\log n}{m}}\right)$ in the following manner: each entry of $G_{i,j}$ is empty with probability $1-\alpha$, and, with probability α, it contains a sample of $N(\mu_i, \sigma_j^2)$.

Our algorithm for this setting will just mimic the one in Section 5.1. Two generic users j and k will have an expected number of commonly rated items equal to $E[c_{jk}] = m \cdot \alpha^2 = \Omega(\log n)$. By the Chernoff bound, with probability $1 - n^{-2}$ for each $1 \leq j < k \leq n$ the number c_{jk} will be at least $\Omega(\log n)$. Therefore, we can use the algorithm in Section 5.1 to estimate the variances of the users to within a constant factor.

Now consider a generic item i. If this item is not rated by the user with smallest variance, then we have a good approximation of all the variances of its raters. Therefore, we can use the known-variance estimate to get an optimal loss of $O\left(\sqrt{\frac{1}{\sum_{j=1}^n \frac{1}{\sigma_{i_j}^2}}}\right)$ for a set of raters $i_1, \ldots, i_k$. If the item was rated by the user with smallest variance then, as in Theorem 5.2, we lose the contribution of this best user, so we match what the known variance algorithm would do if it did not have access to the best user.

The guarantee of the algorithm for the sparse random G is give in the following Theorem 5.3. The proof is similar to that of Theorem 3.1 and hence is omitted.

THEOREM 5.3. *The loss $E[|\mu_i - \hat{\mu}_i|]$ obtained on item i in $G(n, m, \alpha)$ (with $m = \Omega(\log n)$ and $\alpha = \Omega\left(\sqrt{\frac{\log n}{m}}\right)$), is equal to the (known-variance optimum) $O\left(\left(\sum_{t=1}^k \sigma_{j_t}^{-2}\right)^{-1/2}\right)$, if i was rated by the users $j_1, \ldots, j_k$, and if none of them was the user with smallest variance. If, instead, j_1 is the user with smallest variance, then the loss $E[|\mu_i - \hat{\mu}_i|]$ on item i is $O\left(\left(\sum_{t=2}^k \sigma_{j_t}^{-2}\right)^{-1/2}\right)$.*

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

[1] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proc. COLT*, pages 458–469, 2005.

[2] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proc. STOC*, pages 247–257, 2001.

[3] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Proc. FOCS*, pages 103–112, 2010.

[4] M. Belkin and K. Sinha. Toward learning Gaussian mixtures with arbitrary separation. In *Proc. COLT*, pages 407–419, 2010.

[5] A. C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49:122–136, 1941.

[6] B. Carpenter. A hierarchical Bayesian model of crowdsourced relevance coding. In *Proc. TREC*, 2011.

[7] H. Cremér. *Mathematical Methods of Statistics (PMS-9)*. Princeton University Press, 1999.

[8] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. Aggregating crowdsourced binary ratings. In *Proc. WWW*, pages 285–294, 2013.

[9] S. Dasgupta. Learning mixtures of Gaussians. In *Proc. FOCS*, pages 634–644, 1999.

[10] O. Dekel and O. Shamir. Vox populi: Collecting high-quality labels from a crowd. In *Proc. COLT*, 2009.

[11] C.-G. Esseen. A moment inequality with an application to the central limit theorem. *Skand. Aktuarietidskr.*, 39:160–170, 1956.

[12] A. Ghosh, S. Kale, and R. P. McAfee. Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content. In *Proc. EC*, pages 167–176, 2011.

[13] J. Hartung, G. Knapp, and B. K. Sinha. *Statistical Meta-Analysis with Applications*. Wiley, 2011.

[14] P. J. Huber. *Robust Statistics*. Springer, 2011.

[15] P. G. Ipeirotis and P. K. Paritosh. Managing crowdsourced human computation: a tutorial. In *Proc. WWW (Companion Volume)*, pages 287–288, 2011.

[16] A. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *Proc. STOC*, pages 553–562, 2010.

[17] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SICOMP*, 38(3):1141–1156, 2008.

[18] D. Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In *Proc. NIPS*, pages 1953–1961, 2011.

[19] Q. Liu, J. Peng, and A. Ihler. Variational inference for crowdsourcing. In *Proc. NIPS*, 2012.

[20] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics*. J. Wiley, 2006.

[21] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proc. FOCS*, pages 93–102, 2010.

[22] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *JMLR*, 11:1297–1322, 2010.

[23] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. KDD*, pages 614–622, 2008.

[24] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Proc. NIPS*, pages 1085–1092, 1995.

[25] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *JCSS*, 68(4):841–860, 2004.

[26] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *Proc. WWW*, pages 679–688, 2012.

[27] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Proc. NIPS*, pages 2424–2432, 2010.

[28] D. Zhou, J. Platt, S. Basu, and Y. Mao. Learning from the wisdom of crowds by minimax entropy. In *Proc. NIPS*, pages 2204–2212, 2012.