Computer Science 2: Advanced Algorithms Fall 2026

Lecture 10 — September 21st, 2026

Prof. Jarosław Błasiok *Scribes: Alessandro Butti, Simone Toso*

Robust Mean Estimation

Statement on AI use

We used LLMs to aid the transcription process from our notes, especially in the last section concerning the Golden-Thompson inequality, and to look for the referenced papers.

1 Overview

Estimating the mean of an unknown distribution from i.i.d. samples is a basic statistical problem. Under a bounded-covariance assumption, the empirical mean achieves small error with sufficiently many samples. However, it is not robust: even a single corrupted observation can move it arbitrarily far from the true mean.

This lecture studies *robust mean estimation*, where an adversary may replace an ε fraction of the sample [1, 2]. The key idea is that the uncorrupted points still have bounded empirical covariance. We relax the unknown good set to a large weighted subset and show that bounded covariance, together with closeness in total variation, forces its mean to be close to the good-sample mean. This leads to a polynomial-time alternating algorithm with error $O(\sqrt{\varepsilon})$, where ε is the fraction of samples that may be corrupted. We conclude with a brief introduction to matrix concentration.

2 Mean estimation, with and without outliers

Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$, with mean μ and covariance matrix

$$\Sigma := \mathop{\mathbb{E}}_{x\sim\mathcal{D}}(x-\mu)(x-\mu)^\top.$$

For a unit vector $v \in \mathbb{R}^d$, the scalar $\langle x, v\rangle$ is the projection of x onto that direction, and its variance is

$$\mathrm{Var}_{x\sim\mathcal{D}}\big(\langle x, v\rangle\big) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\big(\langle x, v\rangle - \langle \mu, v\rangle\big)^2 = v^\top \Sigma v.$$

Throughout, we assume $\Sigma \preceq I$, where I is the $d \times d$ identity matrix. Equivalently,

$$v^\top \Sigma v \le 1 \qquad \text{for every } v \in \mathbb{R}^d \text{ with } \|v\|_2 = 1.$$

Thus, the distribution has variance at most 1 in every unit direction.

Our goal is to estimate μ from samples. We begin with the uncorrupted case, where the empirical mean provides a useful baseline.

2.1 The empirical mean

Let $x_1, \dots, x_n \sim \mathcal{D}$ be i.i.d. samples and define the empirical (arithmetic) mean

$$\widehat{\mu} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Since each sample has mean μ,

$$\mathbb{E}[\widehat{\mu}] = \mu,$$

so the empirical mean is an unbiased estimator of the population mean.

To quantify its error, define the centered variables

$$z_i := x_i - \mu.$$

Then $\mathbb{E}[z_i] = 0$ and

$$\widehat{\mu} - \mu = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

Therefore,

$$\begin{aligned} \mathbb{E} \left\| \widehat{\mu} - \mu \right\|_2^2 &= \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} z_i \right\|_2^2 \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E} \left\| z_i \right\|_2^2 \\ &= \frac{1}{n} \mathbb{E} \left\| z \right\|_2^2 \\ &= \frac{\operatorname{Tr} \Sigma}{n} \leq \frac{d}{n}, \end{aligned}$$

where the mixed terms vanish because the z_i are independent and centered, while $\operatorname{Tr} \Sigma \leq d$ follows from $\Sigma \preceq I$. Thus, if

$$n \geq \frac{d}{\varepsilon^2},$$

then

$$\mathbb{E} \left\| \widehat{\mu} - \mu \right\|_2^2 \leq \varepsilon^2.$$

Hence the natural sample complexity for squared error ε^2 is of order d/ε^2.

This is an expectation guarantee, for subgaussian distributions the same error scale holds with high probability.

2.2 Why the empirical mean is not robust

Suppose that a single observation is moved by a vector of norm R. Since each observation has weight $1/n$, the empirical mean moves by R/n. Because R can be arbitrarily large, even one corrupted point can destroy the estimate. This motivates using covariance as a certificate of reliability: large displacements create large variance in some direction, whereas small displacements have a correspondingly limited effect on the mean.

2.3 Subgaussian samples and the contamination model

The bound

$$\mathbb{E}\,\|\widehat{\mu}-\mu\|_2^2 \le \frac{d}{n}$$

requires only independence and bounded covariance. For robust estimation, we additionally need a high-probability bound on the empirical covariance matrix of the clean sample.

We therefore assume that $\mathcal{D}$ is subgaussian. Namely, there exists $K = O(1)$ such that, for every $v \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \exp\big(\lambda\,\langle x-\mu, v\rangle\big) \le \exp\left(\frac{K^2\lambda^2\|v\|_2^2}{2}\right).$$

This moment-generating-function bound implies the tail estimate

$$\mathbb{P}(|\,\langle x-\mu, v\rangle\,| \ge t) \le 2\exp\left(-\frac{t^2}{2K^2\|v\|_2^2}\right).$$

This condition implies that large deviations are exponentially unlikely in every direction, ensuring that the tails die off fast enough, which is needed to bound the empirical second moments of the clean sample with high probability.

Contamination model Let $\widetilde{x}_1,\dots,\widetilde{x}_n \sim \mathcal{D}$ be i.i.d. clean samples. After observing the distribution, the sample, and the algorithm, an adversary may replace at most εn points arbitrarily. The algorithm observes only the resulting sample $x_1,\dots,x_n$.

Let

$$G := \{i \in [n] : x_i = \widetilde{x}_i\}$$

be the unknown set of uncorrupted indices. Then

$$|G| \ge (1-\varepsilon)n.$$

Our goal is a polynomial-time estimator satisfying

$$\|\widehat{\mu}-\mu\| = O(\sqrt{\varepsilon}), \qquad n = \widetilde{O}(d/\varepsilon^2).$$

Thus, although the number of required samples grows with the ambient dimension d, the error due to an ε fraction of corruptions does not.

3 The good set has small covariance

3.1 Concentration of the clean sample

We use the following standard concentration result for subgaussian samples without proof; see, for example, [3].

Lemma 1 (Concentration)**.** *Let $\widetilde{x}_1,\dots,\widetilde{x}_n$ be i.i.d. samples from a subgaussian distribution $\mathcal{D}$ with mean μ and covariance $\Sigma \preceq I$. For*

$$n = \widetilde{O}(d/\varepsilon^2),$$

with probability at least $3/4$*, there exists a universal constant* C *such that*

$$\frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i-\mu)(\widetilde{x}_i-\mu)^\top \preceq CI. \tag{1}$$

Equivalently, for every unit vector v,

$$\frac{1}{n}\sum_{i=1}^{n}\big(\langle \widetilde{x}_i-\mu, v\rangle\big)^2 \le C.$$

Thus the clean sample has constant-order empirical second moment in every direction. This is the only point in the argument where subgaussianity and the sample-size requirement are used.

3.2 What survives on G

Since $(\widetilde{x}_i-\mu)(\widetilde{x}_i-\mu)^\top \succeq 0$, removing the corrupted indices can only decrease the sum in PSD order. Hence

$$\frac{1}{n}\sum_{i\in G}(x_i-\mu)(x_i-\mu)^\top \preceq CI.$$

Since $|G|\ge(1-\varepsilon)n$,

$$\frac{1}{|G|}\sum_{i\in G}(x_i-\mu)(x_i-\mu)^\top \preceq \frac{C}{1-\varepsilon}I =: C'I.$$

Define

$$\mu_G := \frac{1}{|G|}\sum_{i\in G}x_i, \qquad \Sigma_G := \frac{1}{|G|}\sum_{i\in G}(x_i-\mu_G)(x_i-\mu_G)^\top.$$

The recentering identity gives

$$\frac{1}{|G|}\sum_{i\in G}(x_i-\mu)(x_i-\mu)^\top = \Sigma_G + (\mu_G-\mu)(\mu_G-\mu)^\top,$$

and therefore

$$\Sigma_G \preceq C'I.$$

Thus

$$\sigma_G^2 := \|\Sigma_G\|_{\mathrm{op}} = O(1). \tag{2}$$

Importantly, this argument remains valid even if the adversary chooses the corrupted indices after observing the clean sample, since it only removes positive semidefinite terms.

3.3 Looking for G: soft subsets

Since the good set G is unknown, we cannot select it directly. Instead, we introduce weights $\alpha_i\in[0,1]$ that indicate how much each observation is included in the candidate subset, with

$$\sum_{i=1}^{n}\alpha_i = (1-\varepsilon)n.$$

The corresponding weighted mean is

$$\mu_\alpha := \frac{1}{(1-\varepsilon)n} \sum_{i=1}^{n} \alpha_i x_i.$$

We look for weights satisfying

$$\frac{1}{(1-\varepsilon)n} \sum_{i=1}^{n} \alpha_i (x_i - \mu_\alpha)(x_i - \mu_\alpha)^\top \preceq BI, \tag{$*$}$$

for some constant $B = O(1)$.

We can show that it is possible to find such weights. Indeed, choose any

$$G' \subseteq G, \qquad |G'| = (1-\varepsilon)n,$$

and set

$$\alpha_i = \mathbf{1}_{\{i \in G'\}}.$$

Then the weights select only good points, and the same deletion, renormalization, and recentering argument used for G shows that their covariance is bounded by BI.

Allowing the weights to take values in $[0, 1]$ gives a *soft subset*, a continuous relaxation of the unknown good subset.

3.4 Three distributions on the observed sample

We now compare the soft subset with the unknown good set G. Let us use the index i to indicate the individual elements of our sample.

Consider

$$\mathrm{unif}([n]), \qquad \mathrm{unif}(G), \qquad p_\alpha(i) := \frac{\alpha_i}{(1-\varepsilon)n},$$

where the first two distributions are uniform on all sample indices and on the good indices, respectively. Since $\sum_i \alpha_i = (1-\varepsilon)n$, p_α is a probability distribution. If $i \sim p_\alpha$, then the random observed point x_i has mean

$$\mathop{\mathbb{E}}_{i \sim p_\alpha}[x_i] = \frac{1}{(1-\varepsilon)n} \sum_i \alpha_i x_i = \mu_\alpha,$$

and covariance equal to the matrix in constraint $(*)$:

$$\frac{1}{(1-\varepsilon)n} \sum_{i=1}^{n} \alpha_i (x_i - \mu_\alpha)(x_i - \mu_\alpha)^\top$$

We compare these distributions using their *total variation distance*. For distributions p, q on a finite set,

$$\mathrm{TV}(p, q) := \frac{1}{2} \sum_i |p(i) - q(i)| = 1 - \sum_i \min\{p(i), q(i)\}.$$

Thus $1 - \mathrm{TV}(p, q)$ is the total probability mass shared by the two distributions.

Since $|G| \geq (1-\varepsilon)n$, the overlap between $\mathrm{unif}([n])$ and $\mathrm{unif}(G)$ is at least $1-\varepsilon$. Hence

$$\mathrm{TV}\big(\mathrm{unif}([n]), \mathrm{unif}(G)\big) = 1 - \frac{|G|}{n} \leq \varepsilon.$$

Similarly, since $\alpha_i \leq 1$,

$$p_\alpha(i) \leq \frac{1}{(1-\varepsilon)n},$$

or equivalently

$$\mathrm{unif}([n])(i) \geq (1-\varepsilon)p_\alpha(i).$$

Therefore the overlap between $\mathrm{unif}([n])$ and p_α is at least $1-\varepsilon$, and

$$\mathrm{TV}\big(\mathrm{unif}([n]), p_\alpha\big) \leq \varepsilon.$$

By the triangle inequality,

$$\mathrm{TV}\big(\mathrm{unif}(G), p_\alpha\big) \leq 2\varepsilon. \tag{3}$$

Thus, whatever admissible weights α we choose, the induced soft subset cannot differ from the good sample by more than 2ε in total variation.

4 Small variance pins down the mean

We now show that, if two distributions have small TV and bounded variance, we can define an upper bound to the distance between the means.

Lemma 2 (Total variation and bounded variance)**.** *Let P_1, P_2 be distributions on $\mathbb{R}^d$, with means μ_{P_1}, μ_{P_2} and covariance matrices satisfying*

$$\|\Sigma_{P_1}\|_{\mathrm{op}} \leq \sigma_1^2, \qquad \|\Sigma_{P_2}\|_{\mathrm{op}} \leq \sigma_2^2.$$

If

$$\mathrm{TV}(P_1, P_2) \leq \epsilon \leq \frac{1}{10},$$

then

$$\|\mu_{P_1} - \mu_{P_2}\| \leq 2\sqrt{\epsilon}\,(\sigma_1 + \sigma_2).$$

Why, in one picture. Let us consider a distribution with variance bounded by σ^2. If you want to move the mean by δ, you need ϵ of the mass at a distance $t = \delta/\epsilon$. costs variance

$$\epsilon t^2 = \frac{\delta^2}{\epsilon} \leq \sigma^2 \tag{4}$$

In other words, you cannot move the mean too far away without breaking the bound on the variance.

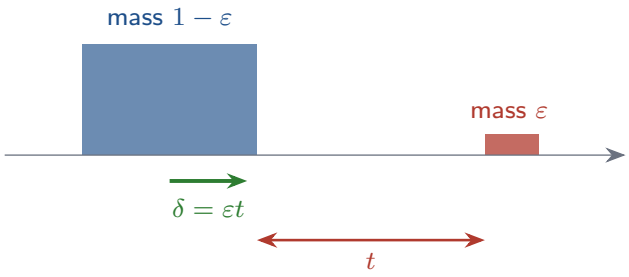


Figure 1: Moving a fraction ϵ of the mass by distance t changes the mean by order ϵt, while contributing a second moment of order ϵt^2. This motivates the $\sqrt{\epsilon}$ scale in Lemma 2.

Proof. First of all, we show that it is sufficient to prove the Lemma on $\mathbb{R}$.

Let us fix a unit vector $v \in \mathbb{R}^d$ and let us project the two distributions along v, getting two distributions on $\mathbb{R}$ that we call Q_1 and Q_2. We will still have $\mathrm{Var}(Q_i) \le \sigma_i^2$ for $i = 1, 2$. We also use the fact that the total variation TV cannot increase under a map, so $TV(Q_1, Q_2) \le \epsilon$ still.

Now we solve it in $\mathbb{R}$. Let us choose a maximal coupling (Y_1, Y_2) of Q_1 and Q_2, so that

$$\mathbb{P}[Y_1 \neq Y_2] = \mathrm{TV}(Q_1, Q_2) \le \epsilon.$$

By using the relation $|\mathbb{E}[X]| \le \mathbb{E}[|X|]$ and using the Cauchy-Schwartz inequality, we get

$$\begin{aligned} |\mu_{P_1} - \mu_{P_2}| &= |\mathbb{E}[Y_1 - Y_2]| \\ &\le \mathbb{E}[|Y_1 - Y_2|] \\ &= \mathbb{E}[|Y_1 - Y_2| \mathbf{1}_{Y_1 \neq Y_2}] \\ &\overset{C.S.}{\le} ||Y_1 - Y_2||_2 \sqrt{P(Y_1 \neq Y_2)} \\ &\le |||Y_1 - Y_2||_2 \sqrt{\epsilon} \end{aligned}$$

Let us define $\Delta := |\mu_{P_1} - \mu_{P_2}|$. By the triangle inequality in L^2, we have

$$\begin{aligned} \|Y_1 - Y_2\|_{L^2} &\le \|Y_1 - \mu_{P_1}\|_{L^2} + \Delta + \|Y_2 - \mu_{P_2}\|_{L^2} \\ &\le \sigma_1 + \Delta + \sigma_2. \end{aligned}$$

Putting the two inequalities together

$$\Delta \le \sqrt{\epsilon}\,(\sigma_1 + \sigma_2 + \Delta),$$

and therefore

$$(1 - \sqrt{\epsilon})\Delta \le \sqrt{\epsilon}\,(\sigma_1 + \sigma_2).$$

Since $\epsilon \le 1/10$,

$$\Delta \le \frac{\sqrt{\epsilon}}{1 - \sqrt{\epsilon}}(\sigma_1 + \sigma_2) \le 2\sqrt{\epsilon}\,(\sigma_1 + \sigma_2).$$

Thus, for every unit vector v,

$$\left|\langle v, \mu_{P_1} - \mu_{P_2}\rangle\right| \le 2\sqrt{\epsilon}\,(\sigma_1 + \sigma_2).$$

Taking the supremum over unit vectors gives

$$\|\mu_{P_1} - \mu_{P_2}\| = \sup_{\|v\|_2=1} \left|\langle v, \mu_{P_1} - \mu_{P_2}\rangle\right| \le 2\sqrt{\epsilon}(\sigma_1 + \sigma_2),$$

which proves the lemma. □

4.1 Consequence: any valid soft subset is already good

We can now apply Lemma 2 to a soft subset such as the ones described in section **??**.

Suppose α satisfies the weight constraints and constraint $(*)$. By (3),

$$\mathrm{TV}\big(p_\alpha, \mathrm{unif}(G)\big) \le 2\varepsilon.$$

Moreover,

$$\|\Sigma_{p_\alpha}\|_{\mathrm{op}} \le B, \qquad \|\Sigma_G\|_{\mathrm{op}} = \sigma_G^2 = O(1).$$

Since $2\varepsilon \le 1/10$, Lemma 2 yields

$$\|\mu_\alpha - \mu_G\| \le 2\sqrt{2\varepsilon}\,(\sigma_G + \sqrt{B}) = O(\sqrt{\varepsilon}). \tag{5}$$

This shows that we can get away with not knowing G itself. The mean of any sufficiently large weighted subset with bounded covariance already gets close within $O(\sqrt{\varepsilon})$ to μ_G.

5 An algorithm: alternating minimization

We now describe an algorithm which, when reiterated, leads to a stable estimate of the mean.

Let us call μ_k the current estimate of the mean. We keep the mean fixed, find weights whose second moment around μ_k is as small as possible, and then use their weighted mean as the next center:

$$\mu_k \longrightarrow \text{optimize the weights} \longrightarrow \mu_{k+1} \longrightarrow \text{repeat.}$$

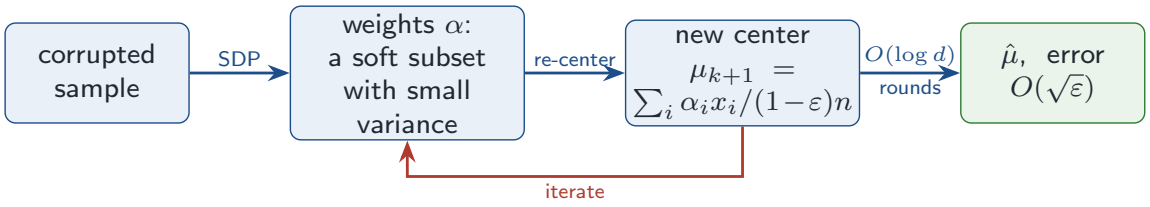


Figure 2: Each iteration selects a low-variance weighted subset around the current center and then moves the center to its weighted mean.

We will now show the rationale behind this procedure.

5.1 One iteration

For a fixed center μ_k, solve

$$\begin{aligned} \text{minimize}_{\alpha,\gamma} \quad & \gamma \\ \text{subject to} \quad & \frac{1}{(1-\varepsilon)n}\sum_{i=1}^{n}\alpha_i(x_i-\mu_k)(x_i-\mu_k)^\top \preceq \gamma I, \\ & 0 \le \alpha_i \le 1 \quad (i \in [n]), \\ & \sum_{i=1}^{n}\alpha_i = (1-\varepsilon)n. \end{aligned} \tag{6}$$

Let $\alpha^{(k+1)}$ be an optimal solution and define

$$\mu_{k+1} := \frac{1}{(1-\varepsilon)n}\sum_{i=1}^{n}\alpha_i^{(k+1)}x_i. \tag{7}$$

Once μ_k is fixed, the matrices $(x_i-\mu_k)(x_i-\mu_k)^\top$ are fixed and positive semidefinite. The matrix constraint in 6 is therefore linear in (α,γ), so this optimization problem is a semidefinite program.

5.2 Why the iteration makes progress

Let $\gamma^\star$ be the optimum of (6). The feasible weights α^G defined above, which were uniform over the good set and null elsewhere, provide a comparison point. For these weights,

$$\begin{aligned} \frac{1}{(1-\varepsilon)n}\sum_{i=1}^{n}\alpha_i^G(x_i-\mu_k)(x_i-\mu_k)^\top &= \frac{1}{|G|}\sum_{i\in G}(x_i-\mu_k)(x_i-\mu_k)^\top \\ &= \Sigma_G + (\mu_G-\mu_k)(\mu_G-\mu_k)^\top. \end{aligned}$$

Consequently,

$$\gamma^\star \le \sigma_G^2 + \|\mu_k-\mu_G\|^2. \tag{8}$$

Thus the objective is small whenever the current center is close to μ_G.

Let p_{k+1} be the distribution induced by the optimal weights $\alpha^{(k+1)}$. Its mean is μ_{k+1}. Its covariance about its own mean is no larger than its second moment about μ_k, so

$$\|\Sigma_{p_{k+1}}\|_{\mathrm{op}} \le \gamma^\star.$$

Moreover, the same total-variation argument as before gives

$$\mathrm{TV}\big(p_{k+1}, \mathrm{unif}(G)\big) \le 2\varepsilon.$$

Therefore, Lemma 2 and equation (8) imply

$$\|\mu_{k+1}-\mu_G\| \lesssim \sqrt{\varepsilon}\,\big(\sigma_G + \|\mu_k-\mu_G\|\big). \tag{9}$$

5.3 Contraction

For sufficiently small ε, (9) implies that there is a universal constant C_0 such that

$$\|\mu_k - \mu_G\| \geq C_0\sqrt{\varepsilon}\,\sigma_G \quad \Longrightarrow \quad \|\mu_{k+1} - \mu_G\| \leq \frac{1}{2}\|\mu_k - \mu_G\|. \tag{10}$$

Hence the error decreases geometrically until it reaches $O(\sqrt{\varepsilon})\sigma_G = O(\sqrt{\varepsilon})$.

5.4 Where to start

To bound the number of iterations, we need an initial estimate that is not too far from μ_G. A convenient choice is the coordinate-wise median μ_0 of the observed sample. For each coordinate, the empirical variance of the good points is at most σ_G^2. Hence at least three quarters of the good points lie within $2\sigma_G$ of the corresponding coordinate of μ_G. Since $|G| \geq (1-\varepsilon)n$ and $\varepsilon \leq 1/20$, more than half of all observed points lie in this interval. Therefore every coordinate-wise median satisfies

$$|(\mu_0)_j - (\mu_G)_j| \leq 2\sigma_G,$$

and consequently

$$\|\mu_0 - \mu_G\| \leq 2\sigma_G\sqrt{d} = O(\sqrt{d}). \tag{11}$$

Combined with the geometric contraction above, this shows that $O(\log(d/\varepsilon))$ iterations suffice to reach the target $O(\sqrt{\varepsilon})$ scale.

5.5 The algorithm

The resulting procedure is:

1. initialize μ_0 using the coordinate-wise median;
2. for $O(\log(d/\varepsilon))$ rounds, solve the SDP problem (6) at the current center and update it according to (7);
3. output the final center $\widehat{\mu}$.

The running time is polynomial, since each round solves a semidefinite program and the number of rounds is logarithmic in the relevant parameters.

On the event described in Lemma 1, the contraction argument show in equation 10 gives

$$\|\widehat{\mu} - \mu_G\| = O(\sqrt{\varepsilon}),$$

while (5) gives

$$\|\mu_G - \mu\| = O(\sqrt{\varepsilon}).$$

Therefore

$$\|\widehat{\mu} - \mu\| \leq \|\widehat{\mu} - \mu_G\| + \|\mu_G - \mu\| = O(\sqrt{\varepsilon}).$$

Altogether, for

$$n = \widetilde{O}(d/\varepsilon^2),$$

the algorithm returns an estimator satisfying

$$\|\widehat{\mu} - \mu\| = O(\sqrt{\varepsilon})$$

with probability at least 3/4.

6 Remarks and outlook

With only the bounded-covariance assumption $\Sigma \preceq I$, the $\Theta(\sqrt{\varepsilon})$ dependence is optimal, even with infinitely many samples [**?**]. For subgaussian distributions, the optimal dependence on the contamination rate is

$$\Theta\big(\varepsilon\sqrt{\log(1/\varepsilon)}\big),$$

which can be obtained by more refined versions of the same scheme.

Filtering methods replace the solution of the SDP problem with repeated computation of the top eigenvector [2]. Subsequent work obtains nearly linear running time for robust mean estimation [**?**].

More broadly, the same "small variance then recenter" principle extends to robust covariance estimation and other high-dimensional robust estimation problems.

7 Matrix concentration

In the previous sections we showed Lemma 1, which states that, with high probability and under certain conditions, we can pose an upper bound on the empirical second order moments and to the gap between empirical and real mean. This result is a matrix-valued analogue of familiar scalar concentration inequalities.

We now briefly discuss why passing from scalar random variables to random matrices introduces a new difficulty and why logarithmic dependence on the dimension naturally appears. For systematic treatments, see [4, 3].

The goal here is only to develop the main ideas. We do not prove a complete matrix concentration theorem.

7.1 From scalars to vectors and matrices

Consider first independent centered scalar random variables $z_1, \dots, z_n$ satisfying $|z_i| \le 1$. Then

$$\mathbb{E}\left(\sum_{i=1}^{n} z_i\right)^2 = \sum_{i=1}^{n} \mathbb{E}\, z_i^2 \le n,$$

where the $z_i z_j$ terms vanish because the z_i variables are independent and with zero mean. Hence

$$\mathbb{E}\left|\sum_{i=1}^{n} z_i\right| \le \sqrt{n}.$$

A similar calculation works for independent centered random vectors $z_i \in \mathbb{R}^d$ satisfying $\|z_i\|_2 \le 1$:

$$\mathbb{E}\left\|\sum_{i=1}^n z_i\right\|_2^2 = \sum_{i=1}^n \mathbb{E}\,\|z_i\|_2^2 \le n,$$

and therefore

$$\mathbb{E}\left\|\sum_{i=1}^n z_i\right\|_2 \le \sqrt{n}.$$

This bound has no explicit dependence on the ambient dimension d.

However, in the case of matrices the situation is different. Let $Z_1, \dots, Z_n \in \mathbb{R}^{d\times d}$ be independent centered symmetric random matrices with $\|Z_i\|_{\mathrm{op}} \le 1$, and define the matrix variance parameter

$$\nu := \left\|\sum_{i=1}^n \mathbb{E}\, Z_i^2\right\|_{\mathrm{op}}.$$

A known inequality gives, up to universal constants,

$$\mathbb{E}\left\|\sum_{i=1}^n Z_i\right\|_{\mathrm{op}} \lesssim \sqrt{\nu \log d} + \log d.$$

Since $\nu \le n$ under the norm bound above, the leading term is $\sqrt{n \log d}$ when n is at least of logarithmic order in d. The new feature is the logarithmic dependence on the matrix dimension [4].

7.2 The matrix moment-generating function

For scalar random variables, exponential moments are a standard tool for controlling large deviations. If $z_1, \dots, z_n$ are independent, then

$$\mathbb{E}\exp\left(\lambda \sum_{i=1}^n z_i\right) = \prod_{i=1}^n \mathbb{E}\exp(\lambda z_i).$$

This factorization is one of the key ingredients in the usual moment-generating-function argument.

To imitate this approach for matrices, we first define the matrix exponential. If Z is symmetric, the spectral theorem gives

$$Z = U\Lambda U^\top,$$

where U is orthogonal and Λ is diagonal. We define

$$\exp(Z) := U \exp(\Lambda) U^\top,$$

where $\exp(\Lambda)$ is obtained by exponentiating each diagonal entry. Equivalently, we can write a series expansion of the exponential matrix, similarly to what we do for scalar exponentials:

$$\exp(Z) = \sum_{k=0}^\infty \frac{Z^k}{k!}.$$

The main obstacle is that, for matrices, the commutative property does not apply in general. For scalars,

$$e^{a+b} = e^a e^b,$$

while for matrices in general

$$\exp(A+B) \neq \exp(A)\exp(B).$$

Thus the scalar factorization of exponential moments cannot be transferred directly to sums of random matrices.

7.3 Golden–Thompson and the role of the trace

A first way to partially overcome the failure of $\exp(A+B) = \exp(A)\exp(B)$ is to work with the trace of the matrix exponential. For symmetric matrices A and B, the Golden–Thompson inequality states that

$$\operatorname{Tr}\exp(A+B) \leq \operatorname{Tr}\big(\exp(A)\exp(B)\big).$$

The trace exponential controls the largest eigenvalue. If H is symmetric, write its eigenvalues as

$$\lambda_1(H), \dots, \lambda_d(H).$$

Then

$$\operatorname{Tr}\exp(H) = \sum_{j=1}^{d} e^{\lambda_j(H)} \geq e^{\lambda_{\max}(H)}.$$

Thus a bound on $\operatorname{Tr}\exp(H)$ controls $\lambda_{\max}(H)$; applying the same argument to $-H$ also controls the most negative eigenvalue and hence $\|H\|_{\mathrm{op}}$.

For example, if A and B are independent random symmetric matrices, Golden–Thompson gives

$$\begin{aligned}
\mathbb{E}\operatorname{Tr}\exp(A+B) &\leq \mathbb{E}\operatorname{Tr}\big(\exp(A)\exp(B)\big) \\
&= \operatorname{Tr}\mathbb{E}\big[\exp(A)\exp(B)\big] \\
&= \operatorname{Tr}\left(\mathbb{E}\exp(A)\ \mathbb{E}\exp(B)\right),
\end{aligned}$$

where the last equality uses independence.

The trace also explains the appearance of the dimension. Since

$$\operatorname{Tr}\exp(H) = \sum_{j=1}^{d} e^{\lambda_j(H)},$$

one effectively pays for up to d eigenvalue directions. In an exponential tail argument this produces a multiplicative factor of order d; after taking logarithms, it becomes a term of order $\log d$. This is the basic mechanism behind the $\sqrt{\log d}$ dependence in matrix concentration. A complete proof for sums of many noncommuting random matrices beyond this introductory discussion can be found in [4].

References

[1] Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016.

[2] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 999–1008, 2017.

[3] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[4] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.