Mathematical methods for machine learning and signal processing SS 19

Solutions to problem set 4

Prof. Veniamin Morgenshtern

Solver: Erwin Riegler, Veniamin Morgenshtern

Problem 1: Orthogonal matching pursuit

See `omp.ipynb`

Problem 2: Recovery of approximately sparse signal using l_1-minimization

Let $\mathbf{x}^*$ be the solution to (P1) and define $\mathbf{h} = \mathbf{x}^* - \mathbf{x}$. By definition, it holds that $\mathbf{h} \in \mathcal{N}(\mathbf{D})$ since $\mathbf{Dh} = \mathbf{Dx}^* - \mathbf{Dx} = \mathbf{0}$. Moreover, we have $\|\mathbf{x}^*\|_1 \leq \|\mathbf{x}\|_1$ because $\mathbf{x}^*$ minimizes $\|\hat{\mathbf{x}}\|_1$ under the constraint that $\mathbf{D}\hat{\mathbf{x}} = \mathbf{Dx}$. If $\mathbf{h} \neq \mathbf{0}$, we have

$$\begin{aligned}
0 \geq \|\mathbf{x}^*\|_1 - \|\mathbf{x}\|_1 &= \|\mathbf{x} + \mathbf{h}\|_1 - \|\mathbf{x}\|_1 \\
&= \|\mathbf{x}_\mathcal{S} + \mathbf{h}_\mathcal{S}\|_1 + \|\mathbf{x}_{\mathcal{S}^c} + \mathbf{h}_{\mathcal{S}^c}\|_1 - \|\mathbf{x}\|_1 \\
&\geq (\|\mathbf{x}_\mathcal{S}\|_1 - \|\mathbf{h}_\mathcal{S}\|_1) + (\|\mathbf{h}_{\mathcal{S}^c}\|_1 - \|\mathbf{x}_{\mathcal{S}^c}\|_1) - \|\mathbf{x}\|_1 \\
&= (\|\mathbf{x}\|_1 - \|\mathbf{x}_{\mathcal{S}^c}\|_1) - \|\mathbf{h}_\mathcal{S}\|_1 + \|\mathbf{h}_{\mathcal{S}^c}\|_1 - \|\mathbf{x}_{\mathcal{S}^c}\|_1 - \|\mathbf{x}\|_1 \\
&= \|\mathbf{h}_{\mathcal{S}^c}\|_1 - \|\mathbf{h}_\mathcal{S}\|_1 - 2\|\mathbf{x}_{\mathcal{S}^c}\|_1 \\
&= \|\mathbf{h}\|_1 - 2\|\mathbf{h}_\mathcal{S}\|_1 - 2\|\mathbf{x}_{\mathcal{S}^c}\|_1 \\
&= \|\mathbf{h}\|_1 \left(1 - 2\frac{\|\mathbf{h}_\mathcal{S}\|_1}{\|\mathbf{h}\|_1}\right) - 2\|\mathbf{x}_{\mathcal{S}^c}\|_1 \\
&\geq \|\mathbf{h}\|_1 (1 - 2C(\mathbf{D}, \mathcal{S})) - 2\|\mathbf{x}_{\mathcal{S}^c}\|_1.
\end{aligned}$$

Note that for every vector $\mathbf{v} \in \mathbb{C}^N$, we can decompose the l_1-norm $\|\mathbf{v}\|_1$ into

$$\|\mathbf{v}\|_1 = \sum_{k=1}^{N} |v_k| = \sum_{k \in \mathcal{S}} |v_k| + \sum_{k \in \mathcal{S}^c} |v_k| = \|\mathbf{v}_\mathcal{S}\|_1 + \|\mathbf{v}_{\mathcal{S}^c}\|_1,$$

which we used several times above.

Under the condition $C(\mathbf{D}, \mathcal{S}) < 1/2$, we can conclude that

$$\|\mathbf{x}^* - \mathbf{x}\|_1 \leq \frac{2\|\mathbf{x} - \mathbf{x}_\mathcal{S}\|_1}{1 - 2C(\mathbf{D}, \mathcal{S})}. \tag{1}$$

In the case of exactly s-sparse signals, we have $\mathbf{x} = \mathbf{x}_\mathcal{S}$. Therefore, the right-hand side in (1) is equal to zero. This implies $\mathbf{x}^* = \mathbf{x}$. We thus recover the result derived in the lecture.

Problem 3: Restricted isometry property and coherence

1. If $\mathbf{A}$ satisfies the RIP of order s, then there exists $\delta \in (0, 1)$ such that

$$(1 - \delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1 + \delta)\|\mathbf{x}\|_2^2 \tag{2}$$

for all $\mathbf{x} \in \mathbb{C}^N$ such that $\left|\mathrm{supp}(\mathbf{x})\right| \leq s$. So a fortiori, since $s' \leq s$, (3) holds for all $\mathbf{x} \in \mathbb{C}^N$ such that $\left|\mathrm{supp}(\mathbf{x})\right| \leq s' \leq s$, meaning that $\mathbf{A}$ satisfies the RIP of order $s' \leq s$.

2. By denoting $\mathbf{A}_{\mathcal{S}}$ the submatrix consisting of the columns of $\mathbf{A}$ indexed by the set $\mathcal{S}$, the definition of the RIP amounts to

$$\left|\|\mathbf{A}_{\mathcal{S}}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \delta_s \|\mathbf{x}\|_2^2$$

for all $\mathbf{x} \in \mathbb{C}^s$ and $\mathcal{S} \subset [1, N]$ such that $\left|S\right| \leq s$. The term on the left-hand side can be equivalently written as $\left|\left\langle(\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}} - \mathbf{I}_s)\mathbf{x}, \mathbf{x}\right\rangle\right|$. Taking the supremum over all $\mathbf{x} \in \mathbb{C}^s$ with unit norm $\|\mathbf{x}\|_2 = 1$ yields the operator norm $\|\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}} - \mathbf{I}_s\|_{2\to2}$. We can then take the maximum over all subsets $\mathcal{S} \subset [1, N]$ of cardinality at most s to obtain the desired result.

3. The expression derived above shows that all eigenvalues of $\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}}$ are contained in the interval $[1 - \delta_s, 1 + \delta_s]$, which bounds the condition number of $\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}}$ by $\frac{1+\delta_s}{1-\delta_s}$. This means that the RIP requires that all column submatrices of $\mathbf{A}$ of size s are well-conditioned.

4. Since $\mathbf{A}$ has normalized column, the matrix $\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}} - \mathbf{I}_s$ has zero on the diagonal. The operator norm $\|\cdot\|_{1\to1}$ then yields

$$\|\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}} - \mathbf{I}_s\|_{1\to1} = \max_{j\in\mathcal{S}} \sum_{k\in\mathcal{S}\setminus\{j\}} |\langle\mathbf{a}_j, \mathbf{a}_k\rangle| \leq (s-1)\mu(\mathbf{A}).$$

Using 2., we have then

$$\delta_s = \max_{\mathcal{S}\subset[1,N],|\mathcal{S}|\leq s} \|\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}} - \mathbf{I}_s\|_{2\to2} \leq \max_{\mathcal{S}\subset[1,N],|\mathcal{S}|\leq s} \|\mathbf{A}_{\mathcal{S}}^{\mathsf{H}}\mathbf{A}_{\mathcal{S}} - \mathbf{I}_s\|_{1\to1} \leq (s-1)\mu(\mathbf{A}).$$

Problem 4: Restricted isometry property: a counterexample

Assume that $\mathbf{A}$ satisfies the RIP of order $s \geq 2$. Then, there exists $\delta \in (0, 1)$ such that

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2 \tag{3}$$

holds for all $\mathbf{x} \in \mathbb{C}^N$ such that $\left|\mathrm{supp}(\mathbf{x})\right| \leq s$. In particular, it holds for the vector $\mathbf{x} \in \mathbb{C}^N$ defined such that $x_l = 1$ for all $l \in [1, s]$ and $x_l = 0$ for $l \in [s+1, N]$. We have

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \sum_{k=1}^{M} \left|\sum_{l=1}^{N} a_{k,l} x_l\right|^2 = \sum_{k=1}^{M} \left|\sum_{l=1}^{s} \frac{1}{\sqrt{M}}\right|^2 = s^2 \quad \|\mathbf{x}\|_2^2 = s,$$

and therefore, (3) implies that $(1-\delta)s \leq s^2 \leq (1+\delta)s$, which, given that $\delta \in (0, 1)$, gives $0 < s < 2$. This contradicts the fact that $s \geq 2$.

Problem 5: Normal vector via PCA

See `normal_estimation_pca.ipynb`

Problem 6: Dual norm of the spectral norm is the nuclear norm

We seek to prove that

$$\sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \langle \mathbf{X}, \mathbf{Z}\rangle = \sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \operatorname{trace}(\mathbf{X}^\top \mathbf{Z}) = \sum_i \sigma_i(\mathbf{Z}).$$

First prove that $\sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \operatorname{trace}(\mathbf{X}^\top \mathbf{Z}) \ge \sum_i \sigma_i(\mathbf{Z})$:
Let $\mathbf{Z} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ be the singular value decomposition of $\mathbf{A}$, and define $\bar{\mathbf{X}} = \mathbf{U}\mathbf{V}^\top = \mathbf{U}\mathbf{I}\mathbf{V}^\top$. $\bar{\mathbf{X}}$ is unitary, so all of its singular values are 1, hence $\sigma_{\max}(\bar{\mathbf{X}}) = 1$ and

$$\operatorname{trace}(\bar{\mathbf{X}}^\top \mathbf{Z}) = \operatorname{trace}(\mathbf{V}\mathbf{U}^\top\mathbf{U}\Sigma\mathbf{V}^\top) = \operatorname{trace}(\mathbf{V}^\top\mathbf{V}\mathbf{U}^\top\mathbf{U}\Sigma) = \operatorname{trace}(\Sigma) = \sum_i \sigma_i(\mathbf{Z}),$$

where we used that $\operatorname{trace}(\mathbf{ABC}) = \operatorname{trace}(\mathbf{CAB})$. Since the supremum cannot be smaller than this single instance, we have

$$\sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \operatorname{trace}(\mathbf{X}^\top \mathbf{Z}) \ge \operatorname{trace}(\bar{\mathbf{X}}^\top \mathbf{Z}) = \sum_i \sigma_i(\mathbf{Z}).$$

Second prove the other direction:

$$\begin{aligned}
\sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \operatorname{trace}(\mathbf{X}^\top \mathbf{Z}) &= \sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \operatorname{trace}(\mathbf{X}^\top \mathbf{U}\Sigma\mathbf{V}^\top) \\
&= \sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \operatorname{trace}(\mathbf{V}^\top \mathbf{X}^\top \mathbf{U}\Sigma) \\
&= \sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \sum_i \sigma(\mathbf{Z})_i (\mathbf{U}\mathbf{X}\mathbf{V}^\top)_{ii} \\
&= \sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \sum_i \sigma(\mathbf{Z})_i \mathbf{u}_i \mathbf{X} \mathbf{v}_i^\top \\
&\le \sup_{\mathbf{X}:\sigma_{\max}(\mathbf{X})\le 1} \sum_i \sigma(\mathbf{Z})_i \sigma_{\max}(\mathbf{X}) \\
&= \sum_i \sigma(\mathbf{Z})_i.
\end{aligned}$$

The inequality comes from the fact that $\|\mathbf{u}_i\| = \|\mathbf{v}_i\| = 1$, and

$$\mathbf{u}_i \mathbf{X} \mathbf{v}_i^\top \le \sup_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \mathbf{u}\mathbf{X}\mathbf{v}^\top = \sigma_{\max}(\mathbf{X}).$$

Problem 7: Subgradient of a norm

Take $\mathbf{g} \in \mathcal{G}$ and let $\mathbf{y}$ be arbitrary. Then

$$\|\mathbf{x}\| + \langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle = \|\mathbf{x}\| + \langle \mathbf{g}, \mathbf{y}\rangle - \langle \mathbf{g}, \mathbf{x}\rangle = \langle \mathbf{g}, \mathbf{y}\rangle \le \|\mathbf{g}\|_* \|\mathbf{y}\| \le \|\mathbf{y}\|,$$

where in the first inequality we used Holder's inequality that states $\left|\langle \mathbf{x}, \mathbf{y}\rangle\right| \le \|\mathbf{x}\|\|\mathbf{y}\|_*$ for any dual pair of norms. Therefore, $\mathbf{g} \in \partial\|\mathbf{x}\|$ by definition of subgradient.

Problem 8: Subgradient of a nuclear norm

Take $\mathbf{G} = \mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W} \in \mathcal{G}$. We will use the characterization of subgradient from the previous exercise: if

$$\mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W} \in \{\mathbf{F} : \langle \mathbf{F}, \mathbf{M} \rangle = \|\mathbf{M}\|_*,\ \|\mathbf{F}\| \leq 1\} \tag{4}$$

then $\mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W} \in \partial\|\mathbf{M}\|_*$. Hence, it is sufficient to demonstrate (4). To do so, first observe:

$$\begin{aligned}
\left\langle \mathbf{M}, \mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W} \right\rangle &= \left\langle \mathbf{M}, \mathbf{U}\mathbf{V}^\mathsf{T} \right\rangle + \langle \mathbf{M}, \mathbf{W} \rangle \\
&= \operatorname{trace}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{V}^\mathsf{T}) + \underbrace{\operatorname{trace}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{W})}_{0} \\
&= \operatorname{trace}(\Sigma) \\
&= \|\mathbf{M}\|_*.
\end{aligned}$$

It remains to show that $\|\mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W}\| \leq 1$. To show this, take $\mathbf{x} \in \mathbb{R}^n$. Then

$$(\mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W})\mathbf{x} = (\mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W})(\mathcal{P}_\mathbf{V}\mathbf{x} + \mathcal{P}_\mathbf{V}^\perp\mathbf{x}) = \mathbf{U}\mathbf{V}^\mathsf{T}\mathcal{P}_\mathbf{V}\mathbf{x} + \mathbf{W}\mathcal{P}_\mathbf{V}^\perp\mathbf{x}.$$

Taking norm squared, and using that $\left\langle \mathbf{U}\mathbf{V}^\mathsf{T}\mathcal{P}_\mathbf{V}\mathbf{x}, \mathbf{W}\mathcal{P}_\mathbf{V}^\perp\mathbf{x} \right\rangle = \left\langle \mathbf{V}^\mathsf{T}\mathcal{P}_\mathbf{V}\mathbf{x}, \mathbf{U}^\mathsf{T}\mathbf{W}\mathcal{P}_\mathbf{V}^\perp\mathbf{x} \right\rangle = 0$

$$\|\mathbf{U}\mathbf{V}^\mathsf{T}\mathcal{P}_\mathbf{V}\mathbf{x} + \mathbf{W}\mathcal{P}_\mathbf{V}^\perp\mathbf{x}\|^2 = \|\mathbf{U}\mathbf{V}^\mathsf{T}\mathcal{P}_\mathbf{V}\mathbf{x}\|^2 + \|\mathbf{W}\mathcal{P}_\mathbf{V}^\perp\mathbf{x}\|^2 \leq \|\mathcal{P}_\mathbf{V}\mathbf{x}\|^2 + \|\mathcal{P}_\mathbf{V}^\perp\mathbf{x}\|^2 = \|\mathbf{x}\|^2,$$

which shows that $\|\mathbf{U}\mathbf{V}^\mathsf{T} + \mathbf{W}\| \leq 1$.