The sparse coding problem

- unknown vectors $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$

$$n \begin{bmatrix} a_1 & A \end{bmatrix}^m$$

- input: random sparse combination $y = Ax = \sum_{i=1}^m x_i a_i$

$X$ has $\leq k$ nonzero entries
(for simplicity, will assume nonzero entries are far from 0)

- goal: recover $A, x$ with enough samples.

Simpler problem: compressed sensing

Given $A$, $y = Ax$, find sparse $x$

- easy when $n \geq m$ (system of equations)
- more interesting when $n < m$.

- when $n < m$, problem is hard in general.
  but when $\{a_i\}$'s have nice properties there are many algorithms

Incoherence: the set of vectors $\{a_i\}$ is $\mu$-incoherent, if $\forall i \neq j$

$$|\langle a_i, a_j \rangle| \leq \frac{\mu}{\sqrt{n}}$$

(usually $\mu = \text{constant or } \text{log } n$)

Intuition: vectors are almost orthogonal
"looks like" orthonormal basis for sparse $x$.

Easy Decoding: Lemma: If $\{a_i\}$'s are $\mu$-incoherent, then

$$|\langle y, a_i \rangle - x_i| \leq M \mu$$
- Easy reading. Lemma: \( \mathbf{E} \) entries are \( \mu \)-incoherent, then

\[
| \langle y, a_i \rangle - x_i | \leq \frac{\mu k}{\sqrt{n}}
\]

in particular, when \( \frac{\mu k}{\sqrt{n}} \ll 1 \) can find the nonzero entries!

**Proof:** \( \langle y, a_i \rangle = \langle \sum_{j=1}^{m} x_j a_j, a_i \rangle \)

\[
= x_i + \sum_{j \neq i} x_j \langle a_j, a_i \rangle \quad \text{\( \leq \)} \quad \text{sparse, \( \leq k \) nonzero} \quad \leq \frac{\mu}{\sqrt{n}}
\]

\[
= x_i \pm \frac{\mu k}{\sqrt{n}}.
\]

- **Approximate Gradient Descent for Sparse coding.**

**Objective function**

\[
f(B, X) = \| Y - BX \|_F^2
\]

\[
\begin{bmatrix}
Y_i \\
p
\end{bmatrix}
= \begin{bmatrix}
A \\
m
\end{bmatrix} \begin{bmatrix}
X_i \\
p
\end{bmatrix}
\]

want: \( \min B \) \( f(B, X) \)

s.t. \( X \) is sparse.

- Observation: \( f \) is “degree 4” in both \( BX \).

  if we fix \( B \) or \( X \), \( f \) is a quadratic and convex.

- Alternating minimization:

  Fix \( B \), decode \( X \), then fix \( X \), find \( B \).

  hope \( \frac{\partial}{\partial B} f(B, X) \approx \frac{\partial}{\partial B} f(B, X^*) \)

  \( \text{gradient of an} \)
- The algorithm:
  Given current matrix $B$ (hopefully $B \approx A$)

  for each sample $y$
  find support $S$ of $x^*$ (support = set of nonzero entries)
  (let $x_S = B_S^Ty$)

  estimate the gradient

  \[
  \frac{\partial}{\partial B} f(B, x) = -E \left[ z (y - Bx)^T x^* \right] = -E \left[ z (A_Sx_S^* - B_SB_S^TA_Sx_S^*) x^T \right] \quad (x)
  \]

- recall
  - Def: $g$ is $(\alpha, \beta, \varepsilon)$-correlated if
    \[
    \langle \nabla g(z^t), z^t - z^* \rangle \geq \alpha \| z^t - z^* \|^2 + \beta \| \nabla g(z^t) \| ^2 - \varepsilon
    \]

  - gradient $(x)$ satisfies this, but not very easy to proof.

---

**Initialization**

- idea: need to extract information about $A$ or $x$, even though we only know $y$.

- plan: find all samples that share the same component, then the top singular direction for these samples is close to $A$.
- Observation:

\[ \langle y^i, y^j \rangle = \langle Ax^i, Ax^j \rangle \]
\[ = \sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} x^i_{\alpha} x^j_{\beta} \langle a^\alpha, a^\beta \rangle \]

only \( k^2 \) nonzero entries.

so \( |\langle y^i, y^j \rangle| \) large if \( x^i, x^j \) share one nonzero entry

\( |\langle y^i, y^j \rangle| \) small if \( \langle x^i, x^j \rangle = 0 \)