Non-Separable Relaxations of a Class of Rank Penalties

Carl Olsson

June 10, 2021

Outline

Background:

- Structure from Motion (SfM) and Factorization

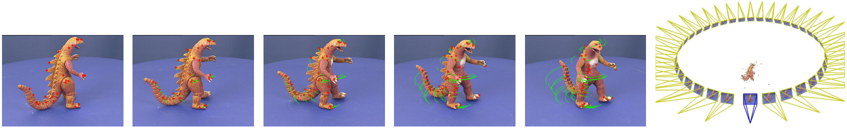
Relaxations of Non-Separable Rank/Sparsity Penalties:

- Framework
- Relaxations
- Shrinking bias, non-separable regularization
- Theoretical results under RIP

Bilinear Parameterization of Rank Penalties:

- Approach
- Theoretical results
- Algorithm Overview
- The pOSE formulation
- SfM results

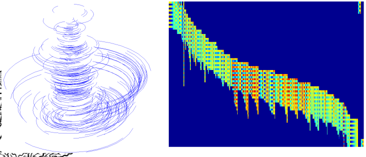
Structure from Motion and Factorization



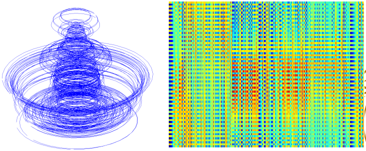
Affine camera model:

$$M = \underbrace{\begin{bmatrix} P_1 \\ P_2 \\ \vdots \end{bmatrix}}_{\text{camera matrices}} \underbrace{\begin{bmatrix} X_1 & X_2 & \dots \end{bmatrix}}_{\text{3D points}}$$

$W \odot M$

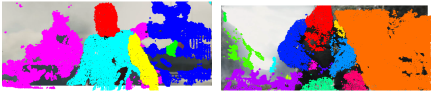


M



Non-Rigid SfM

Use higher rank for non-rigid scenes.



Hard problem, low rank, structured missing data. Primarily interested in recovering the factors.

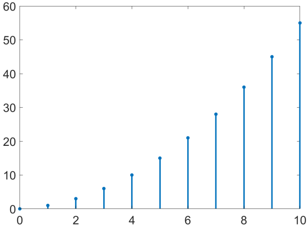
Framework

Sparsity problem:

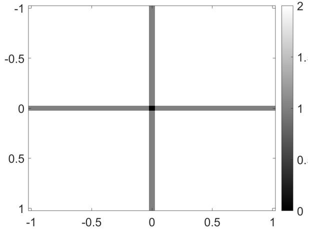
$$G(\text{card}(\boldsymbol{x})) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2, \quad \text{where } G(k) = \sum_{i=1}^{k} g_i,$$

with $0 \leq g_1 \leq g_2 \leq \ldots \leq g_n \leq \infty$. ($g_i = \infty$ is allowed for $i > 0$).

$G(k), k = 0, \ldots, 10$



$G(card(\boldsymbol{x})), \boldsymbol{x} \in \mathbb{R}^2$



Framework

Low rank problem:

$$G(\text{rank}(X)) + \|\mathcal{A}X - \boldsymbol{b}\|^2, \quad \text{where } G(k) = \sum_{i=1}^{k} g_i,$$

with $0 \leq g_1 \leq g_2 \leq \ldots \leq g_n \leq \infty$. ($g_i = \infty$ is allowed for $i > 0$).
Examples:

1. Soft rank penalty $g_i = \mu$.

$$\mu \text{rank}(X) + \|\mathcal{A}X - b\|^2.$$

2. The fixed rank problem $g_i = \begin{cases} 0 & i \leq k \\ \infty & i > k \end{cases}.$

$$\min_{\text{rank}(X) \leq k} \|\mathcal{A}X - b\|^2.$$

Bias

Some general regularizer:

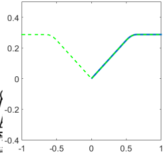
$$r(|x|) + (x - b)^2$$

Minimizer is either 0 or solution to

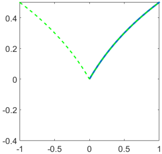
$$x = b - \frac{r'(|x|)}{2}\text{sign}(x)$$

Derivative r' needs to be zero to recover $x = b$ when b is large.

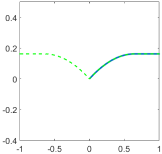
SCAD:



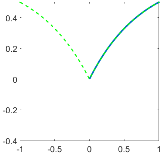
Log:



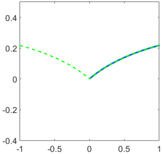
MCP:



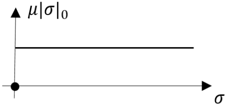
ETP:



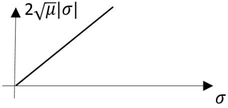
Geman:



1D versions:

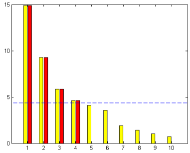


$\text{rank}(X) = \sum_i |\sigma_i(X)|_0$

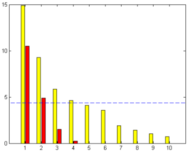


$\|X\|_* = \sum_i \sigma_i(X)$

Singular value thresholding:



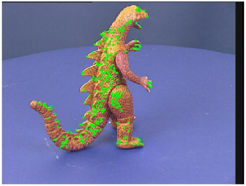
$\mu\text{rank}(X) + \|X - X_0\|_F^2$

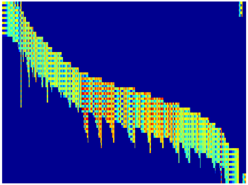


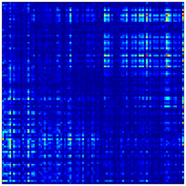
$2\sqrt{\mu}\|X\|_* + \|X - X_0\|_F^2$

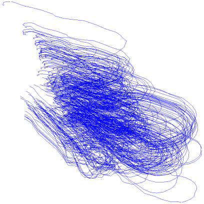
Dino Example

Data set:	Errors:	Trajectories:









Is there a bias free formulation without local minima?

Relaxation

The quadratic envelope:

- Add quadratic $f(x) := G(\text{card}(\boldsymbol{x})) + \|\boldsymbol{x}\|^2$.
- Compute convex envelope f^{**} of f.
- Subtract quadratic $r_g(\boldsymbol{x}) := f^{**}(\boldsymbol{x}) - \|\boldsymbol{x}\|^2$.

Replace $G(\text{card}(x))$ with $r_g(\boldsymbol{x})$:

$$r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2.$$

Remarks:

Vector case: $r_g(\boldsymbol{x}) = r_g(\tilde{\boldsymbol{x}})$, where $\tilde{\boldsymbol{x}}$ are sorted magnitudes or elements in $\boldsymbol{x}$.

Matrix case: $r_g(X) = r_g(\tilde{\boldsymbol{x}})$, where $\tilde{\boldsymbol{x}}$ are sorted singular values of X.

Evaluating the Relaxation

Evaluation via optimization problem:

$$r_g(\boldsymbol{x}) = \max_{\tilde{\boldsymbol{z}}} \left(\sum_{i=1}^{n} \min\left(g_i,\ \tilde{z}_i\right) - \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{x}}\|^2 \right).$$

Concave maximization. Can be solved exactly by searching linear (in the singular values) number of candidate points.

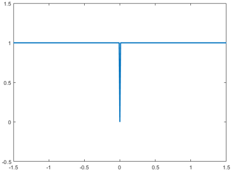
Proximal operator evaluated similarly.

1D-toy example

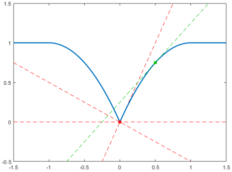
If $G_a(x) = \begin{cases} 0 & x = 0 \\ 1 & x \neq 0 \end{cases}$ then $r_g(x) = 1 - \max(1 - |x|, 0)^2$.

Solve $min_x r_g(x) + (x - b)^2$.

$G_a(x)$:



$r_g(x)$:



$G_a(x) = r_g(x)$ if $x \notin (0, 1)$

In general $G(\text{card}(\tilde{\boldsymbol{x}})) = r_g(\tilde{\boldsymbol{x}})$ if $\tilde{x}_i \notin (0, \sqrt{g_i})$, $\forall i$.

Separable vs. Non-separable

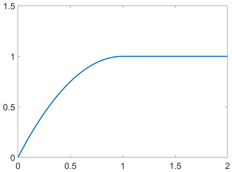
Examples of relaxations:

$$G_a(x) = \begin{cases} 0 & x = 0 \\ 1 & x \neq 0 \end{cases}$$

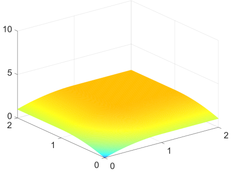
$$G_b(\mathbf{x}) = G_a(x_1) + G_a(x_2)$$

$$G_c(\mathbf{x}) = \begin{cases} 0 & x_1 = x_2 = 0 \\ 1 & x_1 = 0, x_2 \neq 0 \\ 1 & x_1 \neq 0, x_2 = 0 \\ \infty & x_1 \neq 0 \text{ and } x_2 \neq 0 \end{cases}$$

Scalar relaxation:

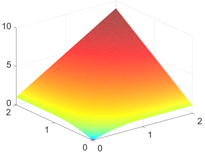


Separable relaxation:



Uninformative (high card)

Non-seprable relaxation



Strong gradient (high card)

Why this approach?

- $r_g(\boldsymbol{x})$ continuous.
- $r_g(\boldsymbol{x}) + \|\boldsymbol{x} - \boldsymbol{b}\|^2$ convex envelop of $g(\text{card}(\boldsymbol{x})) + \|\boldsymbol{x} - \boldsymbol{b}\|^2$. (Same minimizer if unique.)
- $r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ relaxation of $g(\text{card}(\boldsymbol{x})) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ have same global minizers if $\|A\| < 1$ (Carlsson, 2018).
- Any local minimum of $r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ is a local minimum of $g(\text{card}(\boldsymbol{x})) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ if $\|A\| < 1$ (Carlsson, 2018).

Analysis under RIP (Candes etal):

$$(1 - \delta_k)\|\boldsymbol{x}\|^2 \leq \|A\boldsymbol{x}\|^2 \leq (1 + \delta_k)\|\boldsymbol{x}\|^2,$$

for all $\boldsymbol{x}$ with $\text{card}(\boldsymbol{x}) \leq k$

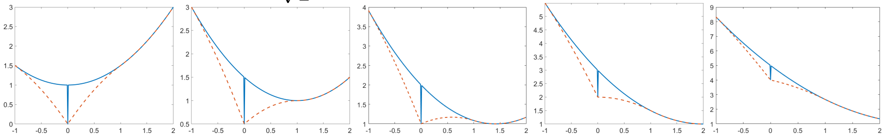
Intuition: "$\|A\boldsymbol{x}\|^2$ behaves similar to $\|\boldsymbol{x}\|^2$"

Goal

Study stationary points of $r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$.

What kind of results can we expect?
Ex. $r_g(x) + \left(\frac{1}{2}x - b\right)^2$, $g_1 = 1$:

$b = 0$ $\quad b = \frac{1}{\sqrt{2}}$ $\quad b = 1$ $\quad b = \sqrt{2}$ $\quad b = 2$



Ambiguous data will give multiple local minima.

Stationary Points

$$r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2 = \underbrace{r_g(\boldsymbol{x}) + \|\boldsymbol{x}\|^2}_{=f^{**}(\boldsymbol{x})} + \underbrace{\|A\boldsymbol{x} - \boldsymbol{b}\|^2 - \|\boldsymbol{x}\|^2}_{:=h(\boldsymbol{x})}$$

$\bar{\boldsymbol{x}}$ stationary iff $-\nabla h(\bar{\boldsymbol{x}}) \in \partial f^{**}(\bar{\boldsymbol{x}})$

$$-\nabla h(\bar{\boldsymbol{x}}) = \underbrace{2(I - A^T A)\bar{\boldsymbol{x}} + 2A^T b}_{:=2\bar{z}}$$

Easy to show that $\bar{\boldsymbol{x}}$ stationary iff

$$\bar{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}} r_g(\boldsymbol{x}) + \|\boldsymbol{x} - \bar{\boldsymbol{z}}\|^2$$

Properties of $\bar{\boldsymbol{z}}$ determines if the stationary point is unique.

Main Result

Theorem (Uniqeness of Sparse Stationary Point)

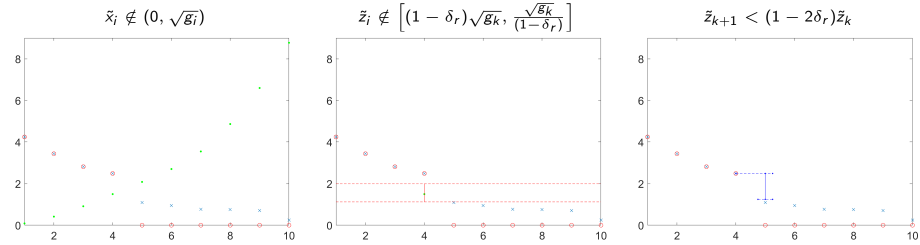
Suppose $2\boldsymbol{z} \in \partial f^{**}(\boldsymbol{x})$ *with* $\boldsymbol{z} = (I - A^T A)\boldsymbol{x} + A^T \boldsymbol{b}$, *where* A *fulfills RIP. If* $card(\boldsymbol{x}) = k$, $\tilde{x}_i \notin (0, \sqrt{g_i})$ *and* $\tilde{\boldsymbol{z}}$ *fulfills*

$$\tilde{z}_i \notin \left[(1-\delta_r)\sqrt{g_k}, \frac{\sqrt{g_k}}{(1-\delta_r)}\right] \text{ and } \tilde{z}_{k+1} < (1-2\delta_r)\tilde{z}_k, \tag{1}$$

then any other stationary point $\boldsymbol{x}'$ *has* $card(\boldsymbol{x}') > r - k$. *If in addition* $k < \frac{r}{2}$ *then* $\boldsymbol{x}$ *solves*

$$\min_{card(\boldsymbol{x}) < \frac{r}{2}} r_g(\tilde{\boldsymbol{x}}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2. \tag{2}$$

Remark: Only uses lower estimate $(1-\delta_r)\|\boldsymbol{x}\|^2 \le \|A\boldsymbol{x}\|^2$.



x - $\tilde{z}_i$

o - $\tilde{x}_i$

· - $\sqrt{g_i}$

Noisy Recovery

Theorem (Exact Recovery of Oracle Solution)

Suppose that $\mathbf{b} = A\mathbf{y} + \epsilon$, for some $\mathbf{y}$ with card$(\mathbf{y}) = k$, $\|A\| < 1$, $\delta_{2k} < \frac{1}{2}$. If

$$\tilde{y}_k > \frac{5}{(1-2\delta_{2k})\sqrt{1-\delta_{2k}}}\|\epsilon\|, \tag{3}$$

then there is a stationary point $\mathbf{x}$, with card$(\mathbf{x}) = k$, that fulfills (1) for all choices of g where

$$\sqrt{g_k} < (1-\delta_k)\left(\tilde{y}_k - \frac{2\|\epsilon\|}{\sqrt{1-\delta_{2k}}}\right) \text{ and } \sqrt{g_{k+1}} > \frac{3(1-\delta_k)}{\sqrt{1-\delta_{2k}}}\|\epsilon\|. \tag{4}$$

Remark: $\|A\| < 1$ restrictive

Hard Constraints

So far only results for sparse vectors/low rank matrices. Why?

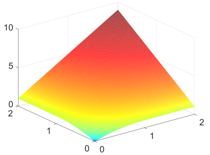
- RIP only holds for sparse vectors.
- Unbiased separable formulations are uninformative for high cardinality.

Are there high rank local minima?

Ex. $\min_{\mathbf{x}} \sum_i \left(\mu - \max(\sqrt{\mu} - \tilde{x}_i, 0)^2\right) + \|A\mathbf{x} - \mathbf{b}\|^2$

- Let $\mathbf{x}_p \in \arg\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|^2$.
- Take dense vector $\mathbf{x}_h$ in nullspace of A.
- $\mathbf{x}_p + t\mathbf{x}_h$ (t large) minimizes $\|A\mathbf{x} - \mathbf{b}\|^2$, with all elements $> \sqrt{\mu}$.

Solution add hard constraints: $g_i = \infty$ if $i \geq k_{\max}$.



Corollary (Unique Local Minimizer)

Suppose that $\mathbf{x}$ is a stationary point fulfilling the assumptions of Theorem 1 with $r = 2k$. If $\|A\| < 1$ and $g_i = \infty$ for $i \geq k$ then $\mathbf{x}$ is the unique local minimizer (and therefore the global minimizer).

Corollary (Noisy Recovery)

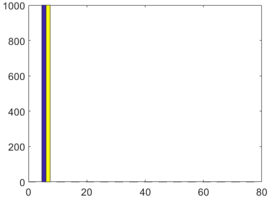
If $\|A\| < 1$ and $g_i = \infty$ for $i \geq k$ then under the assumptions of Theorem 2 the problem has a unique local minimizer.

Some Preliminary Experiments

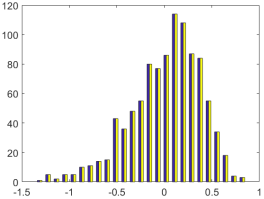
Optimization of $F(\boldsymbol{x}) = r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ with

$g_i = \mu$ for all i (blue) vs. $g_i = \begin{cases} \mu & i \leq 10 \\ \infty & i > 10 \end{cases}$ (*yellow*)

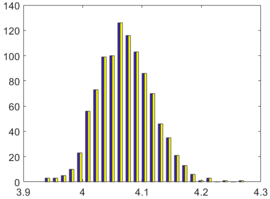
$\mathrm{card}(\boldsymbol{x})$



$\log(\|\boldsymbol{x} - \boldsymbol{x}_0\|)$



$\log(F(\boldsymbol{x}))$



A - random 60×80.

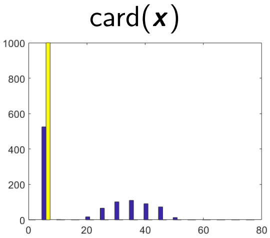
$\mathrm{card}(\boldsymbol{x}_0) = 5$, $\boldsymbol{b} = A\boldsymbol{x}_0 + \epsilon$.

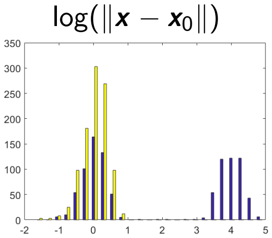
Starting point 0.

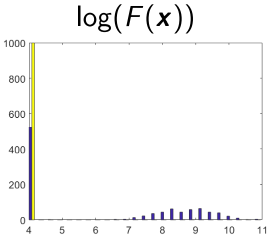
Some Preliminary Experiments

Optimization of $F(\boldsymbol{x}) = r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ with

$$g_i = \mu \text{ for all } i \text{ (blue) vs. } g_i = \begin{cases} \mu & i \leq 10 \\ \infty & i > 10 \end{cases} \text{ (yellow)}$$







A - random 60×80.

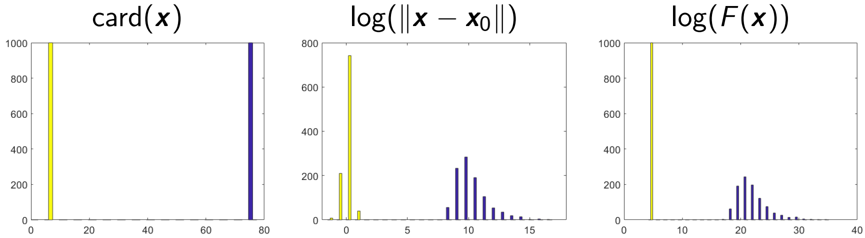
card($\boldsymbol{x}_0$) $= 5$, $\boldsymbol{b} = A\boldsymbol{x}_0 + \epsilon$.

Starting point $A \backslash \boldsymbol{b}$.

Some Preliminary Experiments

Optimization of $F(\boldsymbol{x}) = r_g(\boldsymbol{x}) + \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ with

$g_i = \mu$ for all i (blue) vs. $g_i = \begin{cases} \mu & i \leq 10 \\ \infty & i > 10 \end{cases}$ (*yellow*)



A - random 60×80.

$\text{card}(\boldsymbol{x}_0) = 5,\ b = A\boldsymbol{x}_0 + \epsilon.$

Starting point $A\backslash\boldsymbol{b} + \boldsymbol{v},\ \boldsymbol{v} \in \text{null}(A)$.

Bilinear Parameterization

Most common approach if rank is known?

$$X = BC^T, \quad B \in \mathbb{R}^{m \times r},\ C \in \mathbb{R}^{n \times r} \Rightarrow \text{rank}(X) \leq r.$$

Smooth objective in B, C:

$$\|\mathcal{A}(BC^T) - \boldsymbol{b}\|^2$$

Minimize with 2nd order methods.
(SOTA in SfM is VarPro, Hong etal. 2015, 2016, 2017, 2018.)

Can we do the same for soft penalties?

Low Rank Estimation

Slightly more general framework:

$$\min_{X} H(\boldsymbol{\sigma}(X)) + \|\mathcal{A}X - b\|^2.$$

- $H(\boldsymbol{\sigma}(X)) = \sum_{i=1}^{\text{rank}(X)} h_i \sigma_i(X) + g_i$.
- h_i, g_i, non-negative and non-decreasing.

Quadratic envelope $r_h(X)$ computed in Valtonen-Örnhag 2020.

Example:

1. Weak nuclear norm $g_i = 0$

$$\min \boldsymbol{h}^T \boldsymbol{\sigma}(X) + \|\mathcal{A}X - b\|^2.$$

Goal: Optimize with second order methods.

Approach

The variational form nuclear norm:

$$\min \|X\|_* + \|\mathcal{A}X - b\|^2 \Leftrightarrow \min \frac{\|B\|_F^2 + \|C\|_F^2}{2} + \|\mathcal{A}(BC^T) - b\|^2$$

No need to compute singular values.

General approach: If $X = BC^T = \sum_i B_i C_i^T$ replace $\sigma_i(X)$ with

$$\gamma_i(B_i, C_i) := \frac{\|B_i\|_F^2 + \|C_i\|_F^2}{2}$$

Bilinear Parameterization

Results

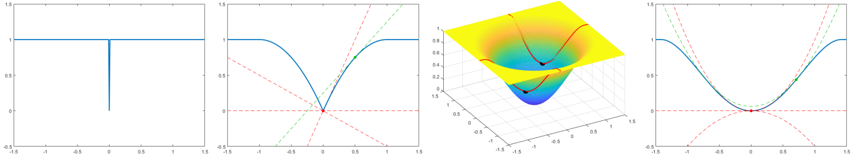
- Iglesias etal 2020. For any X we have

$$\boldsymbol{h}^T \boldsymbol{\sigma}(X) = \min_{BC^T = X} \boldsymbol{h}^T \gamma(B, C)$$

  if $h_1, h_2, \ldots$ is increasing.
- Valtonen-Örnhag etal 2021. For any X we have

$$r_h(\boldsymbol{\sigma}(X)) = \min_{BC^T = X} r_h(\gamma(B, C)).$$

Bilinear Parameterization



(a): $H(x) = \begin{cases} 0 & x = 0 \\ 1 & x \neq 0 \end{cases}$.

(b): $r_h(x)$ continuous

(c): $r_h(\frac{b^2+c^2}{2})$ differentiable (a.e two times).

(d): Slice of $r_h(\frac{b^2+c^2}{2})$ along $c = 0$

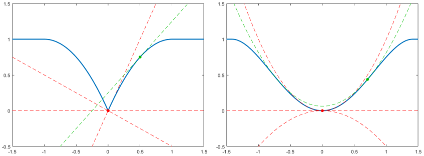
Algorithm Overview

Approximation at iteration t: $\boldsymbol{\eta} = \gamma(B^{(t)}, C^{(t)})$

$$r_h^{(t)}(\gamma(B, C)) = \sum_{i=1}^{n} w_i^{(t)} \frac{\|B_i\|^2 + \|C_i\|^2}{2}$$

$$w_i^{(t)} = 2(z_i - \eta_i)$$

where $z \in \partial f^{**}(\boldsymbol{\eta})$ with $z_i = z_{i-1}$ (z-maximal) when $\eta_i = 0$



Algorithm Overview

1. Given $(B^{(t)}, C^{(t)})$ compute the maximal subgradient $\boldsymbol{z} \in \partial f^{**}(\gamma(B^{(t)}, C^{(t)}))$.
2. Compute the approximation $r_h^{(t)}(\gamma(B, C))$.
3. Run one iteration of VarPro to obtain $(B^{(t+1)}, C^{(t+1)})$.
4. Optional: Compute the SVD $X^{(t+1)} = U\Sigma V^T$, where $X^{(t+1)} = B^{(t+1)}(C^{(t+1)})^T$, and set

$$B^{(t+1)} := U\sqrt{\Sigma}$$
$$C^{(t+1)} := V\sqrt{\Sigma}$$

Empirical observation: SVD can be omitted if $h_i \neq 0$.

Issues

- Slow iterations.
- Hard to increase rank.
- Local minima if $h_i = 0$. (Seem to be removed by SVD step.)

The pOSE Formulation. Hong & Zach 2018

Pinhole Projection:

$$\mathcal{O}_{\text{ML}} = \sum_{i,j} \| \frac{1}{z_{ij}} \boldsymbol{x}_{ij} - \boldsymbol{m}_{ij} \|^2, \qquad \boldsymbol{X}_{ij} = \begin{bmatrix} \boldsymbol{x}_{ij} \\ z_{ij} \end{bmatrix}.$$

Object Space Error:

$$\mathcal{O}_{\text{OSE}} = \sum_{i,j} \| \boldsymbol{x}_{ij} - \boldsymbol{m}_{ij} z_{ij} \|^2.$$

- Perpendicular distance from viewing ray to X_{ij}.
- Linear residuals. (Bilinear least squares in P, U.)
- Not scale invariant (trivial minimizer).

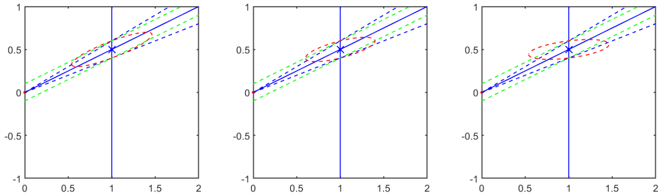
The pOSE Formulation. Hong & Zach 2018

Affine term:

$$\mathcal{O}_{\text{Affine}} = \sum_{i,j} \|\boldsymbol{x}_{ij} - \boldsymbol{m}_{ij}\|^2.$$

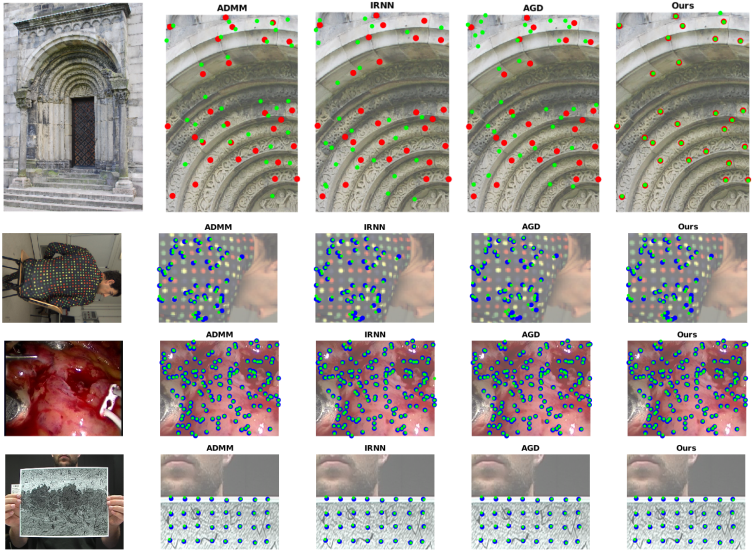
Pseudo Object Space Error:

$$\mathcal{O}_{\text{pOSE}} = (1-\eta)\mathcal{O}_{\text{OSE}} + \eta\mathcal{O}_{\text{Affine}}.$$



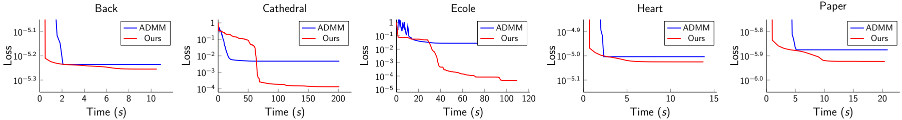
$\eta = 0.25$ $\eta = 0.5$ $\eta = 0.75$

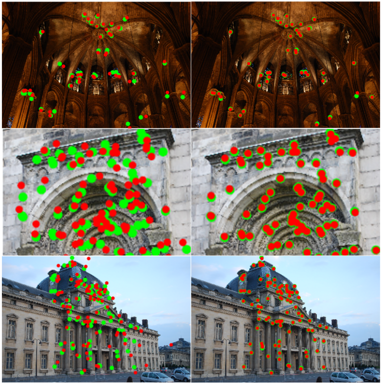
Results


ADMM
IRNN
AGD
Ours
ADMM
IRNN
AGD
Ours
ADMM
IRNN
AGD
Ours
ADMM
IRNN
AGD
Ours

Results

Comparison to ADMM on some data.





Some References

- Carlsson, *On convex envelopes and regularization of non-convex functionals without moving global minima*, Journal of Optimization Theory and Applications, 2019.
- Olsson, Gerosa, Carlsson, *Relaxations for Non-Separable Cardinality/Rank Penalties*, Preprint.
- Hong, Zach, *pOSE: Pseudo Object Space Error for Initialization-Free Bundle Adjustment*, CVPR 2018.
- Hong, Zach, Fitzgibbon, *Revisiting the Variable Projection Method for Separable Nonlinear Least Squares Problems*, CVPR 2017.
- Hong, Zach, Fitzgibbon, Chipola, *Projective Bundle Adjustment from Arbitrary Initialization Using the Variable Projection Method*, CVPR 2017.
- Valtonen-Örnhag, Olsson, *A Unified Optimization Framework for Low-Rank Inducing Penalties*, CVPR 2020.
- Iglesias, Olsson, Valtonen-Örnhag, *Accurate Optimization of Weighted Nuclear Norm for Non-Rigid Structure from Motion*, ECCV 2020
- Valtonen-Örnhag, Olsson, Igelsias, *Bilinear Parameterization for Non-Separable Singular Value Penalties*, CVPR 2021
- Larsson, Olsson, *Convex Low Rank Approximation*, IJCV 2016.

The End