0.0.1 Approximate Eigenvectors in Other Norms

Algorithm ?? is specific to the $\ell_2$-norm. Let us suppose that we care about recovering an estimate $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|$ is small in some norm other than $\ell_2$ (such as the $\ell_1$-norm, which may be more appropriate for some combinatorial problems). It turns out that an analog of bounded covariance is sufficient to enable estimation with the typical $O(\sigma \sqrt{\tau})$ error, as long as we can approximately solve the analogous eigenvector problem. To formalize this, we will make use of the dual norm:

**Definition 0.1.** Given a norm $\| \cdot \|$, the dual norm $\| \cdot \|_*$ is defined as

$$\|u\|_* = \sup_{\|v\|_2 \leq 1} \langle u, v \rangle. \quad (1)$$

As some examples, the dual of the $\ell_2$-norm is itself, the dual of the $\ell_1$-norm is the $\ell_\infty$-norm, and the dual of the $\ell_\infty$-norm is the $\ell_1$-norm. An important property (we omit the proof) is that the dual of the dual is the original norm:

**Proposition 0.2.** If $\| \cdot \|$ is a norm on a finite-dimensional vector space, then $\| \cdot \|_* = \| \cdot \|_*$.

For a more complex example: let $\|v\|(k)$ be the sum of the $k$ largest coordinates of $v$ (in absolute value). Then the dual of $\| \cdot \|(k)$ is $\max(\|u\|_\infty, \|u\|_1/k)$. This can be seen by noting that the vertices of the constraint set $\{ u \mid \|u\|_\infty \leq 1, \|u\|_1 \leq k \}$ are exactly the $k$-sparse $\{-1, 0, +1\}$-vectors.

Let $G_{\text{cov}}(\sigma, \| \cdot \|)$ denote the family of distributions satisfying $\max_{\|v\|_* \leq 1} v^\top \text{Cov}_p[X] v \leq \sigma^2$. Then $G_{\text{cov}}$ is resilient exactly analogously to the $\ell_2$-case:

**Proposition 0.3.** If $p \in G_{\text{cov}}(\sigma, \| \cdot \|)$ and $r \leq \frac{p}{1 - \sigma}$, then $\|\mu(r) - \mu(p)\| \leq \frac{2\sigma}{1 - \sigma}$. In other words, all distributions in $G_{\text{cov}}(\sigma, \| \cdot \|)$ are $(\epsilon, O(\sigma \sqrt{\tau}))$-resilient.

**Proof.** We have that $\|\mu(r) - \mu(p)\| = \langle \mu(r) - \mu(p), v \rangle$ for some vector $v$ with $\|v\|_* = 1$. The result then follows by resilience for the one-dimensional distribution $\langle X, v \rangle$ for $X \sim p$. \qed

When $p^* \in G_{\text{cov}}(\sigma, \| \cdot \|)$, we will design efficient algorithms analogous to Algorithm ?? . The main difficulty is that in norms other than $\ell_2$, it is generally not possible to exactly solve the optimization problem $\max_{\|v\|_* \leq 1} v^\top \Sigma v$ that is used in Algorithm ?? . We instead make use of a $\kappa$-approximate oracle:

**Definition 0.4.** A function $A(\Sigma)$ is a $\kappa$-approximate oracle if for all $\Sigma$, $M = A(\Sigma)$ is a positive semidefinite matrix satisfying

$$\langle M, \Sigma \rangle \geq \sup_{\|v\|_* \leq 1} v^\top \Sigma v, \text{ and } \langle M, \Sigma' \rangle \leq \kappa \sup_{\|v\|_* \leq 1} v^\top \Sigma' v \text{ for all } \Sigma' \preceq 0. \quad (2)$$

Thus a $\kappa$-approximate oracle over-approximates $\langle vv^\top, \Sigma \rangle$ for the maximizing vector $v$ on $\Sigma$, and it underapproximates $\langle vv^\top, \Sigma' \rangle$ within a factor of $\kappa$ for all $\Sigma' \neq \Sigma$. Given such an oracle, we have the following analog to Algorithm ?? :

**Algorithm 1 FilterNorm**

1. Initialize weights $c_1, \ldots, c_n = 1$.
2. Compute the empirical mean $\hat{\mu}_c$ of the data, $\hat{\mu}_c \overset{\text{def}}{=} \frac{\sum_{i=1}^n c_i x_i}{\sum_{i=1}^n c_i}$.
3. Compute the empirical covariance $\hat{\Sigma}_c \overset{\text{def}}{=} \frac{\sum_{i=1}^n c_i (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^\top}{\sum_{i=1}^n c_i}$.
4. Let $M = A(\hat{\Sigma}_c)$ be the output of a $\kappa$-approximate oracle.
5. If $\langle M, \hat{\Sigma}_c \rangle \leq 20\kappa \sigma^2$, output $q(c)$.
6. Otherwise, let $\tau_i = (x_i - \hat{\mu}_c)^\top M (x_i - \hat{\mu}_c)$, and update $c_i \leftarrow c_i \cdot (1 - \tau_i / \tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$.
7. Go back to line 2.

Algorithm 1 outputs an estimate of the mean with error $O(\sigma \sqrt{\kappa \tau})$. The proof is almost exactly the same as Algorithm ?? : the main difference is that we need to ensure that $\langle \Sigma, M \rangle$, the inner product of $M$ with the true covariance, is not too large. This is where we use the $\kappa$-approximation property. We leave the detailed proof as an exercise, and focus on how to construct a $\kappa$-approximate oracle $A$. 

[Lecture 9]
Semidefinite programming. As a concrete example, suppose that we wish to estimate $\mu$ in the $\ell_1$-norm $\|v\| = \sum_{j=1}^{d} |v_j|$. The dual norm is the $\ell_\infty$-norm, and hence our goal is to approximately solve the optimization problem

$$\text{maximize } v^\top \Sigma v \text{ subject to } \|v\|_\infty \leq 1. \quad (3)$$

The issue with (3) is that it is not concave in $v$ because of the quadratic function $v^\top \Sigma v$. However, note that $v^\top \Sigma v = (\Sigma, vv^\top)$. Therefore, if we replace $v$ with the variable $M = vv^\top$, then we can re-express the optimization problem as

$$\text{maximize } \langle \Sigma, M \rangle \text{ subject to } M_{jj} \leq 1 \text{ for all } j, \ M \succeq 0, \ \text{rank}(M) = 1. \quad (4)$$

Here the first constraint is a translation of $\|v\|_\infty \leq 1$, while the latter two constrain $M$ to be of the form $vv^\top$.

This is almost convex in $M$, except for the constraint rank($M$) = 1. If we omit this constraint, we obtain the optimization

$$\text{maximize } \langle \Sigma, M \rangle \text{ subject to } M_{jj} = 1 \text{ for all } j, \ M \succeq 0. \quad (5)$$

Note that here we replace the constraint $M_{jj} \leq 1$ with $M_{jj} = 1$; this can be done because the maximizer of (5) will always have $M_{jj} = 1$ for all $j$. For brevity we often write this constraint as $\text{diag}(M) = 1$.

The problem (5) is a special instance of a semidefinite program and can be solved in polynomial time (in general, a semidefinite program allows arbitrary linear inequality or positive semidefinite constraints between linear functions of the decision variables; we discuss this more below).

The optimizer $M^*$ of (5) will always satisfy $(\Sigma, M^*) \geq \sup_{\|v\|_\infty \leq 1} v^\top \Sigma v$ because and $v$ with $\|v\|_\infty \leq 1$ yields a feasible $M$. The key is to show that it is not too much larger than this. This turns out to be a fundamental fact in the theory of optimization called Grothendieck’s inequality:

**Theorem 0.5.** If $\Sigma \succeq 0$, then the value of (5) is at most $\frac{\pi}{2} \sup_{\|v\|_\infty \leq 1} v^\top \Sigma v$.

See? for a very well-written exposition on Grothendieck’s inequality and its relation to optimization algorithms. In that text we also see that a version of Theorem 0.5 holds even when $\Sigma$ is not positive semidefinite or indeed even square. Here we produce a proof based on [todo: cite] for the semidefinite case.

**Proof of Theorem 0.5.** The proof involves two key relations. To describe the first, given a matrix $X$ let $\arcsin[X]$ denote the matrix whose $i, j$ entry is $\arcsin(X_{ij})$ (i.e. we apply arcsin element-wise). Then we have (we will show this later)

$$\max_{\|v\|_\infty \leq 1} v^\top \Sigma v = \max_{X \succeq 0, \text{diag}(X) = 1} \frac{2}{\pi} \langle \Sigma, \arcsin[X] \rangle. \quad (6)$$

The next relation is that

$$\arcsin[X] \succeq X. \quad (7)$$

Together, these imply the approximation ratio, because we then have

$$\max_{M \succeq 0, \text{diag}(M) = 1} \langle \Sigma, M \rangle \leq \max_{M \succeq 0, \text{diag}(M) = 1} \langle \Sigma, \arcsin[M] \rangle = \frac{\pi}{2} \max_{\|v\|_\infty \leq 1} v^\top \Sigma v. \quad (8)$$

We will therefore focus on establishing (6) and (7).

To establish (6), we will show that any $X$ with $X \succeq 0$, $\text{diag}(X) = 1$ can be used to produce a probability distribution over vectors $v$ such that $\mathbb{E}[v^\top \Sigma v] = \frac{2}{\pi} \langle \Sigma, \arcsin[X] \rangle$.

First, by Graham/Cholesky decomposition we know that there exist vectors $u_i$ such that $M_{ij} = \langle u_i, u_j \rangle$ for all $i, j$. In particular, $M_{ii} = 1$ implies that the $u_i$ have unit norm. We will then construct the vector $v$ by taking $v_i = \text{sign}(\langle u_i, g \rangle)$ for a Gaussian random variable $g \sim \mathcal{N}(0, I)$.

We want to show that $\mathbb{E}_g[v_i v_j] = \frac{2}{\pi} \arcsin(\langle u_i, u_j \rangle)$. For this it helps to reason in the two-dimensional space spanned by $v_i$ and $v_j$. Then $v_i v_j = -1$ if the hyperplane induced by $g$ cuts between $u_i$ and $u_j$, and +1 if it does not. Letting $\theta$ be the angle between $u_i$ and $u_j$, we then have $\mathbb{P}[v_i v_j = -1] = \frac{\theta}{\pi}$ and hence

$$\mathbb{E}_g[v_i v_j] = (1 - \frac{\theta}{\pi}) - \frac{\theta}{\pi} = \frac{2}{\pi} (\frac{\pi}{2} - \theta) = \frac{2}{\pi} \arcsin(\langle u_i, u_j \rangle), \quad (9)$$

2
as desired. Therefore, we can always construct a distribution over $v$ for which $\mathbb{E}[v^\top \Sigma v] = \frac{2}{\pi} \langle \Sigma, \arcsin[M] \rangle$, hence the right-hand-side of (6) is at most the left-hand-side. For the other direction, note that the maximizing $v$ on the left-hand-side is always a $\{-1, +1\}$ vector by convexity of $v^\top \Sigma v$, and for any such vector we have $\frac{2}{\pi} \arcsin[vv^\top] = vv^\top$. Thus the left-hand-side is at most the right-hand-side, and so the equality (6) indeed holds.

We now turn our attention to establishing (7). For this, let $X^\otimes k$ denote the matrix whose $i,j$ entry is $X^k_{ij}$ (we take element-wise power). We require the following lemma:

**Lemma 0.6.** For all $k \in \{1, 2, \ldots\}$, if $X \succeq 0$ then $X^\otimes k \succeq 0$.

**Proof.** The matrix $X^\otimes k$ is a submatrix of $X^\otimes k$, where $(X^\otimes k)_{i_1 \cdots i_k, j_1 \cdots j_k} = X_{i_1, j_1} \cdots X_{i_k, j_k}$. We can verify that $X^\otimes k \succeq 0$ (its eigenvalues are $\lambda_{i_1} \cdots \lambda_{i_k}$ where $\lambda_i$ are the eigenvalues of $X$), hence so is $X^\otimes k$ since submatrices of PSD matrices are PSD.

With this in hand, we also make use of the Taylor series for $\arcsin(z)$: $\arcsin(z) = \sum_{n=0}^\infty \frac{(2n)!}{(2^n n!)^2} \frac{2n+1}{2n+1} z^{2n+1} = z + \frac{z^3}{6} + \cdots$. Then we have

$$\arcsin[X] = X + \sum_{n=1}^\infty \frac{(2n)!}{(2^n n!)^2} \frac{1}{2n+1} X^{\otimes (2n+1)} \succeq X,$$

as was to be shown. This completes the proof.

**Alternate proof (by Mihaela Curmei):** We can also show that $X^\otimes k \succeq 0$ more directly. Specifically, we will show that if $A, B \succeq 0$ then $A \odot B \succeq 0$, from which the result follows by induction. To show this let $A = \sum_i \lambda_i u_i u_i^\top$ and $B = \sum_j \nu_j v_j v_j^\top$ and observe that

$$A \odot B = \left(\sum_i \lambda_i u_i u_i^\top\right) \odot \left(\sum_j \nu_j v_j v_j^\top\right)$$

$$= \sum_{i,j} \lambda_i \nu_j (u_i u_i^\top) \odot (v_j v_j^\top)$$

$$= \sum_{i,j} \lambda_i \nu_j (u_i \odot v_j)(u_i \odot v_j)^\top \geq 0,$$

from which the claim follows. Here the key step is that for rank-one matrices the $\odot$ operation behaves nicely: $(u_i u_i^\top) \odot (v_j v_j^\top) = (u_i \odot v_j)(u_i \odot v_j)^\top$. 
