What is the best online algorithm for PCA?

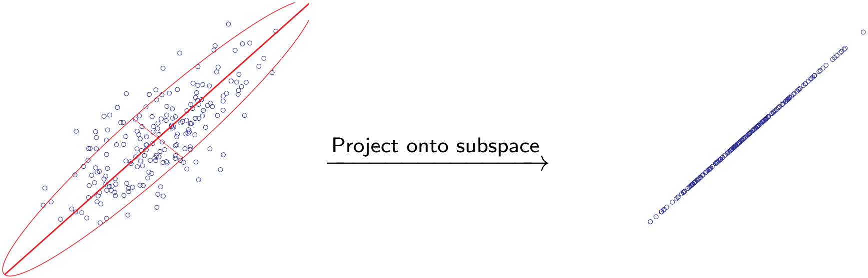
Jiazhong Nie
Wojciech Kotlowski
Manfred K. Warmuth

University of California - Santa Cruz

July 5, 2013

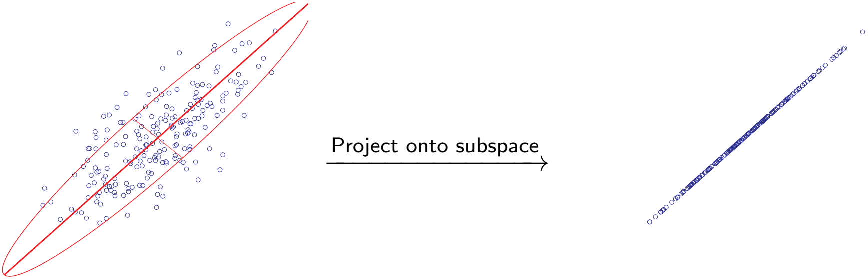
Online PCA

- Batch PCA



Online PCA

- Batch PCA



- Online PCA
  - Before see next point, learner picks random subspace
  - Then suffers compression loss
- Keep uncertainty over subspaces with density matrix
- Two main algorithms: EG and GD

EG seems best for online PCA

- Known regret bound for EG:

$$\sqrt{k \ln \frac{n}{k} \ \text{loss of best } k \text{ subspace}}$$

EG seems best for online PCA

- Known regret bound for EG:

$$\sqrt{k \ln \frac{n}{k} \ \text{loss of best } k \text{ subspace}}$$

- Only logarithmic dependence on the dimension n
- EG seems best for simplex-like constraints
- Mathematically elegant
  - Quantum Relative Entropy, matrix log, matrix exponential
  - But computationally expensive

What about GD for PCA?

Motivated by the squared Frobenius norm $\|\mathbf{W}\|_F^2 = \sum_{1 \le i,j \le n} (w_{i,j})^2$.

$$\mathbf{W}_{t+1} = \inf_{\substack{\mathbf{W} \text{ dens. matrix} \\ \text{w. eigenvals} \le \frac{1}{n-k}}} \|\mathbf{W} - \mathbf{W}_t\|_F^2 + \eta \, \mathrm{tr}(\mathbf{W}\mathbf{x}_t\mathbf{x}_t^\top),$$

What about GD for PCA?

Motivated by the squared Frobenius norm $\|\mathbf{W}\|_F^2 = \sum_{1 \le i,j \le n} (w_{i,j})^2$.

$$\mathbf{W}_{t+1} = \inf_{\substack{\mathbf{W} \text{ dens. matrix} \\ \text{w. eigenvals} \le \frac{1}{n-k}}} \|\mathbf{W} - \mathbf{W}_t\|_F^2 \; + \; \eta \, \mathrm{tr}(\mathbf{W} \mathbf{x}_t \mathbf{x}_t^\top),$$

Rewrite into two steps:

Unconstrained min.: $\widehat{\mathbf{W}}_t = \mathbf{W}_t - \eta \, \mathbf{x}_t \mathbf{x}_t^\top$

Projection: $\mathbf{W}_{t+1} = \inf_{\substack{\mathbf{W} \succeq \mathbf{0} \\ \text{w. eigenvals} \le \frac{1}{n-k} \\ tr(\mathbf{W}) = 1}} \|\mathbf{W} - \widehat{\mathbf{W}}_t\|_F^2$

Regret bound for GD

$$\leq \sqrt{2\frac{k(n-k)}{n}T}$$

$$\overset{n \gg k}{\approx} \sqrt{kT}$$

Regret bound for GD

$$\leq \sqrt{2\frac{k(n-k)}{n}T}$$
$$\overset{n \gg k}{\approx} \quad \sqrt{kT}$$

- No dependence on n?
- Better than EG?

Improved regret bound for EG for the PCA case

New: Exploit the fact that instances are sparse

$$\sqrt{2(n-k)\underbrace{\log\frac{n}{n-k}}_{\leq\frac{k}{(n-k)}}\quad\underbrace{\text{loss of best } k \text{ subspace}}_{\leq\frac{n-k}{n}T}}+(n-k)\log\frac{n}{n-k}$$

$$\leq\sqrt{2k\ T}+k$$

Improved regret bound for EG for the PCA case

New: Exploit the fact that instances are sparse

$$\sqrt{2(n-k)\underbrace{\log\frac{n}{n-k}}_{\le\frac{k}{(n-k)}}\;\underbrace{\text{loss of best } k \text{ subspace}}_{\le\frac{n-k}{n}T}} + (n-k)\log\frac{n}{n-k}$$

$$\le \sqrt{2k\ T} + k$$

- Same bound as GD - but not better?

Improved regret bound for EG for the PCA case

New: Exploit the fact that instances are sparse

$$\sqrt{2(n-k)\underbrace{\log\frac{n}{n-k}}_{\leq\frac{k}{(n-k)}}\;\underbrace{\text{loss of best } k \text{ subspace}}_{\leq\frac{n-k}{n}T}}+(n-k)\log\frac{n}{n-k}$$

$$\leq\sqrt{2k\;T}+k$$

- Same bound as GD - but not better?
- Are these Quantum Relative Entropies unnecessary?

EG's revenge

Budget bound for EG instead of **time bound**

$$\leq \sqrt{2k \text{ loss of best } k \text{ subspace}} + k$$

EG's regret is optimal

Regret of GD

$$\geq k\sqrt{\text{loss of best } k \text{ subspace}}$$

EG's revenge

Budget bound for EG instead of **time bound**

$$\leq \sqrt{2k \text{ loss of best } k \text{ subspace}} + k$$

EG's regret is optimal

Regret of GD

$$\geq k\sqrt{\text{loss of best } k \text{ subspace}}$$

EG is better than GD by a factor of $\sqrt{k}$

EG's revenge

Budget bound for EG instead of **time bound**

$$\leq \sqrt{2k \text{ loss of best } k \text{ subspace}} + k$$

EG's regret is optimal

Regret of GD

$$\geq k\sqrt{\text{loss of best } k \text{ subspace}}$$

EG is better than GD by a factor of $\sqrt{k}$

Synopsis:

- **Time bounds are crude - focus on high loss case**

EG's revenge

Budget bound for EG instead of **time bound**

$$\leq \sqrt{2k \text{ loss of best } k \text{ subspace}} + k$$

EG's regret is optimal

Regret of GD

$$\geq k\sqrt{\text{loss of best } k \text{ subspace}}$$

EG is better than GD by a factor of $\sqrt{k}$

Synopsis:

- **Time bounds are crude - focus on high loss case**
- **Budget bounds are more refined - can focus on small loss case**

GD seems to be hurt by "forgetting"

Alternate **Incremental Off-line** GD:

$$\mathbf{W}_{t+1} = \inf_{\substack{\mathbf{W}\text{ dens. matrix}\\ \text{w. eigenvals} \le \frac{1}{n-k}}} \underbrace{\|\mathbf{W}-\mathbf{W}_0\|_F^2}_{\text{Divergence from first parameter}} + \eta \underbrace{\sum_{q=1}^{t} \operatorname{tr}(\mathbf{W}\mathbf{x}_q\mathbf{x}_q^\top)}_{\text{all data up to trail } t}$$

GD seems to be hurt by "forgetting"

Alternate **Incremental Off-line** GD:

$$\mathbf{W}_{t+1} = \inf_{\substack{\mathbf{W} \text{ dens. matrix} \\ \text{w. eigenvals} \leq \frac{1}{n-k}}} \underbrace{\|\mathbf{W} - \mathbf{W}_0\|_F^2}_{\text{Divergence from first parameter}} + \eta \underbrace{\sum_{q=1}^{t} \mathrm{tr}(\mathbf{W}\mathbf{x}_q\mathbf{x}_q^\top)}_{\text{all data up to trail } t}$$

- **Incremental Offline** GD might have a budget bound as good as EG ?