Laura Balzano, University of Michigan

Stephen Wright, University of Wisconsin

Local Convergence of an Incremental Algorithm for Subspace Identification

✧Incremental Gradient

- ✧ When a cost function can be written as a sum of costs on "data blocks," Incremental gradient performs cost function optimization one "data block" at a time.
- ✧ Great for real-time or big data applications.
- ✧ Convergence rates are poor within a local region of the solution, as compared to steepest descent or second-order methods.

✧Manifold Optimization

- ✧ When a non-linear constraint set can be written as a Riemannian manifold, we can use manifold methods for optimization.
- ✧ Convergence results require armijo step which sometimes adds a large computational burden.

✧Incremental Gradient

- When a cost function can be written as a sum of costs on "data blocks," Incremental gradient performs cost function optimization one "data block" at a time.

Consider a least-squares problem of the form

$$\text{minimize}_x f(x) = \sum_{i=1}^{n} \|g_i(x)\|^2 .$$

✧Incremental Gradient

$$\text{minimize}_x f(x) = \sum_{i=1}^{n} \|g_i(x)\|^2 \ .$$

Now consider the same problem but where $g_i(x)$ is a linear function of data block i, $i = 1, \ldots, m$ and the incremental gradient algorithm given by [Bertsekas 99, p116] with step size α_k at iteration k. Let x^* be the optimal solution corresponding to this problem. Then:

1. There exists $\bar{\alpha} > 0$ such that if α_k is equal to some constant $\alpha \in (0, \bar{\alpha}]$ for all k, the sequence x_k converges to some vector $x(\alpha)$. Furthermore, the error $\|x_k - x(\alpha)\|$ converges to 0 linearly. Finally, we have $\lim_{\alpha \to 0} x(\alpha) = x^*$.

2. If $\alpha_k > 0$ for all k, and

$$\alpha_k \to 0, \ \sum_{k=0}^{\infty} \alpha_k = \infty \ ,$$

then $\{x_k\}$ converges to x^*.

✧Optimization on Manifolds

Consider any optimization problem on a Riemannian manifold $\mathcal{M}$ with a retraction given from the tangent space of $\mathcal{M}$ to $\mathcal{M}$. Perform any gradient-related descent algorithm using the Armijo step size on a manifold [Absil, Mahony, Sepulchre 08, p62].

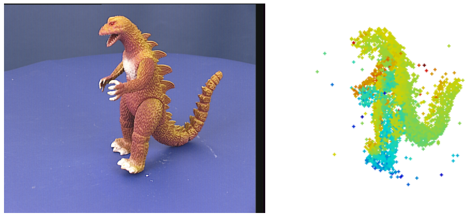
Then every limit point of the sequence of iterates is a critical point of the cost function; i.e. $\nabla f = 0$.

Outline

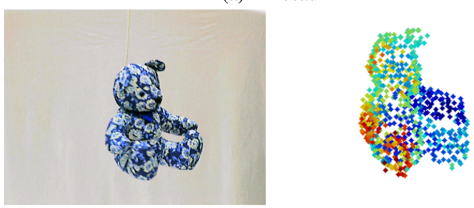
MICHIGAN ENGINEERING
UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

- ✧Subspace Tracking with Missing Data
- ✧GROUSE algorithm convergence rate in the full-data case
- ✧GROUSE algorithm convergence rate with missing data
- ✧Equivalence of grouse to a kind of missing-data incremental SVD

Applications that use Subspaces of R^n

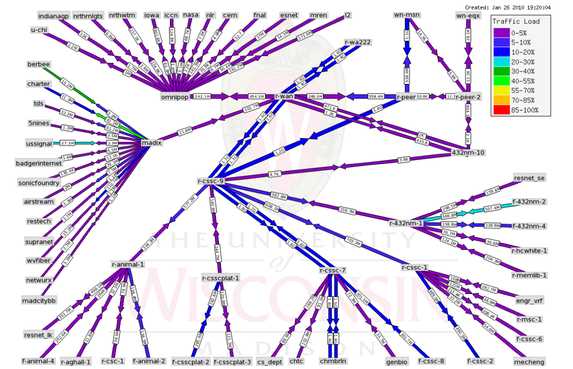


(a) Dinosaur



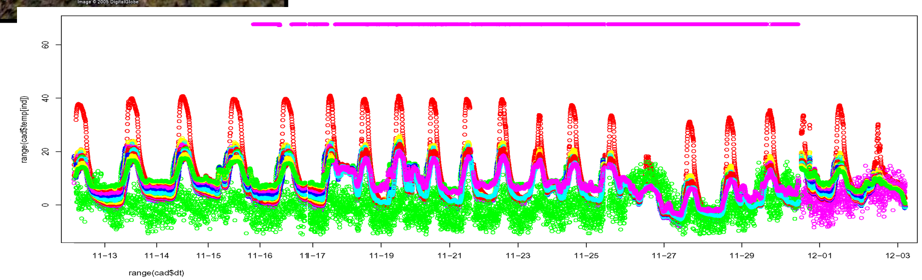
(b) Teddy Bear

3D object modeling: when points are matched across frames, they lie in a 3D subspace.



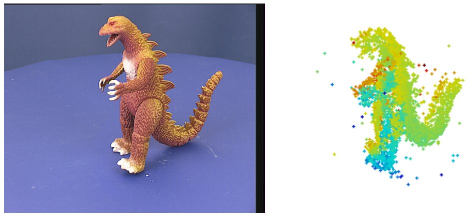
Network data analysis: due to network connectivity constraining the flows, traffic data lie in a low dimensional subspace

Ranking based on human assessment: people's preferences have been demonstrated to lie near a low-dimensional manifold; we are using a handful of factors only

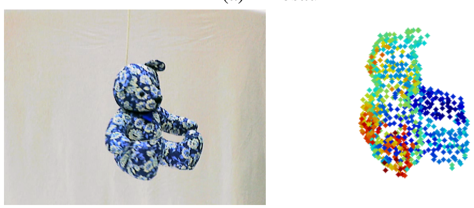


Sensor network data analysis: very spatially correlated data lie near a low-dimensional subspace

These Applications all have Missing Data

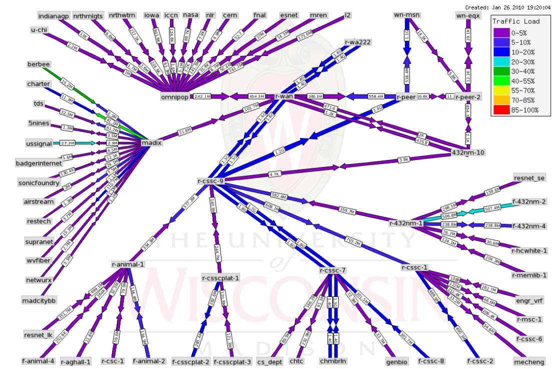


(a) Dinosaur



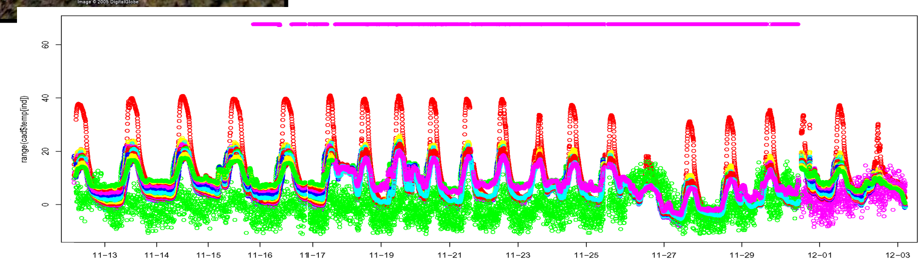
(b) Teddy Bear

3D object modeling: missing data due to obstruction from different camera angles



Network data analysis: missing data due to massive throughput

Ranking based on human assessment: missing data due to impossibility of considering all alternatives

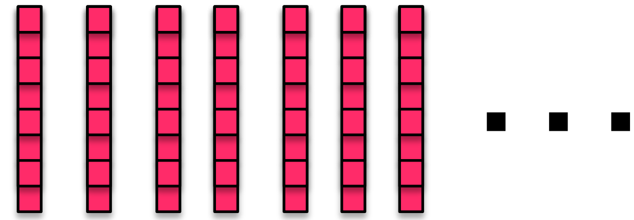


Sensor network data analysis: missing data due to cheap sensors and crummy communication links

Subspace Identification: Full Data

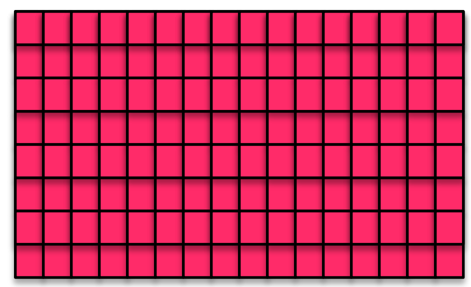
Suppose we receive a sequence of length-n vectors that lie in a d-dimensional subspace S:

$$v_1, v_2, \ldots, v_t, \ldots, \in S \subset \mathbb{R}^n$$



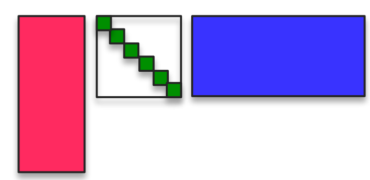
And then we collect T of these vectors into a matrix,

$$X = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \ldots & v_T \\ | & | & & | \end{bmatrix}$$



If S is static, we can identify it as the column space of this matrix by performing the SVD:

$$X = U \Sigma V^T \, .$$

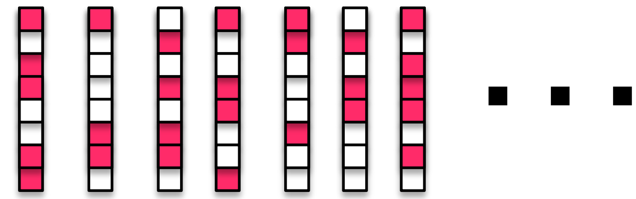


The orthogonal columns of U span the subspace S.

Subspace Identification: Missing Data

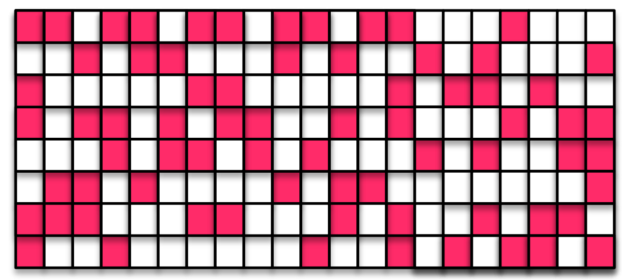
Suppose we receive a sequence of incomplete length-n vectors that lie in a d-dimensional subspace S, and $\Omega_t \subset \{1, \ldots, n\}$ refers to the observed indices:

$$[v_1]_{\Omega_1}, [v_2]_{\Omega_2}, \ldots, [v_t]_{\Omega_t}, \ldots, \in S \subset \mathbb{R}^n$$



And then we collect T of these vectors into a matrix:

$$X = \begin{bmatrix} | & | & & | \\ [v_1]_{\Omega_1} & [v_2]_{\Omega_2} & \cdots & [v_T]_{\Omega_T} \\ | & | & & | \end{bmatrix}$$



~~If S is static, we can identify it as the column space of this matrix by performing the SVD:~~

$$X = U \Sigma V^T \, .$$

The orthogonal columns of U span the subspace S.

Problem Set-Up

- Seek subspace $S \subset \mathbb{R}^n$ of known dimension $d \ll n$.
- Know certain components $\Omega_t \subset \{1, 2, \dots, n\}$ of vectors $v_t \in S$, $t = 1, 2, \dots$ — the subvector $[v_t]_{\Omega_t}$.
- Assume that $\mathcal{S}$ is incoherent w.r.t. the coordinate directions.

We'll also assume for purposes of analysis that

- $v_t = \bar{U} s_t$, where $\bar{U}$ is an $n \times d$ orthonormal spanning $\mathcal{S}$ and the components of $s_t \in \mathbb{R}^d$ are i.i.d. normal with mean 0.
- Sample set Ω_t is independent for each t with $|\Omega_t| \geq q$, for some q between d and n.
- Observation subvectors $[v_t]_{\Omega_t}$ contain no noise.

Problem Set-Up

We take an incremental gradient approach to minimizing over $\mathcal{S}$ the function

$$F(\mathcal{S}) = \sum_{i=1}^{T} \|[v_i - P_{\mathcal{S}} v_i]_{\Omega_i}\|_2^2 \ .$$

Since the variable is a subspace we optimize on the Grassmannian.

GROUSE

Given current estimate U_t and partial data vector $[v_t]_{\Omega_t}$, where $v_t = \bar{U}s_t$:

$$\begin{aligned}
w_t &:= \arg\min_w \|[U_t w - v_t]_{\Omega_t}\|_2^2; \\
p_t &:= U_t w_t; \\
[r_t]_{\Omega_t} &:= [v_t - U_t w_t]_{\Omega_t}; \quad [r_t]_{\Omega_t^c} := 0; \\
\sigma_t &:= \|r_t\|\|p_t\|; \\
\text{Choose } \eta_t &> 0; \\
U_{t+1} &:= U_t + \left[(\cos\sigma_t\eta_t - 1)\frac{p_t}{\|p_t\|} + \sin\sigma_t\eta_t\frac{r_t}{\|r_t\|}\right]\frac{w_t^T}{\|w_t\|};
\end{aligned}$$

We focus on the (locally acceptable) choice

$$\eta_t = \frac{1}{\sigma_t}\arcsin\frac{\|r_t\|}{\|p_t\|}, \quad \text{which yields } \sigma_t\eta_t = \arcsin\frac{\|r_t\|}{\|p_t\|} \approx \frac{\|r_t\|}{\|p_t\|}.$$

To measure the discrepancy between the current estimate span(U_t) and $\mathcal{S}$, we use the angles between the two subspaces. There are d angles between two d-dimensional subspaces, and we call them $\phi_{t,i}$, $i = 1, \ldots, d$, where

$$\cos \phi_{t,i} = \sigma_i(U_t^T \bar{U}) \ ,$$

where σ_i denotes the i^{th} singular value. Define

$$\epsilon_t := \sum_{i=1}^{d} \phi_{t,i} = d - \sum_{i=1}^{d} \sigma_i(U_t^T \bar{U})^2 = d - \|U_t^T \bar{U}\|_F^2 \ .$$

We seek a bound for $\mathbb{E}[\epsilon_{t+1} | \epsilon_t]$, where the expectation is taken over the random vector s_t for which $v_t = \bar{U} s_t$.

Outline

- Subspace Tracking with Missing Data
- GROUSE algorithm convergence rate in the full-data case
- GROUSE algorithm convergence rate with missing data
- Equivalence of grouse to a kind of missing-data incremental SVD

Full-data case vastly simpler to analyze than the general case. Define

- $\theta_t := \arccos(\|p_t\|/\|v_t\|)$ is the angle between $R(U_t)$ and $\mathcal{S}$ that is revealed by the update vector v_t;
- Define $A_t := U_t^T \bar{U}$, $d \times d$, nearly orthogonal when $R(U_t) \approx \mathcal{S}$. We have $\epsilon_t = d - \|A_t\|_F^2$.

Lemma

$$\epsilon_t - \epsilon_{t+1} = \frac{\sin(\sigma_t \eta_t) \sin(2\theta_t - \sigma_t \eta_t)}{\sin^2 \theta_t} \left(1 - \frac{s_t^T A_t^T A_t A_t^T A_t s_t}{s_t^T A_t^T A_t s_t} \right),$$

The right-hand side is nonnegative for $\sigma_t \eta_t \in (0, 2\theta_t)$, *and zero if* $v_t \in R(U_t) = \mathcal{S}_t$ *or* $v_t \perp \mathcal{S}_t$.

MICHIGAN ENGINEERING
UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

Theorem

Suppose that $\epsilon_t \leq \bar{\epsilon}$ *for some* $\bar{\epsilon} \in (0, 1/3)$. *Then*

$$E\left[\epsilon_{t+1} \,|\, \epsilon_t\right] \leq \left(1 - \left(\frac{1-3\bar{\epsilon}}{1-\bar{\epsilon}}\right)\frac{1}{d}\right)\epsilon_t.$$

Since the sequence $\{\epsilon_t\}$ is decreasing, by the earlier lemma, we have $\epsilon_t \downarrow 0$ with probability 1 when started with $\epsilon_0 \leq \bar{\epsilon}$.

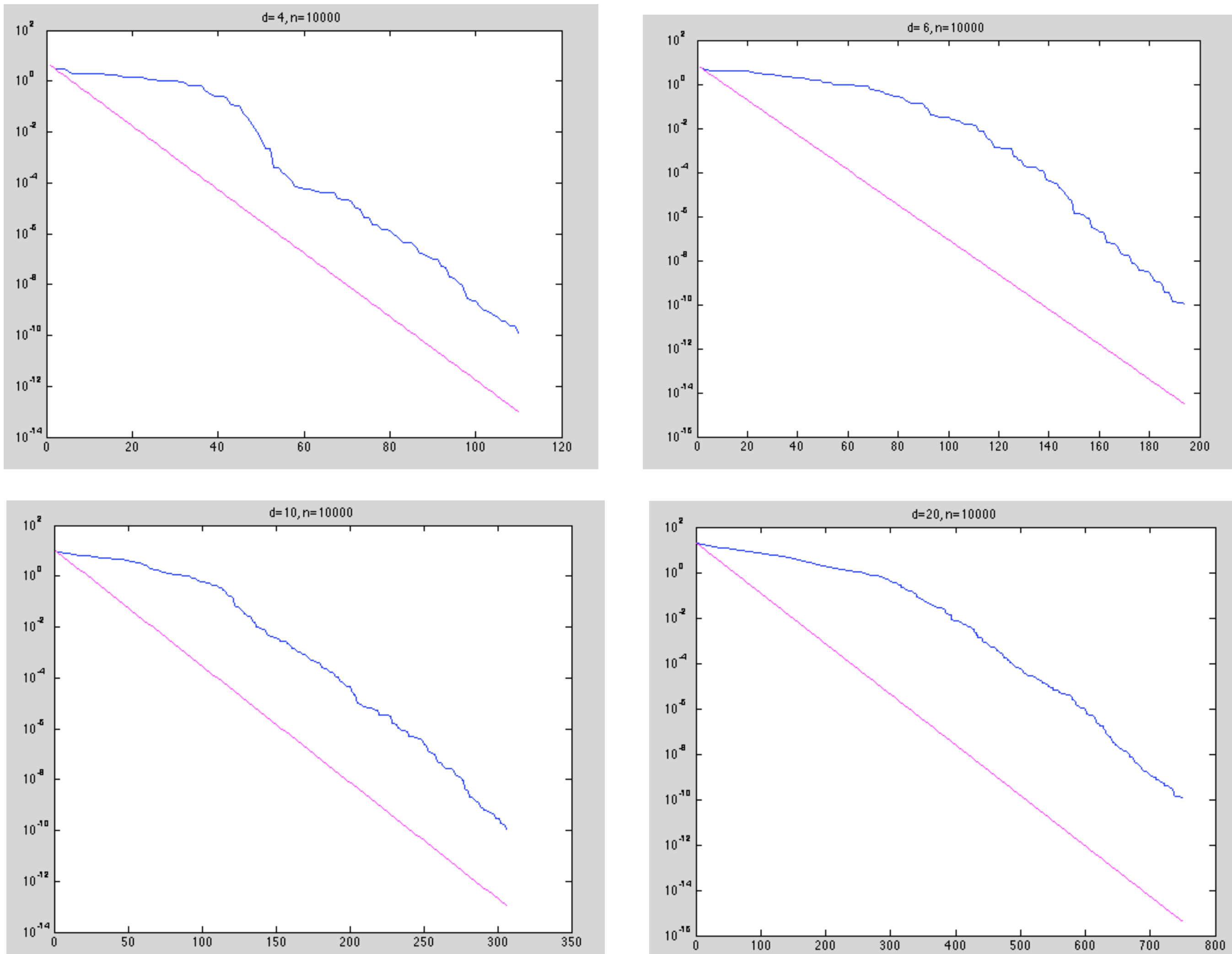
Linear convergence rate is asymptotically $1 - 1/d$.

- For $d = 1$, get near-convergence in one step (thankfully!)
- Generally, in d steps (which is number of steps to get the exact solution using SVD), improvement factor is

$$(1 - 1/d)^d < \frac{1}{e}.$$

ε_t versus 1-1/d

n=10000
d=4, 6, 10, 20



Outline

MICHIGAN ENGINEERING
UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

- Subspace Tracking with Missing Data
- GROUSE algorithm convergence rate in the full-data case
- GROUSE algorithm convergence rate with missing data
- Equivalence of grouse to a kind of missing-data incremental SVD

Our Result for the General Case

Michigan Engineering
University of Michigan ▪ College of Engineering

Recall, n is the ambient dimension, d the inherent dimension, we have $|\Omega| > q$ samples per vector. We have assumptions on the number of samples, the coherence in the subspaces and in the residual vectors, and we require that these assumptions hold with probability $1-\delta$ for $\delta \in (0, .6)$. Then for

$$\epsilon_t \leq (8 \times 10^{-6})(.6-\delta)^2 \frac{q^3}{n^3 d^2}$$

we have

$$\mathbb{E}[\epsilon_{t+1} | \epsilon_t] \leq \left(1 - (.16)(.6-\delta)\frac{q}{nd}\right)\epsilon_t \ .$$

Comments

$$\epsilon_t \leq (8 \times 10^{-6})(.6 - \delta)^2 \frac{q^3}{n^3 d^2}$$

$$\mathbb{E}[\epsilon_{t+1} | \epsilon_t] \leq \left(1 - (.16)(.6 - \delta)\frac{q}{nd}\right) \epsilon_t \ .$$

The decrease constant is not too far from that observed in practice; we see a factor of about

$$1 - X\frac{q}{nd}$$

where X is not much less than 1.

The threshold condition on ϵ_t, however, is quite pessimistic. Linear convergence behavior is seen at much higher values.

Outline

- GROUSE algorithm convergence rate in the full-data case
- GROUSE algorithm convergence rate with missing data
- Equivalence of grouse to a kind of missing-data incremental SVD

The standard iSVD

MICHIGAN ENGINEERING
UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

Algorithm 2 iSVD: Full Data

Given U_0, an arbitrary $n \times d$ orthonormal matrix, with $0 < d < n$; Σ_0, a $d \times d$ diagonal matrix of zeros which will later hold the singular values, and V_0, an arbitrary $n \times d$ orthonormal matrix.

for $t = 0, 1, 2, \ldots$ **do**

Take the current data column vector v_t;

Define $w_t := \arg\min_w \|U_t w - v\|_2^2 = U_t^T v_t$;

Define

$$p_t := U_t w_t; \quad r_t := v_t - p_t;$$

Noting that

$$\begin{bmatrix} U_t \Sigma_t V_t^T & v_t \end{bmatrix} = \begin{bmatrix} U_t & \frac{r_t}{\|r_t\|} \end{bmatrix} \begin{bmatrix} \Sigma_t & w_t \\ 0 & \|r_t\| \end{bmatrix} \begin{bmatrix} V_t & 0 \\ 0 & 1 \end{bmatrix}^T,$$

we compute the SVD of the update matrix:

$$\begin{bmatrix} \Sigma_t & w_t \\ 0 & \|r_t\| \end{bmatrix} = \hat{U}\hat{\Sigma}\hat{V}^T,$$

and set

$$U_{t+1} := \begin{bmatrix} U_t & \frac{r_t}{\|r_t\|} \end{bmatrix} \hat{U}, \quad \Sigma_{t+1} = \hat{\Sigma}, \quad V_{t+1} = \begin{bmatrix} V_t & 0 \\ 0 & 1 \end{bmatrix} \hat{V}.$$

end for

How do we incorporate missing data?

- We could put zeros into the matrix
  - Very interesting recent results from Sourav Chatterjee on one-step “Universal Singular Value Thresholding” show that zero-filling followed by SVD reaches the minimax lower bound on MSE.
  - But in the average case, we see that convergence of the zero-filled SVD is very very slow.
- Let’s instead replace the missing entries with our prediction using the existing model

iSVD with missing data 2

Algorithm 4 iSVD: Partial Data, Forget singular values

Given U_0, an $n \times d$ orthonormal matrix, with $0 < d < n$;

for $t = 0, 1, 2, \ldots$ **do**

Take Ω_t and v_{Ω_t} from (2.1);

Define $w_t := \arg\min_w \|U_{\Omega_t} w - v_{\Omega_t}\|_2^2$;

Define vectors $\tilde{v}_t$, p_t, r_t:

$$(\tilde{v}_t)_i := \begin{cases} v_i & i \in \Omega_t \\ (U_t w_t)_i & i \in \Omega_t^C \end{cases}; \quad p_t := U_t w_t; \quad r_t := \tilde{v}_t - p_t;$$

Noting that

$$\begin{bmatrix} U_t & \tilde{v}_t \end{bmatrix} = \begin{bmatrix} U_t & \frac{r_t}{\|r_t\|} \end{bmatrix} \begin{bmatrix} I & w_t \\ 0 & \|r_t\| \end{bmatrix},$$

we compute the SVD of the update matrix:

$$\begin{bmatrix} I & w_t \\ 0 & \|r_t\| \end{bmatrix} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T,$$

and set $U_{t+1} := \begin{bmatrix} U_t & \frac{r_t}{\|r_t\|} \end{bmatrix} \widetilde{U}_{:,1:d} W_t$, where W_t is an arbitrary $d \times d$ orthogonal matrix.

end for

GROUSE and iSVD equivalence

MICHIGAN ENGINEERING
UNIVERSITY of MICHIGAN ■ COLLEGE of ENGINEERING

Theorem

Suppose we have the same U_t and $[v_t]_{\Omega_t}$ at the t-th iterations of iSVD and GROUSE. Then there exists $\eta_t > 0$ in GROUSE such that the next iterates U_{t+1} of both algorithms are identical, to within an orthogonal transformation by the $d \times d$ matrix

$$W_t := \left[\frac{w_t}{\|w_t\|} \mid Z_t \right],$$

where Z_t is a $d \times (d-1)$ orthonormal matrix whose columns span $N(w_t^T)$.

The precise values for which GROUSE and iSVD are identical are:

$$\lambda = \frac{1}{2}\left[(\|w_t\|^2 + \|r_t\|^2 + 1) + \sqrt{(\|w_t\|^2 + \|r_t\|^2 + 1)^2 - 4\|r_t\|^2} \right]$$

$$\beta = \frac{\|r_t\|^2\|w_t\|^2}{\|r_t\|^2\|w_t\|^2 + (\lambda - \|r_t\|^2)^2}$$

$$\eta_t = \frac{1}{\sigma_t} \arcsin \beta.$$

Future Directions

- Apply GROUSE analysis to ell-1 version, GRASTA
- Re-think the proof from new angles.
  - We see convergence at higher ε.
  - We see monotonic decrease at any random initialization.
  - We see convergence even without incoherence (but good steps are only made when the samples align).

Thank you!

Questions?