non-convex approaches to convex optimization: asymmetric low-rank solutions

constantine caramanis

the university of texas at austin
constantine@utexas.edu

joint work with xinyang yi, yudong chen, dohyung park

June 21, 2016

non-convex problems

a success story of high dimensional statistics: identifying/exploiting non-linear and non-convex structure with low intrinsic dimension.

non-convex problems

many examples of this:

- sparse solutions to regression problems.
- low rank solutions.
- matrix completion
- corrupted low-rank recovery (robust PCA)
- corrupted matrix completion
- etc...

convex optimization to the rescue

for f a convex function:

$$\min : f(X), \quad s.t.\ X \in \mathscr{X}$$

this paradigm has allowed us to solve many problems.

convex optimization to the rescue

key insight: while original problems are non-convex, their solutions are (often, approximate) solutions to convex problems.

robust PCA

separating a low rank from a sparse matrix (robust PCA)

$$Y^* = L^* + S^*$$

L^* a rank r matrix. S^* sparse.

robust PCA

a convex formulation:

$$\begin{aligned} \text{min}: \quad & \|L\|_* + \lambda \|S\|_1 \\ \text{s.t.}: \quad & L + S = Y^*. \end{aligned}$$

solvable by SDP (Chandrasekaran et al. '09, Candes et al. '09)

robust PCA

a convex formulation:

$$\begin{aligned} \text{min}: \quad & \|L\|_* + \lambda \|S\|_1 \\ \text{s.t.}: \quad & L + S = Y^*. \end{aligned}$$

this can recover L^* from a constant fraction of randomly placed corruptions, or $1/r$ fraction of arbitrarily placed corruptions.

robust PCA

but we're never satisfied....

robust PCA

reading in the data: $O(d^2)$.

writing $\hat{L}$: $O(rd)$.

rank-r svd: $O(rd^2)$.

how much more for computation?

computational costs

computation is expensive:[1]

$$\min : \ \|L\|_* + \lambda \|S\|_1, \quad \text{s.t.} : L + S = Y^*.$$

- $O(d^4)$ for SDP.
- $O(d^3)$ Inexact Aug. L. Mult. (Lin, Chen, Ma '13)
- $O(r^2 d^2)$ for AltProj (Netrapalli et al., '14)

convex and non-convex approaches for *convex optimization*.

[1] dependence on incoherence, condition number and error also important.

computational costs

a different path: *non-convex optimization*.

$$\min_{U,V,S} : \|UV^\top + S - Y\|_F^2, \quad \text{s.t.} : U, V \in \mathbb{R}^{d \times r}.$$

- the factorization forces the low-rank constraint.
- only store: $d \times r$ matrices, and sparse $d \times d$ matrix S.
- bilinear (not convex) in U and V.

computational costs

a different path: *non-convex optimization*.

$$\min_{U,V,S} : \|UV^\top + S - Y\|_F^2, \quad \text{s.t.} : U, V \in \mathbb{R}^{d \times r}.$$

algorithm:

- initialize (important/expensive)
- iterate:
  - (projected) gradient descent on U and on V.
  - update S by sorting $(UV^\top - Y)$.

main cost: $r \times d$ matrix multiplication each iteration (no *SVD*).

some results and an application

computational results (details to follow)

- if r (rank) of L is big: $O(rd^2)$.
- if $r \ll d^{1/3}$: $O(r^4 d \log d)$
- this also works *in spite of/because of* many erasures.

foreground-background separation

- *separate foreground and background in video with static background*
- approach: stack frames – background is low-rank, foreground is sparse corruption.

foreground-background separation

video.

foreground-background separation

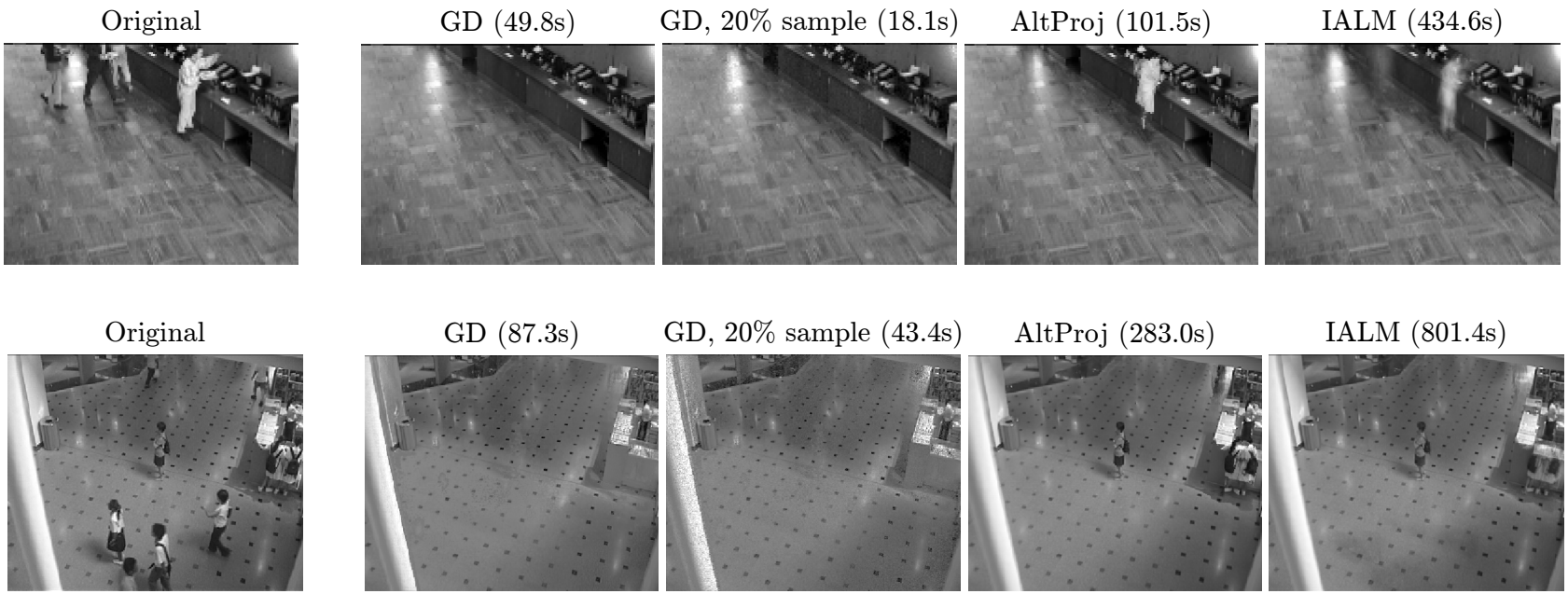

Figure: Foreground-background separation in *Restaurant* and *ShoppingMall* videos.

foreground-background separation

computation time comparison: restaurant video is 101 seconds [2]

- inexact approx. lagrange multipler (Lin et al. '13): 434.6 sec
- alternating projection (Netrapalli et al. '14): 101.5 sec
- factored gradient descent (this work): 49.8 sec
- sparsified gradient descent (this work): 18.1 sec

[2] we found similar results in other experiments.

related work

factored approach for low-rank problems

- alternating minimization: Jain, Netrapalli, Sanghavi '13, Hardt '14, Gu, Wang, Liu '16,...
- gradient descent: Bhojanapalli, Jain, Sanghavi '15, Bhojanapalli, Kyrillidis, Sanghavi '16, Chen, Wainwright '15, Sun, Luo '15, Tu et al. '15, Zhao, Wang, Liu '15, Zheng, Lafferty '15,...
- related work in EM, phase retrieval, and elsewhere.

fast running times, but no (or weak) robustness guarantees.

related work

additional challenges

- L is not symmetric or PSD (prior work requires this).
- dealing with erasures while guaranteeing robustness.

notation and assumptions

- L rank r and $d \times d$.
- incoherence of L^* controlled by μ.
- condition number of L^* controlled by κ.
- corruptions are arbitrary but sparse: αd per row/column.

the algorithm*

cost function:

$$\begin{aligned} \text{min}: \quad & \|UV^\top + S - Y\|_F^2 + \|UU^\top - VV^\top\|_F^2 \\ \text{s.t.}: \quad & S\ \alpha-\text{sparse},\ U, V\ \mu-\text{incoherent}. \end{aligned}$$

initialization:

$$S_{\text{init}} \leftarrow \mathscr{T}_\alpha[Y]$$
$$[L, \Sigma, R] \leftarrow \text{SVD}_r[Y - S_{\text{init}}],\ U_0 \leftarrow L\Sigma^{1/2},\ V_0 \leftarrow R\Sigma^{1/2}.$$

gradient iterations:

$$S_t \leftarrow \mathscr{T}_\alpha\left[Y - U_t V_t^\top\right]$$
$$U_{t+1} \leftarrow \Pi_{\mathscr{U}}\left(U_t - \eta\nabla_U\mathscr{L}(U_t, V_t; S_t) - \tfrac{1}{2}\eta\, U_t(U_t^\top U_t - V_t^\top V_t)\right)$$
$$V_{t+1} \leftarrow \Pi_{\mathscr{V}}\left(V_t - \eta\nabla_V\mathscr{L}(U_t, V_t; S_t) - \tfrac{1}{2}\eta\, V_t(V_t^\top V_t - U_t^\top U_t)\right)$$

analysis framework – gradient descent

locally, the following hold

- (descent) if far from *an optimal solution*, gradient step improvement is proportional to $\|UV^\top - L^*\|_F^2$.
- (smooth) gradient vanishes with $\|UV^\top - L^*\|_F^2$.

consequence: gradient is steep and can take big step size $\sim 1/\sigma_1^*$.

analysis framework – gradient descent

where the two conditions are used:

$$
\begin{aligned}
&\| U_{t+1} - U^t_{\pi^*} \|_F^2 + \| V_{t+1} - V^t_{\pi^*} \|_F^2 \\
&\leq \\
&\| U_t - \eta \nabla_U \mathscr{L}_t - \eta \nabla_U \mathscr{G}_t - U^t_{\pi^*} \|_F^2 + \| V_t - \eta \nabla_V \mathscr{L}_t - \eta \nabla_V \mathscr{G}_t - V^t_{\pi^*} \|_F^2 \\
&\leq \\
&\delta_t - 2\eta \underbrace{\langle\langle \nabla_U \mathscr{L}_t + \nabla_U \mathscr{G}_t,\ U_t - U^t_{\pi^*} \rangle\rangle}_{W_1} - 2\eta \underbrace{\langle\langle \nabla_V \mathscr{L}_t + \nabla_V \mathscr{G}_t,\ V_t - V^t_{\pi^*} \rangle\rangle}_{W_2} \\
&\qquad + \eta^2 \underbrace{\| \nabla_U \mathscr{L}_t + \nabla_U \mathscr{G}_t \|_F^2}_{W_3} + \eta^2 \underbrace{\| \nabla_V \mathscr{L}_t + \nabla_V \mathscr{G}_t \|_F^2}_{W_4},
\end{aligned}
$$

$W_1 + W_2$ big: descent condition.
$W_3 + W_4$ small: smoothness condition.

analysis framework – initialization

- all optimal solutions close to initialization.
- initialization implies the above descent conditions hold.

key technical challenges

establishing local descent and smoothness conditions.
generically it's not clear how much of this agenda goes through.
here: need to use incoherence, and, critically, the sparsity structure.

for partial observations: controlling size of projections onto sparse support reduces step size by factor of μr.

main results: full observation

Thm. In $\mathscr{O}(\kappa \log(1/\varepsilon))$ iterations, output satisfies

$$\|\hat{U}\hat{V}^\top - L^*\|_F \leq \varepsilon \cdot \sigma_r^*,$$

even with α-fraction of arbitrary corruptions, for

$$\alpha \leq c \min\left\{ \frac{1}{\mu\sqrt{\kappa r}^3}, \frac{1}{\mu \kappa^2 r} \right\}.$$

complexity. running time $\sim \mathscr{O}(\kappa r d^2 \log(1/\varepsilon))$.
robustness. can tolerate $\alpha \sim \mathscr{O}(1/\mu r\sqrt{r})$. culprit for extra $\sqrt{r}$: initialization requirement.

main results: partial observations

Thm. In $\mathscr{O}(\kappa\mu r \log(1/\varepsilon))$ iterations, output satisfies

$$\|\hat{U}\hat{V}^\top - L^*\|_F \leq \varepsilon \cdot \sigma_r^*,$$

even with α-fraction of arbitrary corruptions, for

$$\alpha \leq c \min\left\{\frac{1}{\mu\sqrt{\kappa r}^3}, \frac{1}{\mu\kappa^2 r}\right\},$$

and with p-fraction of observations, for

$$p \geq \frac{\kappa^4\mu^2 r^2 \log d}{d}.$$

complexity. running time $\sim \mathscr{O}(\mu^3 r^4 d \log d \log(1/\varepsilon))$. *robustness*. still can tolerate $\alpha \sim \mathscr{O}(1/\mu r\sqrt{r})$.

in summary...

- robust PCA much faster than convex or SVD-based approaches – theoretically and empirically.
- obtain close-to-optimal guarantees for robustness and for erasures.
- name of the game: initialization.

conclusion

find out more from:

`http://users.ece.utexas.edu/~cmcaram/`

or e-mail:

`constantine@utexas.edu`