Review and Discussion of

Xuejun Liao
Department of Electrical and Computer Engineering
Duke University

April 18, 2014
Signal representation with a dictionary

- Signal decomposition: \( x = D\alpha \)
- When \( D \) is overcomplete, \( \alpha \) is not unique.
- Need a penalty \( g(\alpha) \) to indicate the properties expected of \( \alpha \).
- The coefficients are found as

\[
\hat{\alpha} = \arg \min_{\alpha} \frac{1}{2} \| x - D\alpha \|_2^2 + g(\alpha).
\]

- Examples of \( g(\cdot) \)
  - PCA: \( g(\alpha) = \text{Null} \).
  - k-sparsity: \( g(\alpha) = \chi(\|\alpha\|_0 = k) \), where \( \chi(p) = \begin{cases} 0, & \text{if } p \text{ is true,} \\ \infty, & \text{otherwise.} \end{cases} \)
  - Lasso: \( g(\alpha) = \chi(\|\alpha\|_1 \leq \lambda) \).
  - Ridge regression: \( g(\alpha) = \chi(\|\alpha\|_2^2 \leq \lambda) \).
  - Nonnegative matrix factorization: \( g(\alpha) = \chi(\alpha \geq 0) \).
Dictionary learning

- Consider that a distribution $\mathbb{P} \in \mathcal{P}$ governs the generation of $x$.
- Look for a dictionary $D \in \mathcal{D}$ that work for the entire $\mathbb{P}$, by minimizing

$$f_x(D) = \min_{\alpha} \frac{1}{2} \|x - D\alpha\|^2_2 + g(\alpha)$$

over all samples $x \sim \mathbb{P}$.
- Achieve the goal by minimizing the expected loss function,

$$D^* = \arg \min_{D \in \mathcal{D}} \mathbb{E}_{x \sim \mathbb{P}} f_x(D).$$

- Examples of $\mathcal{D}$
  - Column-orthonormal matrices (PCA).
  - Column-normalized matrices (some compressive sensing systems).
  - Sparse matrices.
The problem addressed in the papers

- Given finite samples \( X = [x_1, \ldots, x_n] \), the empirical loss function is

\[
F_X(D) = \frac{1}{n} \sum_{i=1}^{n} f_{x_i}(D) = \frac{1}{n} \sum_{i=1}^{n} \left( \min_{\alpha_i} \frac{1}{2} \| x_i - D\alpha_i \|_2^2 + g(\alpha_i) \right),
\]

based on which one can get

\[
\hat{D}_n = \arg \min_D F_X(D)
\]

- The question is: How does \( \hat{D}_n \) perform compared to \( D^* \)?
- The goal is to bound the performance of \( \hat{D} \) relative to that of \( D^* \), i.e., to find the upper bound of

\[
|F_X(\hat{D}) - E_{x \sim P} f_x(D^*)|.
\]

- Relaxation: uniform convergence bound

\[
\sup_{D \in \mathcal{D}} |F_X(D) - E_{x \sim P} f_x(D)| \leq \text{upper bound.}
\]
Main result

Let $L_f = \sup_{D,D' \in \mathcal{D}} \frac{|f(D) - f(D')|}{\|D - D'\|}$ be the Lipschitz constant of function $f$, and

$$\Lambda_n(L) = \text{Prob} \left( \max(L_{F_X}, F_{E_f}) > L \right),$$
$$\Gamma_n(\gamma) = \sup_{D \in \mathcal{D}} \text{Prob} \left( |F_X(D) - \mathbb{E}_{x \sim \mathbb{P}} f_x(D)| > \gamma \right).$$

Suppose there exists $\exists c, T > 0$ (which depend only on $\mathcal{D}$) such that $\Gamma_n(\gamma) \leq 2e^{-n\tau^2}/c^2$ holds true $\forall \tau \in [0, cT]$.

Let $C, h \geq 1$ be constants depending only on $\mathcal{D}$ and $\beta = h \max(\log \frac{2LC}{c}, 1)$.

Then, given $x \in [0, nT^2 - \beta]$, it holds, with probability exceeding $1 - 2e^{-x} - \Lambda_n(L)$, that

$$\sup_{D \in \mathcal{D}} |F_X(D) - \mathbb{E}_{x \sim \mathbb{P}} f_x(D)| \leq 3c \sqrt{\frac{\beta \log n}{n}} + c \sqrt{\frac{\beta + x}{n}}.$$
A sketch of the proof

Consider an \( \epsilon \)-cover of \( \mathcal{D} \) with a finite covering number

\[
\mathcal{N}(\mathcal{D}, \epsilon) = \min \left\{ \#\mathcal{Q} : \mathcal{D} \subset \bigcup_{q \in \mathcal{Q}} B_\epsilon(q) \right\}.
\]

For a given \( \mathbf{D} \in \mathcal{D} \), let \( \mathbf{D}_j \) be its closest neighbor in \( \mathcal{Q} \). Since \( \|\mathbf{D} - \mathbf{D}_j\| \leq \epsilon \), one has

\[
|F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D})| \leq |F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_j)| + |F_{\mathbf{X}}(\mathbf{D}_j) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D}_j)| + |\mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D}_j) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D})|,
\]

\[
\leq L_{F_X} \|\mathbf{D} - \mathbf{D}_j\| + \max_{1 \leq j \leq \mathcal{N}} |F_{\mathbf{X}}(\mathbf{D}_j) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D}_j)| + L_{\mathbb{E}_{f_{\mathbf{X}}}} \|\mathbf{D} - \mathbf{D}_j\|
\]

\[
\leq L\epsilon + \gamma + L\epsilon.
\]

Next step: find the bound of \( \sup_{\mathbf{D} \in \mathcal{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D})| \). Need to consider all \( \epsilon \)-balls.

- The last inequality holds unless one or more of the following events occur:
  
  (i) \( \max(L_{F_X}, F_{\mathbb{E}_{f_{\mathbf{X}}}}) > L \);  
  
  (ii) \( |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D})| > \gamma \), \( \exists j \).

- By assumption, event (i) occurs with probability \( \Lambda_n(L) \), and even (ii) occurs with probability at most \( \mathcal{N}(\mathcal{D}, \epsilon) \Gamma_n(\gamma) \).

By the union bound, \( \sup_{\mathbf{D} \in \mathcal{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{f_{\mathbf{X}}}(\mathbf{D})| \leq 2L\epsilon + \gamma \) holds with probability at least

\[
1 - \Lambda_n(L) - \mathcal{N}(\mathcal{D}, \epsilon) \Gamma_n(\gamma).
\]

- The final form of the theorem is obtained by a detailed analysis of \( \mathcal{N}(\mathcal{D}, \epsilon) \) and \( \Gamma_n(\gamma) \).
Relevance to our work

- Sample complexity for dictionary learning or GMM training in compressive sensing.
- Challenges – the underlying patches may not be drawn from the same distribution as the training patches.
- Possible new work – how could the in-situ linear measurements help to reduce the sample complexity?