Lecture 17 Sparse coding

Sunday, October 30, 2016 5:38 PM

- The sparse coding problem
  - unknown vectors $a_1, a_2, \dots, a_m \in \mathbb{R}^n$

$$n\left[\begin{array}{c|c} a_i & A \end{array}\right]$$
$$m$$

  - 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 μ-incoherent, if $\forall\, i \neq j$

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

  (usually μ = constant or $\log n$)

  intuition: vectors are almost orthogonal

  "looks like" orthonormal basis for sparse x.
- Easy Decoding: Lemma: If $\{a_i\}$'s are μ-incoherent, then

$$|\langle y, a_i \rangle - x_i| \leq \mu k$$

- Easy decoding: Lemma: If $\{a_i\}$s are μ-incoherent, then

$$\left| \langle y, a_i \rangle - x_i \right| \le \frac{\mu k}{\sqrt{n}}$$

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

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

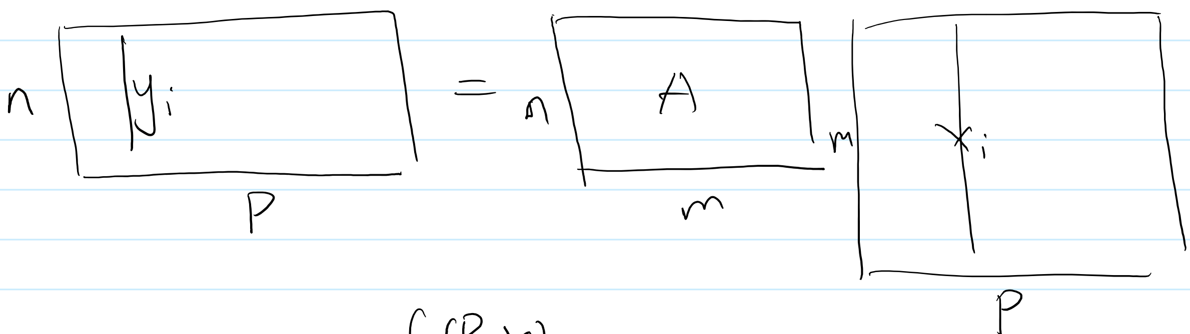
$$= x_i + \sum_{j \ne i} \underbrace{x_j}_{\substack{\text{sparse} \\ \le k \text{ nonzero}}} \underbrace{\langle a_j, a_i \rangle}_{\le \frac{\mu}{\sqrt{n}}}$$

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

- Approximate Gradient Descent for Sparse Coding.

objective function

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



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

s.t. X is sparse.

- observation: f is "degree 4" in both B X.

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

- alternating minimization:

Fix B, decode X, then fix X, find B.

hope $\frac{\partial}{\partial B} f(B, X) \approx \underbrace{\frac{\partial}{\partial B} f(B, X^*)}_{\text{gradient of an}}$

unknown convex function

- 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 $x_S = B_S^T y$

  estimate the gradient

$$\frac{\partial}{\partial B} f(B, x) = -\mathbb{E}\left[z(y - Bx)x^T\right]$$
$$= -\mathbb{E}\left[z(A_S x_S^* - B_S B_S^T A_S x_S^*)x^T\right] \quad (*)$$

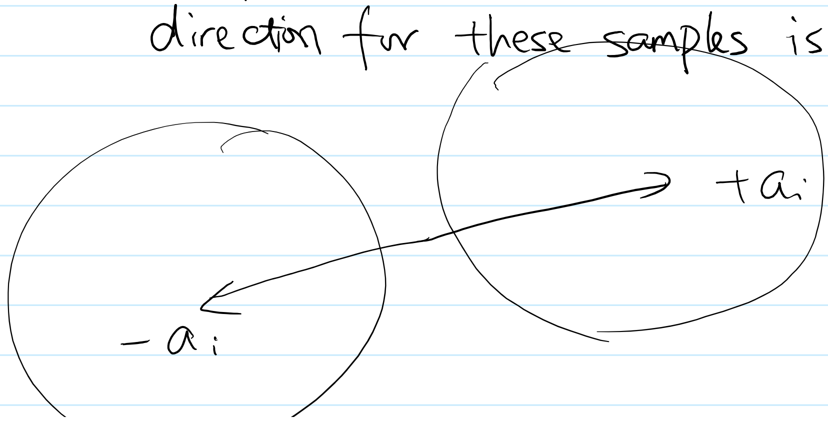
- recall
  - Def: g is $(\alpha, \beta, \varepsilon)$-correlated if

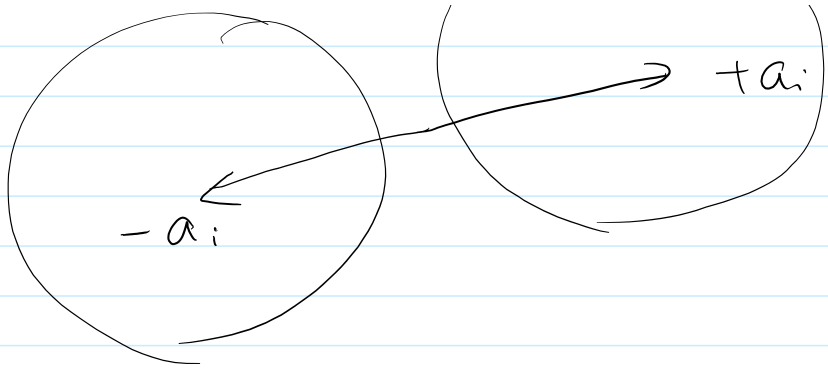
$$\langle \nabla g(z^t), z^t - z^* \rangle \geq \alpha \|z^t - z^*\|^2 + \beta \|\nabla g(z^t)\|^2 - \varepsilon$$

  - gradient $(*)$ 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_i





- observation:

$$\langle y^i, y^j \rangle = \langle Ax^i, Ax^j \rangle$$
$$= \sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} \underbrace{x^i_\alpha x^j_\beta}_{\text{only } k^2 \text{ nonzero entries.}} \langle a_\alpha, a_\beta \rangle$$

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 $\langle x^i, x^j \rangle = 0$