LOW-RANK APPROXIMATION IN INPUT-SPARSITY TIME, WITH REGULARIZATION ~~AND ROBUSTNESS~~

HAIM AVRON
TEL AVIV UNIVERSITY

KEN CLARKSON
IBM RESEARCH, ALMADEN

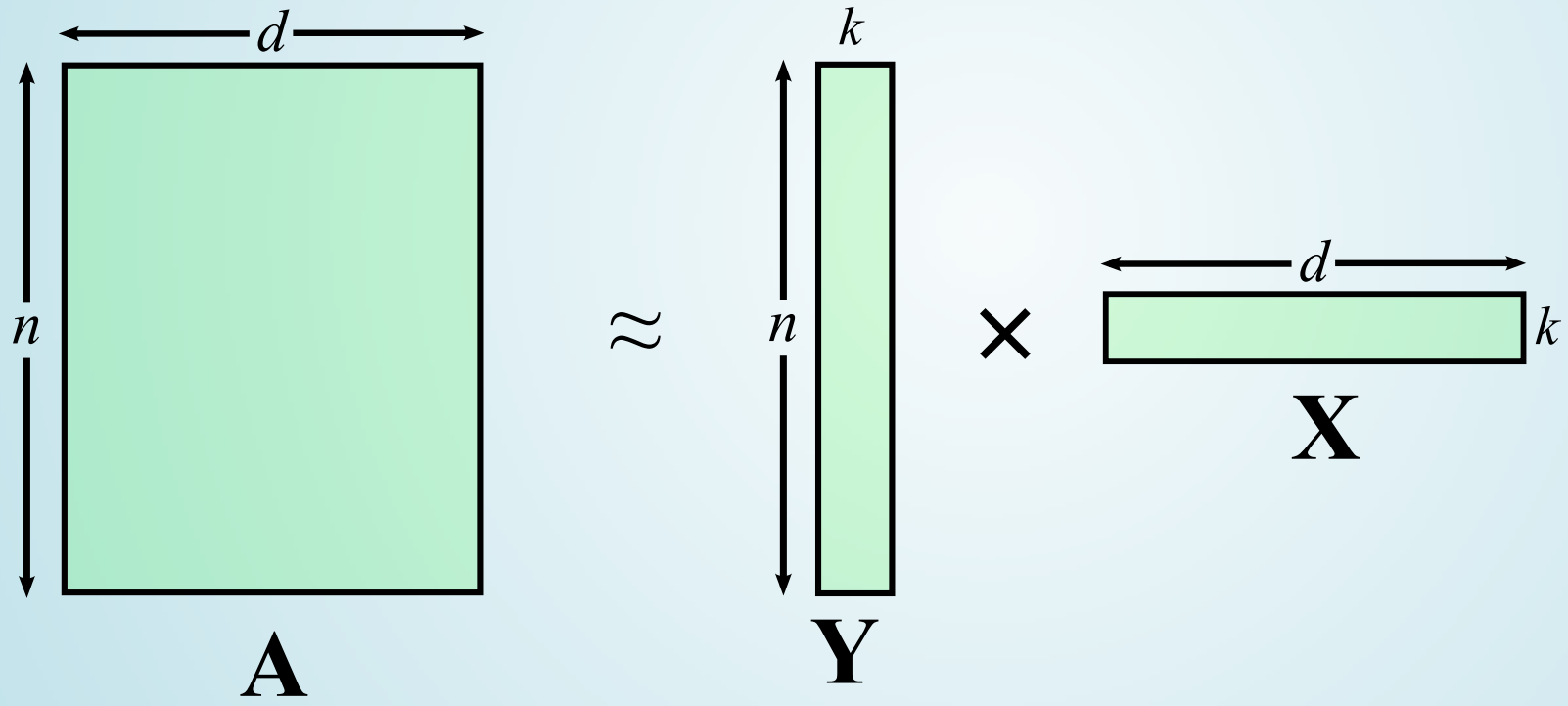
DAVID WOODRUFF
IBM RESEARCH, ALMADEN

THE PROBLEM

Regularized low-rank approximation

Given
$\mathbf{A} \in R^{n \times d}$, integer k, $\lambda \geq 0$,
find

$$\mathbf{Y}_*, \mathbf{X}_* \equiv \operatorname{argmin}_{\substack{\mathbf{Y} \in \mathbf{R}^{n \times k} \\ \mathbf{X} \in \mathbf{R}^{k \times d}}} \|\mathbf{YX} - \mathbf{A}\|_F^2 + \lambda(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2)$$



FUN FACTS

Equivalent to nuclear (trace) norm minimization [SS05]:

$$\operatorname{argmin}_{\substack{\mathbf{Y}\in\mathbf{R}^{n\times k}\\ \mathbf{X}\in\mathbf{R}^{k\times d}}} \|\mathbf{YX}-\mathbf{A}\|_F^2 + 2\lambda\|\mathbf{YX}\|_*$$

For best rank-k $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$, solution is

$$\mathbf{Y}_* = \mathbf{U}_k\mathbf{\Lambda}, \mathbf{X}_* = \mathbf{\Lambda}\mathbf{V}_k^\top,$$

where $\mathbf{\Lambda} \equiv (\mathbf{\Sigma}_k - \lambda\mathbf{I}_k)_+^{1/2}$, and $()_+$ takes negative entries to zero.

The nuclear norm "encourages" low rank;
constraint insists on rank $\leq k$

WHY THIS PROBLEM?

- Suppose
$$\mathbf{A} = \hat{\mathbf{A}} + \mathbf{N},$$
where $\mathrm{rank}(\hat{\mathbf{A}}) \leq k$, and $\mathbf{N}$ is noise
- Then $\sigma_i(\mathbf{A}) \approx \sigma_i(\hat{\mathbf{A}}) + \sigma$,
for $\mathbf{N}$ i.i.d. Gaussian noise appropriate variance
  - Singular values $\sigma_i()$
  - Singular vectors of $\mathbf{A}$ lightly perturbed from $\hat{\mathbf{A}}$
  - So recover $\hat{\mathbf{A}}$ with the right λ
- More generally,
$$\|\mathbf{YX} - \hat{\mathbf{A}}\|_F^2 \leq \|\mathbf{YX} - \mathbf{A}\|_F^2 - \|\mathbf{N}\|_F^2 + 2\|\mathbf{N}\|_2\|\mathbf{YX}\|_*,$$
so minimizing $\|\mathbf{YX} - \mathbf{A}\|_F^2 + \|\mathbf{N}\|_2\|\mathbf{YX}\|_*$ is a plausible proxy for minimizing $\|\mathbf{YX} - \hat{\mathbf{A}}\|_F^2$

APPROXIMATE SOLUTIONS

Regularized low-rank approximation, relative error ϵ

Given
$\epsilon > 0, \lambda \geq 0$
a $(1+\epsilon)$-approximation $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$ has

$$\|\tilde{\mathbf{Y}}\tilde{\mathbf{X}} - \mathbf{A}\|_F^2 + \lambda(\|\tilde{\mathbf{X}}\|_F^2 + \|\tilde{\mathbf{Y}}\|_F^2)$$
$$\leq (1+\epsilon)\left[\|\mathbf{Y}_*\mathbf{X}_* - \mathbf{A}\|_F^2 + \lambda(\|\mathbf{X}_*\|_F^2 + \|\mathbf{Y}_*\|_F^2)\right]$$

input-sparsity-time algorithm

With fixed probability near one,
$(1+\epsilon)$-approximation can be found in

$$O(\mathrm{nnz}(\mathbf{A})) + \tilde{O}(k^2\epsilon^{-1}(n+d)) + \mathrm{poly}(\epsilon^{-1}k)$$

time.

RIDGE REGRESSION

Given $\mathbf{Y}_*$,

$$\mathbf{X}_* = \operatorname{argmin}_{\mathbf{X}\in\mathbf{R}^{k\times d}} \|\mathbf{Y}_*\mathbf{X} - \mathbf{A}\|_F^2 + \lambda\|\mathbf{X}\|_F^2,$$

that is, *multiple-response ridge regression*.

$\mathbf{X}_*$ in rowspan($\mathbf{A}$)

Observation. Suppose $P_\mathbf{A}$ projects to the rowspace of $\mathbf{A}$, so for $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^\top P_\mathbf{A} \in \operatorname{rowspan}(\mathbf{A})$. Then

$$\|(\mathbf{Y}_*\mathbf{X} - \mathbf{A})P_\mathbf{A}\|_F^2 + \lambda\|\mathbf{X}P_\mathbf{A}\|_F^2 \le \|\mathbf{Y}_*\mathbf{X} - \mathbf{A}\|_F^2 + \lambda\|\mathbf{X}\|_F^2,$$

so without loss of generality, each row of $\mathbf{X}_*$ is in $\operatorname{rowspan}(\mathbf{A})$.

To solve these regression problems,
we reduce to a smaller problem using *sparse embeddings*

SPARSE EMBEDDINGS

oblivious subspace embedding

A distribution over matrices $\mathbf{S} \in \mathbb{R}^{m \times n}$
so that for any given $\mathbf{Y} \in \mathbb{R}^{n \times d'}$,
with fixed probability near one,

$$\|\mathbf{SYx}\| \approx_\epsilon \|\mathbf{Yx}\| \text{ for all } \mathbf{x} \in \mathbb{R}^{d'}$$

---

if $\mathbf{S}$ is a sparse embedding for $\mathbf{Y}$,
then for all $\mathbf{X} \in \mathbb{R}^{d' \times d}$,
$\|\mathbf{SYX}\|_F^2 \approx_\epsilon \|\mathbf{YX}\|_F^2$

Also called *sketching* matrices

oblivious affine embeddings

A distribution over matrices $\mathbf{S} \in \mathbb{R}^{m \times n}$
so that for any given $\mathbf{Y} \in \mathbb{R}^{n \times d'}$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$,
with fixed probability near one,

$$\|\mathbf{S}(\mathbf{YX} - \mathbf{A})\|_F^2 \approx_\epsilon \|\mathbf{YX} - \mathbf{A}\|_F^2 \text{ for all } \mathbf{X} \in \mathbb{R}^{d'}$$

---

sparse oblivious affine embeddings

There are distributions for which $\mathbf{S}$ has one non-zero per column,
so $\mathbf{SY}$ can be computed in $\text{nnz}(\mathbf{Y})$ time, and
$\mathbf{S} \in \mathbb{R}^{m_\mathbf{S} \times d'}$ with $m_\mathbf{S}$ small.

MULTIPLE REDUCTIONS

If $\mathbf{S}$ is an affine embedding, then

$$\tilde{\mathbf{X}} = \operatorname{argmin}_{\mathbf{X} \in \mathbf{R}^{k \times d}} \|\mathbf{S}(\mathbf{Y}_* \mathbf{X} - \mathbf{A})\|_F^2 + \lambda \|\mathbf{X}\|_F^2$$

is a good approximation to $\mathbf{X}_*$.

Applying observation above to $\operatorname{rowspan}(\mathbf{SA})$,
WLOG $\tilde{\mathbf{X}} = \mathbf{ZSA}$ for some $\mathbf{Z} \in \mathbb{R}^{k \times m_\mathbf{S}}$

Applying again "on the other side":

two-sided reduction

There are sparse affine embeddings $\mathbf{S}$ and $\mathbf{R}^\top$ so that

$$\mathbf{W}_*, \mathbf{Z}_* \equiv \operatorname{argmin}_{\substack{\mathbf{W} \in \mathbf{R}^{m_\mathbf{R} \times k} \\ \mathbf{Z} \in \mathbf{R}^{k \times m_\mathbf{S}}}} \|\mathbf{ARW\,ZSA} - \mathbf{A}\|_F^2 + \lambda(\|\mathbf{ZSA}\|_F^2 + \|\mathbf{ARW}\|_F^2)$$

yield $(1+\epsilon)$-approximation $\tilde{\mathbf{X}} \equiv \mathbf{Z}_* \mathbf{SA}$, $\tilde{\mathbf{Y}} \equiv \mathbf{ARW}_*$.

With additional affine *and* sparse embeddings $\hat{\mathbf{S}}, \hat{\mathbf{R}}$, can reduce to

$$\operatorname{argmin}_{\substack{\mathbf{W} \in \mathbf{R}^{m_\mathbf{R} \times k} \\ \mathbf{Z} \in \mathbf{R}^{k \times m_\mathbf{S}}}} \|\hat{\mathbf{S}}\mathbf{ARW\,ZSA}\hat{\mathbf{R}} - \hat{\mathbf{S}}\mathbf{A}\hat{\mathbf{R}}\|_F^2 + \lambda(\|\mathbf{ZSA}\hat{\mathbf{R}}\|_F^2 + \|\hat{\mathbf{S}}\mathbf{ARW}$$

a subproblem where all matrices have $\operatorname{poly}(k/\epsilon)$ rows and columns.

RIDGE REGRESSION, REPRISE

Suppose we want to apply sparse embedding machinery to

$$\mathbf{x}_* = \operatorname{argmin}_{\mathbf{x}\in\mathbf{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|^2,$$

that is, solve instead

$$\min_{\mathbf{x}\in\mathbf{R}^d} \|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|^2 + \lambda\|\mathbf{x}\|^2.$$

When λ is very large, $\lambda\|\mathbf{x}\|^2$ dominates, and $\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|^2 \approx \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ can be loose.

So ridge regression should get easier as λ increases.

THE STATISTICAL DIMENSION OF RIDGE REGRESSION

statistical dimension

The statistical dimension (regression degrees of freedom, effective dimension) of ridge regression is

$$\mathrm{sd}_\lambda(\mathbf{A}) \equiv \sum_i \frac{1}{1+\lambda/\sigma_i^2(\mathbf{A})},$$

where $\sigma_i(\mathbf{A})$ is the i'th singular value of $\mathbf{A}$.

Note that
$\mathrm{sd}_\lambda(\mathbf{A}) \leq \mathrm{rank}(A) \leq \min\{n, d\}$, and
$\mathrm{sd}_\lambda(\mathbf{A})$ is decreasing in λ.

SMALLER EMBEDDINGS FOR RIDGE REGRESSION

smaller ridge

A ridge regression problem with n rows can be reduced to one with $\tilde{O}(\epsilon^{-1}\mathtt{sd}_\lambda(\mathbf{A}))$ rows, in $O(\mathrm{nnz}(\mathbf{A})) + \tilde{O}(d(\epsilon^{-1}\mathtt{sd}_\lambda(\mathbf{A}) + \mathtt{sd}_\lambda(\mathbf{A})^2))$ time.
The resulting problem yields a $(1+\epsilon)$-approximation.

- This extends prior work from sampling to sketching
- For low-rank approximation, some $k \to O(\epsilon^{-1}\mathtt{sd}_\lambda(\mathbf{Y}_*))$
- There is another reduction that reduces d to

$$\mathrm{poly}(\mathtt{sd}_\lambda(\mathbf{A})\epsilon^{-1}(\sigma_1(A)^2/\lambda))$$

CONCLUSIONS

- Better dependence on k and ϵ in "lower order terms" for this class
- Algorithms get faster as λ gets large
- We don't actually know $\mathtt{sd}_\lambda(\mathbf{A})$
- Also: can get pre-conditioners for kernel ridge regression
- Results generalize to regularizers $f(\mathbf{X}, \mathbf{Y})$ satisfying some orthogonal invariance conditions

Thank you for your attention!