[Lecture 8]

0.1 Efficient Algorithms

We now turn our attention to efficient algorithms. Recall that previously we considered minimum distance functionals projecting onto sets $\mathcal{G}$ and $\mathcal{M}$ under distances TV and $\widetilde{\mathsf{TV}}$. Here we will show how to approximately project onto the set $\mathcal{G}_{\mathsf{cov}}(\sigma)$, the family of bounded covariance distributions, under TV distance. The basic idea is to write down a (non-convex) optimization problem that tries to find the projection, and then show that the cost landscape of this optimization is nice enough that all local minima are within a constant factor of the global minimum.

To study efficient computation we need a way of representing the distributions $\tilde{p}$ and p^*. To do this we will suppose that $\tilde{p}$ is the empirical distribution over n points $x_1, \dots, x_n$, while p^* is the empirical distribution over some subset S of these points with $|S| \geq (1-\epsilon)n$. Thus in particular p^* is an ϵ-deletion of $\tilde{p}$.

Before we assumed that $\mathsf{TV}(p^*, \tilde{p}) \leq \epsilon$, but taking $p' = \frac{\min(p^*, \tilde{p})}{1-\mathsf{TV}(p^*, \tilde{p})}$, we have $p' \leq \frac{\tilde{p}}{1-\epsilon}$ and $\|\mathsf{Cov}_{p'}[X]\| \leq \frac{\sigma^2}{1-\epsilon} \leq 2\sigma^2$ whenever $\|\mathsf{Cov}_{p^*}[X]\| \leq \sigma^2$. Therefore, taking $p^* \leq \frac{\tilde{p}}{1-\epsilon}$ is equivalent to the TV corruption model from before for our present purposes.

We will construct an efficient algorithm that, given $\tilde{p}$, outputs a distribution q such that $\mathsf{TV}(q, p^*) \leq \mathcal{O}(\epsilon)$ and $\|\mathsf{Cov}_q[X]\|_2 \leq \mathcal{O}(\sigma^2)$. This is similar to the minimum distance functional, in that it finds a distribution close to p^* with bounded covariance; the main difference is that q need not be the projection of $\tilde{p}$ onto $\mathcal{G}_{\mathsf{cov}}$, and also the covariance of q is bounded by $\mathcal{O}(\sigma^2)$ instead of σ^2. However, the modulus of continuity bound from before says that *any* distribution q that is near p^* and has bounded covariance will approximate the mean of p^*. Specifically, we have

$$\|\mu(q) - \mu(p^*)\|_2^2 \leq \mathcal{O}(\max(\|\mathsf{Cov}_q[X]\|, \|\mathsf{Cov}_{p^*}[X]\|) \cdot \mathsf{TV}(p^*, q)) = \mathcal{O}(\sigma^2\epsilon). \tag{1}$$

Actually, we can obtain the tight constants in the $\mathcal{O}(\cdot)$:

Lemma 0.1. *If $\mathsf{TV}(p, q) \leq \epsilon$, then $\|\mu(p) - \mu(q)\|_2 \leq \sqrt{\frac{\|\Sigma_q\|\epsilon}{1-\epsilon}} + \sqrt{\frac{\|\Sigma_p\|\epsilon}{1-\epsilon}}$.*

We will prove Lemma 0.1 at the end of the section.

The main result of this section is the following:

Proposition 0.2. *Suppose $\tilde{p}$ and p^* are empirical distributions as above with $p^* \leq \tilde{p}/(1-\epsilon)$, and further suppose that $\|\mathsf{Cov}_{p^*}[X]\| \leq \sigma^2$ and $\epsilon < 1/3$. Then given $\tilde{p}$ (but not p^*), there is an algorithm with runtime $\mathrm{poly}(n, d)$ that outputs a q with $\mathsf{TV}(p^*, q) \leq \epsilon$ and $\|\mathsf{Cov}_q[X]\|\big(\frac{1-\epsilon}{1-3\epsilon}\big)^2\sigma^2$. In addition, $\|\mu(q) - \mu(p^*)\|_2 \leq 2\frac{\sqrt{\epsilon(1-2\epsilon)}}{1-3\epsilon}\sigma$.*

Note that the conclusion $\|\mu(p^*) - \mu(q)\|_2 \leq \mathcal{O}(\sigma\sqrt{\epsilon})$ follows from the modulus bound on $\mathcal{G}_{\mathsf{cov}}(\sigma)$ together with the property $\mathsf{TV}(p^*, q) \leq \epsilon$.

The algorithm, `MinCovL2`, underlying Proposition 0.2 is given below; it maintains a weighted distribution q, which places weight q_i on point x_i. It then computes the weighted mean and covariance, picking the weights that minimize the norm of the covariance.

The intuition behind Algorithm 1 is as follows: the constraint $q_i \leq \frac{1}{(1-\epsilon)n}$ ensures that q is an ϵ-deletion of the uniform distribution over $X_1, \dots, X_n$. Then, subject to that constraint, Algorithm 1 seeks to minimize the weighted covariance matrix: note the objective is exactly $\|\Sigma_q\|^2$.

Algorithm 1 is non-trivial to analyze, because although the constraint set is convex, the objective is non-convex: both q_i and μ_q are linear in q, and so the overall objective (even for a fixed v) is thus a non-convex cubic in q. On the other hand, for any "reasonable" choice of q, μ_q should be close to μ_{p^*}. If we apply this approximation–substituting μ_{p^*} for μ_q–then the objective becomes convex again. So the main idea behind the proof is to show that this substitution can be (approximately) done.

Before getting into that, we need to understand what stationary points of (2) look like. In general, a stationary point is one where the gradient is either zero, or where the point is at the boundary of the constraint set and the gradient points outward into the infeasible region for the constraints. However, the supremum of v can lead to a non-differentiable function (e.g. $\max(x_1, x_2)$ is non-differentiable when $x_1 = x_2$).

Algorithm 1 `MinCovL2`

```
Input: $x_1, \dots, x_n \in \mathbb{R}^d$.
Find any stationary point $q$ of the optimization problem:
    $\min_q \sup_{\|v\|_2 \le 1} \sum_{i=1}^n q_i \langle v, X_i - \mu_q \rangle^2,$    (2)
    $s.t. \mu_q = \sum_i q_i X_i,$
    $q \ge 0, \sum_i q_i = 1, q_i \le \frac{1}{(1-\epsilon)n}$
Output $\hat{\mu}_q$, the empirical mean for the stationary point $q$.
```

In this case, we can use something called a "Clarke subdifferential", to show that the preceding conditions hold for some v that maximizes the supremum:

Lemma 0.3. *Suppose that q is a stationary point of (2). Then, for any feasible p, there exists a $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ such that*

$$\mathbb{E}_q[\langle v, X - \mu_q \rangle^2] \le \mathbb{E}_p[\langle v, X - \mu_q \rangle^2]. \tag{3}$$

Moreover, v is a maximizer of the left-hand-side, i.e. $v^\top \Sigma_q v = \|\Sigma_q\|$.

Proof. Let $F_v(q) = \mathbb{E}_q[\langle v, X - \mu_q \rangle^2]$. First, compute $\nabla F_v(q)$. We have

$$\begin{aligned} \frac{\partial}{\partial q_i} F(q) &= \frac{\partial}{\partial q_i} \sum_{j=1}^n q_j \langle v, X_j - \mu_q \rangle^2 & (4) \\ &= \langle v, X_i - \mu_q \rangle^2 + 2 \sum_{j=1}^n q_j \langle v, X_j - \mu_q \rangle \frac{\partial \mu_q}{\partial q_i} & (5) \\ &= \langle v, X_i - \mu_q \rangle^2, & (6) \end{aligned}$$

where the last equality is because $\sum_j q_j (X_j - \mu_q) = 0$. Consequently, $\nabla F_v(q)_i = \langle v, X_i - \mu_q \rangle^2$.

Now, let $F(q) = \max_{\|v\|_2 = 1} F_v(q) = \|\Sigma_q\|$. If the maximizing v is unique and equal to v^*, then $\nabla F(q) = \nabla F_{v^*}(q)$, and q is a stationary point if and only if $\sum_i (q_i - p_i) \nabla F_{v^*}(q)_i \le 0$ for all feasible p, or equivalently $\mathbb{E}_q[\angle v^*, X_i - \mu_q \rangle^2] - \mathbb{E}_p[\langle v^*, X_i - \mu_q \rangle^2] \le 0$, which is exactly the condition (3).

Suppose (the harder case) that the maximizing v is not unique. Then F is not differentiable at q, but the Clark subdifferential is the convex hull of $\nabla F_v(q)$ for all maximizing v's. Stationarity implies that $\sum_i (q_i - p_i) g_i \le 0$ for some g in this convex hull, and thus by convexity that $\sum_i (q_i - p_i) \nabla F_{v^*}(q)_i \le 0$ for some maximizing v^*. This gives us the same desired condition as before and thus completes the proof. □

Given Lemma 0.3, we are in a better position to analyze Algorithm 1. In particular, for any p (we will eventually take the global minimizer $\overline{p}$ of (2)), Lemma 0.3 yields

$$\begin{aligned} \|\mathsf{Cov}_q\| &= \mathbb{E}_q[\langle v, X - \mu_q \rangle^2] & (7) \\ &\le \mathbb{E}_p[\langle v, X - \mu_q \rangle^2] & (8) \\ &= \mathbb{E}_p[\langle v, X - \mu_p \rangle^2] + \langle v, \mu_p - \mu_q \rangle^2 & (9) \\ &\le \|\mathsf{Cov}_p\| + \|\mu_p - \mu_q\|_2^2. & (10) \end{aligned}$$

The $\|\mu_p - \mu_q\|_2^2$ quantifies the "error due to non-convexity"–recall that if we replace μ_q with a fixed μ_p in (2), the problem becomes convex, and hence any stationary point would be a global minimizer. The distance $\|\mu_p - \mu_q\|_2^2$ is how much we pay for this discrepancy.

Fortunately, $\mu_p - \mu_q$ is small, precisely due to the modulus of continuity! We can show that any feasible p, q for (2) satisfies $\mathsf{TV}(p, q) \leq \frac{\epsilon}{1-\epsilon}$ (see Lemma 0.4), hence Lemma 0.1 gives $\|\mu_p - \mu_q\|_2 \leq \sqrt{\frac{\|\Sigma_q\|\epsilon}{1-2\epsilon}} + \sqrt{\frac{\|\Sigma_p\|\epsilon}{1-2\epsilon}}$. Plugging back in to (10), we obtain

$$\|\mathsf{Cov}_q\| \leq \|\mathsf{Cov}_p\| + \frac{\epsilon}{1-2\epsilon}\Big(\sqrt{\|\Sigma_p\|} + \sqrt{\|\Sigma_q\|}\Big)^2. \tag{11}$$

For fixed $\|\mathsf{Cov}_p\|$, we can view this as a quadratic inequality in $\sqrt{\|\Sigma_p\|}$. Solving the quadratic then yields

$$\|\mathsf{Cov}_q\| \leq \Big(\frac{1-\epsilon}{1-3\epsilon}\Big)^2 \|\mathsf{Cov}_p\|. \tag{12}$$

In particular, taking p to be the global minimum $\overline{p}$ of (2), we have $\|\mathsf{Cov}_{\overline{p}}\| \leq \|\mathsf{Cov}_{p^*}\| \leq \sigma^2$, so $\|\mathsf{Cov}_q\| \leq \Big(\frac{1-\epsilon}{1-3\epsilon}\Big)^2 \sigma^2$. Plugging back into Lemma 0.1 again, we then have

$$\|\mu_q - \mu_{p^*}\|_2 \leq \sqrt{\frac{\epsilon}{1-2\epsilon}}\Big(\sigma + \frac{1-\epsilon}{1-3\epsilon}\sigma\Big) = 2\frac{\sqrt{1-2\epsilon}}{1-3\epsilon}\sqrt{\epsilon}\sigma, \tag{13}$$

which proves Proposition 0.2.

0.2 Lower Bound (Breakdown at $\epsilon = 1/3$)

The $1 - 3\epsilon$ in the denominator of our bound means that Proposition 0.2 becomes vacuous once $\epsilon \geq 1/3$. Is this necessary? We will show that it indeed is.

Specifically, when $\epsilon = 1/3$, it is possible to have:

$$p^* = \frac{1}{2}\delta_{-a} + \frac{1}{2}\delta_0, \tag{14}$$

$$\tilde{p} = \frac{1}{3}\delta_{-a} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_b, \tag{15}$$

$$q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_b, \tag{16}$$

where q is a stationary point no matter how large b is. In particular, $\mu_q = \frac{b}{2}$ can be arbitrarily far away from the true mean of $-\frac{a}{2}$.

To see this more intuitively, note that an equivalent minimization problem to (2) would be to minimize $\sum_{i=1}^n q_i(x_i - \mu)^2$ with respect to both q_i and μ (since the minimizer for fixed q is always at $\mu = \mu_q$). Therefore, a stationary point is one such that:

- The q_i are concentrated on the $(1-\epsilon)n$ smallest values of $(x_i - \mu)^2$
- μ is equal to μ_q

The distribution q clearly satisfies this: we have $\mu_q = b/2$, and both 0 and b are closer to μ_q than $-a$ is.

This also shows why the breakdown point is $1/3$ and not smaller. If ϵ were slightly smaller than $1/3$, then some of the mass of q would have to remain on δ_{-a}. Then as b increases, the mean μ_q would increase more slowly, and eventually $-a$ would be closer to μ_q than b.

0.3 Auxiliary Lemmas

Proof of Lemma 0.1. Note that the proof of Lemma **??** implies that if E is an event with $q(E) \geq 1 - \epsilon$, then $\|\mathbb{E}_q[X] - \mathbb{E}_q[X \mid E]\| \leq \sqrt{\|\Sigma_q\|\frac{\epsilon}{1-\epsilon}}$. Now if q, p satisfy $\mathsf{TV}(p, q) \leq \epsilon$, there is a midpoint r that is an ϵ-deletion of both p and q. Applying the preceding result, we thus have $\|\mathbb{E}_q[X] - \mathbb{E}_r[X]\|_2 \leq \sqrt{\|\Sigma_q\|\frac{\epsilon}{1-\epsilon}}$. Similarly $\|\mathbb{E}_p[X] - \mathbb{E}_r[X]\|_2 \leq \sqrt{\|\Sigma_p\|\frac{\epsilon}{1-\epsilon}}$. The result then follows by the triangle inequality. □

Lemma 0.4. *Suppose that q, q' are both ϵ-deletions of a distribution p. Then $\mathsf{TV}(q, q') \leq \frac{\epsilon}{1-\epsilon}$.*

Proof. Conceptually, the reason Lemma 0.4 is true is that q' can be obtained from q by first adding an ϵ-fraction of points (to get to p), then deleting an ϵ-fraction. Since $\mathsf{TV} \leq \epsilon$ exactly allows an ϵ-fraction of of additions and deletions, this should yield the result. The reason we get $\frac{\epsilon}{1-\epsilon}$ is because q and q' are only a $(1-\epsilon)$-"fraction" of p, so the ϵ-deletions are more like $\frac{\epsilon}{1-\epsilon}$-deletions relative to q and q'.

To be more formal, for any set A, note that we have

$$q(A) \leq \frac{p(A)}{1-\epsilon} \text{ and } q'(A) \leq \frac{r(A)}{1-\epsilon}. \tag{17}$$

Also, using A^c instead of A, we also get

$$q(A) \geq \frac{p(A)-\epsilon}{1-\epsilon} \text{ and } q'(A) \geq \frac{p(A)-\epsilon}{1-\epsilon}. \tag{18}$$

Combining these inequalities yields

$$q(A) \leq \frac{\epsilon + (1-\epsilon)q'(A)}{1-\epsilon} \leq \frac{\epsilon}{1-\epsilon} + q'(A), \tag{19}$$

and similarly $q'(A) \leq \frac{\epsilon}{1-\epsilon} + q(A)$, which together implies $|q(A) - q'(A)| \leq \frac{\epsilon}{1-\epsilon}$. Since this holds for all A, we obtain our TV distance bound. □