This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Descending, lifting or smoothing: Secrets of robust cost optimization

Christopher Zach[1] and Guillaume Bourmaud[2]

[1] Toshiba Research Europe, Cambridge, United Kingdom
[2] University of Bordeaux, Bordeaux, France

Abstract. Robust cost optimization is the challenging task of fitting a large number of parameters to data points containing a significant and unknown fraction of outliers. In this work we identify three classes of deterministic second-order algorithms that are able to tackle this type of optimization problem: direct approaches that aim to optimize the robust cost directly with a second order method, lifting-based approaches that add so called lifting variables to embed the given robust cost function into a higher dimensional space, and graduated optimization methods that solve a sequence of smoothed cost functions. We study each of these classes of algorithms and propose improvements either to reduce their computational time or to make them find better local minima. Finally, we experimentally demonstrate the superiority of our improved graduated optimization method over the state of the art algorithms both on synthetic and real data for four different problems.

1 Introduction

Robust cost optimization aims to fit parameters to data containing outliers. This generic optimization problem arises in a large number of applications in computer vision such as bundle adjustment [23], optical flow [4], SLAM [8], registration of 3D surfaces [30], etc. In applications where the data contains a small number of outliers, using a convex kernel[3], such as the L_1 kernel or the Huber kernel, sufficiently reduces influence of outliers to obtain a good fit of the parameters to inlier data points. However, when the observations contain a large number of potentially gross outliers, a convex kernel is not "robust" enough and a quasi-convex kernel, such as Tukey's biweight kernel, has to be employed. Optimizing over a sum of quasi-convex kernels produces a highly non-convex cost function with many local minima. In low-dimensional parameter problems poor local minima can be escaped using stochastic/sampling approaches, such as RANSAC [10] or simulated annealing [16]. Nevertheless, such methods are impractical for applications that have a large number of parameters and data points (such as bundle adjustment or optical flow). For these large scale problems deterministic second-order approaches are generally considered to be a good compromise between efficiency and accuracy. In return, special care must be taken to escape poor local minima.

[3] In this paper, the word "kernel" refers to a loss function.

Contributions: In this paper, we start with identifying three classes of such algorithms: *direct approaches*, *lifting-based approaches* and *graduated optimization methods*. Then, we study each of these classes of algorithms and propose improvements either to reduce their computational time or to make them find better local minima. More precisely, we make the following contributions: (i) We show that the *direct approaches* only differ in their quadratic approximation of the quasi-convex kernel. This analysis allows us to outline the limitations and numerical instabilities of some of these algorithms. (ii) We propose to use a convexified Newton approximation to implement *lifting-based approaches* and experimentally demonstrate that this modification leads to better local minima than the classical Gauss-Newton approximation. (iii) We design a novel stopping criterion that allows to significantly speed-up *graduated optimization methods* without harming their ability to avoir poor local minima. (iv) We experimentally demonstrate the superiority of our improved graduated optimization method over the state of the art algorithms both on synthetic and real data for three different problems.

Organization of the paper: The rest of the paper is organized as follows: Section 2 discusses the related work and Section 3 introduces our notations as well as some fundamental definitions. Our contributions are gathered in Sections 4, 5 and 6, where we study three different types of algorithms and make several recommendations to improve their performances. In Section 7 numerical evaluations of the methods discussed in the previous three sections are presented. A summary of our recommendations and future work are provided in Section 8.

2 Related work

In this section, we describe the state of the art approaches for robust cost optimization (e.g. redescending m-estimation [15]) and how they are related to the novel method we propose in this paper. We focus on deterministic second-order methods because they are generally considered to be a good compromise between efficiency and accuracy for problems with large numbers of parameters and data points (such as bundle adjustment). In the following literature review we distinguish *direct approaches*, *graduated optimization methods* and *lifting-based approaches*[4].

Direct approaches aim to optimize the original robust objective, usually by utilizing a surrogate model suitable for a second-order method. IRLS [13], the Triggs correction [23] and "square rooting the kernel" [9] are well-known instances of this class of methods. Consequently, these approaches find the local minimum corresponding to the basin of convergence they were initialized in.

Graduated optimization is a meta-algorithm explicitly designed to avoid poor local minima by building a sequence of successively smoother (and therefore easier to optimize) approximations of the original objective. The optimization algo-

[4] The "Variable Projection" (VarPro) approach, which can be interpreted as the "opposite" of lifting, was recently shown to be successful for matrix factorization [14].

rithm consists in successively optimizing the sequence of cost functions (e.g. by using one of the *direct approaches*), with the solution from the previous objective used as starting point for the next one. Homotopy optimization methods (e.g. [7]) and continuation methods (e.g. [20]) are other terms for the same meta-algorithm. Graduated non-convexity [5], multi-scale methods and Gaussian homotopies [19], and deterministic annealing (e.g. [21]) are specific constructions that belong to this family of methods. One drawback of graduated optimization is that they appear to be inefficient as an entire sequence of optimization problems has to be solved.

Instead of explicitly building a sequence of smoothed cost functions, lifting approaches[5] [31, 29, 26] add so called lifting variables (which can be interpreted as confidence weights) to embed the original robust cost function into a higher dimensional space of unknowns. Initializing the lifting variables to a large value corresponds to smoothing the robust cost function while setting them to their optimal values produces the original robust cost function. The algorithm consists in jointly optimizing over the parameters of interest and the lifting variables. Lifting-based methods can be interpreted as "self-tuned" graduated optimization. One drawback of these methods is, that their performance significantly depends on the initialization of the lifting variables (as demonstrated in our numerical experiments).

3 Background and notations

Robust cost optimization consists in minimizing functions of the form:

$$\min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}) \qquad \text{with} \qquad \Psi(\boldsymbol{\theta}) = \sum_{i=1}^{N} \psi(\|\mathbf{f}_i(\boldsymbol{\theta})\|), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ are the parameters of interest, $\mathbf{f}_i(\boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}^d$ is the i-th vectorial residual function and $\psi(\cdot)$ is a robust kernel function (that will be formally defined hereinafter), that allows to reduce the influence of outlying data points. $\|\cdot\|$ is the usual L_2-norm (leading to isotropic penalization of large residuals). The arguably simplest application of Eq.(1) arises when robustly fitting a "mean" vector $\boldsymbol{\theta}$ to data points $\mathbf{y}_i \in \mathbb{R}^d$ which leads to the following problem: $\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \psi(\|\mathbf{y}_i - \boldsymbol{\theta}\|)$. From a practical point of view, we would like a robust kernel function to convey the idea that large residuals should always have a smaller influence than smaller residuals when estimating the optimal parameters $\boldsymbol{\theta}^*$. We will now translate this idea into formal properties of a robust kernel function: A robust kernel function $\psi : \mathbb{R} \to \mathbb{R}_0^+$ is a symmetric function sufficiently smooth near 0 with the following properties: 1) $\psi(0) = 0$ and $\psi''(0) = 1$. 2) The mapping $\phi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with $\phi(z) := \psi(\sqrt{2z})$ (or $\phi(r^2/2) = \psi(r)$) is concave and monotonically increasing.

[5] Here, the term "lifting" refers to the "multiplicative" version of lifting [11, 29]. We do not consider other types of lifting such as the "additive" version of lifting [12, 28].

In robust cost functions such as Eq. 1 the robust kernel ψ is applied only to non-negative arguments, but it is customary to extend its domain to the entire real line $\mathbb{R}$. The "normalization" property (property 1) allows to compare the robustness of different kernels. Concerning property 2, the fact that ϕ should be monotonically increasing is obvious but the necessity of its concavity requires some justification. To so do, we examine the *weight function* ω associated with a robust kernel ψ that describes how ψ weighs the influence of residuals[6]:

$$\omega(r) := \psi'(r)/r = \phi'(r^2/2). \tag{2}$$

Since we aim for large residuals having a smaller influence than smaller residuals, $\omega(\cdot)$ should be monotonically decreasing in $|r|$, which is guaranteed by the concavity of ϕ. Let us note that this definition of a robust kernel includes both convex and quasi-convex kernels. However, as stated in the introduction, in the experiments we will only consider quasi-convex kernels.

4 Direct methods: IRLS, Triggs correction, $\sqrt{\psi}$

In this section, we review the approaches that aim to (iteratively) minimize the objective $\Psi(\boldsymbol{\theta})$ (see Eq. 1) without explicitly modifying the objective, and we outline that each of these approaches can be interpreted as methods trying to locally approximate ψ with a quadratic function. In order to be computationally efficient, these methods try to cast the original problem in a way that allows non-linear least-squares solvers, such as Gauss-Newton or Levenberg-Marquardt, to be employed. As a consequence, at each iteration these approaches perform the following steps:
1. perform a first order approximation of the vectorial residual function around the current value of the parameters $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$: $\mathbf{f}_i(\bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}) \approx \bar{\mathbf{f}}_i + \mathtt{J}_i \Delta\boldsymbol{\theta}$ where $\mathtt{J}_i$ is the Jacobian of $\mathbf{f}_i(\bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta})$ w.r.t. the increment $\Delta\boldsymbol{\theta}$ evaluated at $\Delta\boldsymbol{\theta} = \mathbf{0}$ and $\bar{\mathbf{f}}_i$ is a short hand notation for $\mathbf{f}_i(\bar{\boldsymbol{\theta}})$,
2. approximate $\psi(\|\bar{\mathbf{f}}_i + \mathtt{J}_i \Delta\boldsymbol{\theta})\|)$ with a quadratic function $\breve{\psi}$ s.t. $\psi(\|\bar{\mathbf{f}}_i\|) = \breve{\psi}(\|\bar{\mathbf{f}}_i\|)$.
While step 1 is the same for all the approaches, step 2 turns out to be very different for each of them.

IRLS [13]: One way to derive the IRLS methods is to interpret it as instance of the majorize-minimize principle (e.g. [17]): given the current solution $\bar{\boldsymbol{\theta}}$, IRLS uses a quadratic majorizer of ψ, i.e $\psi(r) \leq \breve{\psi}_{\text{IRLS}}(r)$:

$$\breve{\psi}_{\text{IRLS}}(\|\bar{\mathbf{f}}_i + \mathtt{J}_i \Delta\boldsymbol{\theta})\|) = \omega(\|\bar{\mathbf{f}}_i\|)\left(\|\bar{\mathbf{f}}_i + \mathtt{J}_i \Delta\boldsymbol{\theta})\|^2/2 - \|\bar{\mathbf{f}}_i\|^2/2\right) + \psi(\|\mathbf{f}_i(\bar{\boldsymbol{\theta}})\|). \tag{3}$$

Since robust kernels are by construction sub-quadratic, a non-degenerate quadratic majorizer always exists. The IRLS algorithm iteratively builds and minimizes the quadratic surrogates, which yields a sequence of solutions $\boldsymbol{\theta}^{(k)}$ with monotonically decreasing costs $\Psi(\boldsymbol{\theta}^{(k)})$.

[6] For instance, for the quadratic kernel (which does not try to reduce the influence of large residuals), we have $\omega(r) = 1$.

Triggs correction [23]: Contrary to IRLS, the Triggs correction performs a second order expansion of $F_i(\boldsymbol{\mathit{\Delta\theta}}) := \phi(\|\mathbf{f}_i + \mathtt{J}_i\boldsymbol{\mathit{\Delta\theta}}\|^2/2)$ around $\boldsymbol{\mathit{\Delta\theta}} = \mathbf{0}$. The resulting approximation of ψ is given by

$$\breve{\psi}_{\text{Triggs}}(\|\bar{\mathbf{f}}_i + \mathtt{J}_i\boldsymbol{\mathit{\Delta\theta}})\|) = \psi(\|\bar{\mathbf{f}}_i\|) + \nabla_{\boldsymbol{\mathit{\Delta\theta}}}F_i(\mathbf{0})^\top\boldsymbol{\mathit{\Delta\theta}} + \boldsymbol{\mathit{\Delta\theta}}^\top\mathtt{H}_{F_i}(\mathbf{0})\boldsymbol{\mathit{\Delta\theta}} \tag{4}$$

with the following expressions for the gradient and Hessian at $\boldsymbol{\mathit{\Delta\theta}} = \mathbf{0}$:

$$\nabla_{\boldsymbol{\mathit{\Delta\theta}}}F_i(\mathbf{0}) = \omega(\|\mathbf{f}_i\|)\mathtt{J}_i^\top\mathbf{f}_i \qquad \mathtt{H}_{F_i}(\mathbf{0}) = \mathtt{J}_i^\top\left(\frac{\omega'(\|\mathbf{f}_i\|)}{\|\mathbf{f}_i\|}\mathbf{f}_i\mathbf{f}_i^\top + \omega(\|\mathbf{f}_i\|)\mathtt{I}\right)\mathtt{J}_i.$$

where we used Eq. 2 as well as $\phi''(z) = (\omega(\sqrt{2z}))' = \frac{\omega'(\sqrt{2z})}{\sqrt{2z}}$. Note that $\mathbf{f}_i$ is an eigenvector for $\frac{\omega'}{\|\mathbf{f}_i\|}\mathbf{f}_i\mathbf{f}_i^\top + \omega\mathtt{I}$ (omitting arguments to ω and ω'):

$$\left(\frac{\omega'}{\|\mathbf{f}_i\|}\mathbf{f}_i\mathbf{f}_i^\top + \omega\mathtt{I}\right)\mathbf{f}_i = \omega'\|\mathbf{f}_i\|\mathbf{f}_i + \omega\mathbf{f}_i = (\omega'\|\mathbf{f}_i\| + \omega)\mathbf{f}_i.$$

Hence, if $\omega + \|\mathbf{f}_i\|\omega' < 0$, then $\mathtt{H}_{F_i}(\mathbf{0})$ is negative-definite and the Triggs correction approach cannot be applied. The popular Ceres solver [1], which supports the Triggs correction for robust cost optimization, reverts to IRLS for the current step in this case.

Square-rooting ψ [9]: A third option consists in square-rooting ψ and performing a first order Taylor expansion of it around $\boldsymbol{\mathit{\Delta\theta}} = \mathbf{0}$. Defining $G_i(\boldsymbol{\mathit{\Delta\theta}}) := g(\bar{\mathbf{f}}_i + \mathtt{J}_i\boldsymbol{\mathit{\Delta\theta}})$ where $g(\mathbf{v}) = \sqrt{\psi(\|\mathbf{v}\|)}\cdot\mathbf{v}/\|\mathbf{v}\|$, we obtain:

$$\breve{\psi}_{\sqrt{}}(\|\bar{\mathbf{f}}_i + \mathtt{J}_i\boldsymbol{\mathit{\Delta\theta}})\|) = \left(g(\bar{\mathbf{f}}_i) + J_{G_i}(\mathbf{0})\boldsymbol{\mathit{\Delta\theta}}\right)^2 \tag{5}$$

with

$$\mathtt{J}_{G_i}(\mathbf{0}) = \mathtt{J}_{g_i}(\bar{\mathbf{f}}_i)\mathtt{J}_i \quad \text{and} \quad \mathtt{J}_{g_i}(\mathbf{v}) = \frac{\sqrt{\psi(\|\mathbf{v}\|)}\|\mathbf{v}\|^2 I - \frac{\gamma(\omega(\mathbf{v}))}{\sqrt{\psi(\|\mathbf{v}\|)}}\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|^3}.$$

where we defined $\gamma(\omega(\mathbf{v})) = \psi(\mathbf{v}) - \omega(\mathbf{v})\frac{\|\mathbf{v}\|^2}{2}$ (more details about that function γ will be provided in Section 5). Despite $\|\mathbf{v}\|$ appearing in the denominator, $g(\mathbf{v})$ is smoothly behaving near $\mathbf{0}$. The $\mathbf{v}/\|\mathbf{v}\|$ term cancels out the non-differentiability induced by the square root. Observe that $\sqrt{\psi(\|\mathbf{v}\|)}$ behaves like $\|\mathbf{v}\|/\sqrt{2}$ near $\mathbf{v} = \mathbf{0}$, hence $g(\mathbf{v}) \approx \mathbf{v}/\sqrt{2}$ for $\mathbf{v} \approx \mathbf{0}$. Consequently, $\lim_{\mathbf{v}\to\mathbf{0}}\mathtt{J}_{g_i}(\mathbf{v}) = \mathtt{I}/\sqrt{2}$.

In Fig.1, we plot the different approximations of $\psi(\|\bar{\mathbf{f}}_i + J_i(\bar{\boldsymbol{\theta}})\boldsymbol{\mathit{\Delta\theta}})\|)$ for the 1-D linear function $f_i(\theta) = \theta$ and $\bar{\theta} = 1$. One can see that both "Square-rooting ψ" and the Triggs correction do not preserve the symmetry of ψ, whereas IRLS does. Moreover, the IRLS approximation is the only function that has its minimum at 0 whereas "Square-rooting ψ" has a tendency to overshoot with a minimum at ≈ -0.7 and the Triggs correction produces a negative second-order derivative. **Contribution:** Our analysis shows that the underlying quadratic models are very different and that the IRLS model has desirable properties which supports what is pointed out in [29]: the Triggs correction performs poorly and "Square-rooting ψ" is often inferior to IRLS. Nevertheless, the *direct approaches* have a major drawback: their limited ability to escape poor local minima. This leads us to studying fundamentally different approaches in the following sections.

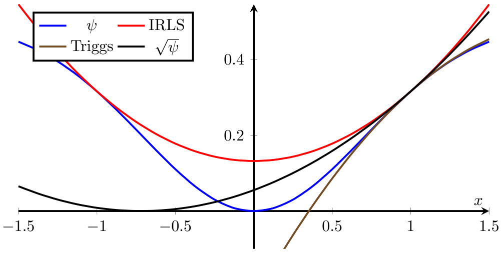

Fig. 1. Quadratic surrogate models used by direct approaches at $\bar{\theta} = 1$ (the x-axis corresponds to $\theta = \bar{\theta} + \Delta\theta$): IRLS (Eq. 3), second order expansion used in the Triggs correction (Eq. 4) (which is concave at $\bar{\theta}$) and "Square-rooting ψ" (Eq. 5). Observe that only the IRLS model preserves the symmetry of ψ and has its minimum at the zero residual.

5 Half-quadratic lifting-based methods

In this section we review the lifting approach for robust cost minimization proposed in [29], and we unify the formulation with a convexified Newton approximation[7]. In analogy with half-quadratic lifting [11] the robust kernel ψ is reformulated as point-wise minimum over a family of convex parabolas,

$$\psi(x) = \min_{v \in [0,1]} v\frac{x^2}{2} + \gamma(v), \tag{6}$$

where $\gamma : [0,1] \to \mathbb{R}_0^+$ is a convex and monotonically decreasing "bias" function in $[0,1]$. For many interesting choices of ψ the bias function γ can be continuously extended to the domain $\mathbb{R}_0^+$ (see e.g. [26]). γ is convex but generally increasing in $\mathbb{R}_{\geq 1}$. In order to avoid the constraint $v \in [0,1]$ (or $v \geq 0$, respectively) we reparametrize $v = w(u)$, where $w : \mathbb{R} \to [0,1]$ or $w : \mathbb{R} \to \mathbb{R}_0^+$. Three sensible choices for w are $w(u) = u^2$, $w(u) = e^u$ and $w = \text{sigmoid}(u)$, where sigmoid is the sigmoid function, e.g. $\text{sigmoid}(u) = 1/(1+e^{-u})$. Note that in the objective Eq. 1 one has to introduce an auxiliary unknowns u_i for each term in the sum, but this only induces a moderate increase in run-time in a second order minimization method (e.g. by leveraging the Schur complement [29]).

Using Eq. 6 we can reformulate Eq. 1 as

$$\Psi(\boldsymbol{\theta}) = \min_{u_1,\dots,u_N} \sum_i \left(w(u_i)\frac{\|\mathbf{f}_i(\boldsymbol{\theta})\|^2}{2} + \gamma(w(u_i)) \right) =: \min_{u_1,\dots,u_N} \tilde{\Psi}(\boldsymbol{\theta}, (u_i)_i) \tag{7}$$

For notational brevity we will write w_i for $w(u_i)$, w_i' for $w'(u_i)$ etc. in the following.

[7] This is different to a so-called "Lifted Newton method" [3], which addresses "deeply nested functions" and thus is not directly applicable to robust optimization.

Gauss-Newton: After linearizing the residual $\mathbf{f}_i(\bar{\boldsymbol{\theta}} + \varDelta\boldsymbol{\theta}) \approx \bar{\mathbf{f}}_i + \mathtt{J}_i\varDelta\boldsymbol{\theta}$ we can rewrite each term of $\tilde{\varPsi}$ as

$$F_i(\varDelta\boldsymbol{\theta}, \varDelta u_i) := \tfrac{w_i}{2}\left\|\bar{\mathbf{f}}_i + \mathtt{J}_i\varDelta\boldsymbol{\theta}\right\|^2 + \gamma(w_i) = \left\|\begin{matrix}\frac{\sqrt{w_i}}{\sqrt{2}}(\mathtt{J}_i\varDelta\boldsymbol{\theta} + \bar{\mathbf{f}}_i) \\ \sqrt{\gamma(w_i)}\end{matrix}\right\|^2. \tag{8}$$

After taking first order derivatives we obtain the Gauss-Newton model for $\tilde{\varPsi}$,

$$\begin{aligned}\tilde{\varPsi}^{GN}(\varDelta\boldsymbol{\theta}, (\varDelta u_i)_i) = \frac{1}{2}\sum_i \begin{pmatrix}\varDelta\boldsymbol{\theta} \\ \varDelta u_i\end{pmatrix}^\top \begin{pmatrix} w_i\mathtt{J}_i^\top\mathtt{J}_i & \frac{w_i'}{2}\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ \frac{w_i'}{2}\bar{\mathbf{f}}_i^\top\mathtt{J}_i & \frac{(w_i')^2}{4w_i}\|\bar{\mathbf{f}}_i\|^2 + \frac{(w_i'\gamma_i')^2}{2\gamma_i}\end{pmatrix}\begin{pmatrix}\varDelta\boldsymbol{\theta} \\ \varDelta u_i\end{pmatrix} \\ + \sum_i \begin{pmatrix} w_i\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ \frac{w_i'}{2}\|\bar{\mathbf{f}}_i\|^2 + w_i'\gamma_i'\end{pmatrix}^\top \begin{pmatrix}\varDelta\boldsymbol{\theta} \\ \varDelta u_i\end{pmatrix} + const.\end{aligned} \tag{9}$$

By construction the matrices

$$\begin{pmatrix} w_i\mathtt{J}_i^\top\mathtt{J}_i & \frac{w_i'}{2}\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ \frac{w_i'}{2}\bar{\mathbf{f}}_i^\top\mathtt{J}_i & \frac{(w_i')^2}{4w_i}\|\bar{\mathbf{f}}_i\|^2 + \frac{(w_i'\gamma_i')^2}{2\gamma_i}\end{pmatrix} \tag{10}$$

are positive semi-definite. The bottom right element has two problematic points: when $w_i \to 0$ (then $(w_i')^2/w_i$ is indeterminate) and when $w_i \to 1$ (in this case $(\gamma_i')^2/\gamma_i$ is indeterminate as $\gamma(1) = 0$). It can be shown [27] that the first order Taylor expansions of $(w')^2/w$ and $(\gamma'(v))^2/\gamma(v)$ at the problematic points are given by

$$\frac{(w'(\varDelta u))^2}{w(\varDelta u)} \approx 2w''(0) + \tfrac{4}{3}w'''(0)\varDelta u \qquad \frac{(\gamma'(1+\varDelta v))^2}{\gamma(1+\varDelta v)} \approx 2\gamma''(1) + \tfrac{4}{3}\gamma'''(1)\varDelta v$$

for $\varDelta u$ and $\varDelta v$ small. Consequently, a Gauss-Newton based method can be implemented generically by providing γ and w and the corresponding derivatives.

Newton: The Newton approximation of F_i (Eq. 8) around $\bar{\boldsymbol{\theta}}$ and u_i is given by

$$\begin{aligned}F_i^N(\varDelta\boldsymbol{\theta}, \varDelta u_i) \approx \frac{1}{2}\begin{pmatrix}\varDelta\boldsymbol{\theta} \\ \varDelta u_i\end{pmatrix}^\top \begin{pmatrix} w_i\mathtt{J}_i^\top\mathtt{J}_i & w_i'\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ w_i'\bar{\mathbf{f}}_i^\top\mathtt{J}_i & \frac{w_i''}{2}\|\bar{\mathbf{f}}_i\|^2 + w_i''\gamma_i' + (w_i')^2\gamma_i''\end{pmatrix}\begin{pmatrix}\varDelta\boldsymbol{\theta} \\ \varDelta u_i\end{pmatrix} \\ + \begin{pmatrix} w_i\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ \frac{w_i'}{2}\|\bar{\mathbf{f}}_i\|^2 + w_i'\gamma_i'\end{pmatrix}\begin{pmatrix}\varDelta\boldsymbol{\theta} \\ \varDelta u_i\end{pmatrix} + const.\end{aligned} \tag{11}$$

In this case the Hessian matrices

$$A_i^N := \begin{pmatrix} w_i\mathtt{J}_i^\top\mathtt{J}_i & w_i'\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ w_i'\bar{\mathbf{f}}_i^\top\mathtt{J}_i & \frac{w_i''}{2}\|\bar{\mathbf{f}}_i\|^2 + w_i''\gamma_i' + (w_i')^2\gamma_i''\end{pmatrix} =: \begin{pmatrix} w_i\mathtt{J}_i^\top\mathtt{J}_i & w_i'\mathtt{J}_i^\top\bar{\mathbf{f}}_i \\ w_i'\bar{\mathbf{f}}_i^\top\mathtt{J}_i & \alpha_i\end{pmatrix} \tag{12}$$

are not guaranteed to be p.s.d. We also denote the bottom right element of A_i^N by $\alpha_i := \frac{w_i''}{2}\|\bar{\mathbf{f}}_i\|^2 + w_i''\gamma_i' + (w_i')^2\gamma_i''$. Assuming that $w_i\mathtt{J}_i^\top\mathtt{J}_i$ is strictly positive definite (not just p.s.d. guaranteed by construction)[8], we obtain via the Schur complement that A_i^N is p.s.d. iff $\alpha_i - \frac{(w_i)'^2}{w_i} \bar{\mathbf{f}}_i^\top \mathtt{J}_i(\mathtt{J}_i^\top \mathtt{J}_i)^{-1} \mathtt{J}_i^\top \bar{\mathbf{f}}_i \geq 0$. In order to enforce that A_i^N is p.s.d., we add a non-negative value δ_i to α_i

$$\alpha_i + \delta_i - \frac{(w_i)'^2}{w_i} \bar{\mathbf{f}}_i^\top \mathtt{J}_i(\mathtt{J}_i^\top \mathtt{J}_i)^{-1} \mathtt{J}_i^\top \bar{\mathbf{f}}_i \geq 0. \tag{13}$$

Since $\mathtt{J}_i(\mathtt{J}_i^\top \mathtt{J}_i)^{-1} \mathtt{J}_i^\top$ is a projection matrix into a respective subspace (the column space of $\mathtt{J}_i$), we deduce that $\bar{\mathbf{f}}_i^\top \mathtt{J}_i(\mathtt{J}_i^\top \mathtt{J}_i)^{-1} \mathtt{J}_i^\top \bar{\mathbf{f}}_i \leq \|\bar{\mathbf{f}}_i\|^2$. Hence, setting $\delta_i = \max\{0, \frac{(w_i)'^2}{w_i} \|\bar{\mathbf{f}}_i\|^2 - \alpha_i\}$ is a sufficient condition for Eq. 13 to be satisfied. Note that $\alpha_i + \delta_i = \max\left\{\alpha_i, \frac{(w_i)'^2}{w_i} \|\bar{\mathbf{f}}_i\|^2\right\}$ and therefore the convexified matrix $\breve{A}_i^N$ is given by

$$\breve{A}_i^N := \begin{pmatrix} w_i \mathtt{J}_i^\top \mathtt{J}_i & w_i' \mathtt{J}_i^\top \bar{\mathbf{f}}_i \\ w_i' \bar{\mathbf{f}}_i^\top \mathtt{J}_i & \max\left\{\alpha_i, \frac{(w_i)'^2}{w_i} \|\bar{\mathbf{f}}_i\|^2\right\} \end{pmatrix}. \tag{14}$$

Thus, the (convexified) Newton model for $\tilde{\Psi}$ finally reads as

$$\begin{aligned} \tilde{\Psi}^N(\Delta\boldsymbol{\theta}, (\Delta u_i)_i) = \frac{1}{2} \sum_i \begin{pmatrix} \Delta\boldsymbol{\theta} \\ \Delta u_i \end{pmatrix}^\top \begin{pmatrix} w_i \mathtt{J}_i^\top \mathtt{J}_i & w_i' \mathtt{J}_i^\top \bar{\mathbf{f}}_i \\ w_i' \bar{\mathbf{f}}_i^\top \mathtt{J}_i & \max\left\{\alpha_i, \frac{(w_i)'^2}{w_i} \|\bar{\mathbf{f}}_i\|^2\right\} \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{\theta} \\ \Delta u_i \end{pmatrix} \\ + \sum_i \begin{pmatrix} w_i \mathtt{J}_i^\top \bar{\mathbf{f}}_i \\ \frac{w_i'}{2} \|\bar{\mathbf{f}}_i\|^2 + w_i' \gamma_i' \end{pmatrix}^\top \begin{pmatrix} \Delta\boldsymbol{\theta} \\ \Delta u_i \end{pmatrix} + const. \end{aligned} \tag{15}$$

Contribution: Our novel Newton-based approach (Eq. 15) suggests different updates for $\Delta\boldsymbol{\theta}$ and $(\Delta u_i)_{i=1,\dots,N}$ than the Gauss-Newton approach (Eq. 9). This is due to the fact that our Newton-based solver leverages second order information. Thus one may expect it to reach better local minima than the Gauss-Newton based solver.

6 Graduated optimization

Graduated optimization aims to avoid poor local minima usually returned by local optimization methods (such as the direct methods presented in Section 4) by iteratively optimizing successively better approximations of the original objective. It therefore relies on a sequence of objectives $(\Psi^0, \dots, \Psi^{k_{\max}})$ such that $\Psi^0 = \Psi$ and Ψ^{k+1} is in some sense easier to optimize than Ψ^k. To our knowledge graduated optimization has not been explored much in the geometric computer vision literature (besides graduated non-convexity, which was specifically developed for a robust and edge-preserving image smoothing method), although it is frequently used in image matching (by leveraging a scale space or image pyramid e.g. [24, 18]). Algorithm 1 illustrates the basic graduated optimization method. The construction of Ψ^k and the choices for a stopping criterion are left unspecified and will be described in the following.

[8] which will be guaranteed in the implementation as we use a damped Newton approach.

Algorithm 1 A generic graduated optimization method.

```
$\hat{\boldsymbol\theta}[k_{\max}] \leftarrow \boldsymbol\theta^0$
for all $k = k_{\max}, \dots, 0$ do                       ▷ Traverse towards original cost
    if $k < k_{\max}$ then $\hat{\boldsymbol\theta}[k] \leftarrow \tilde{\boldsymbol\theta}[k+1]$      ▷ Propagate solution downwards
    repeat
        $\hat{\boldsymbol\theta}[k] \leftarrow$STEP$(\Psi^k, \hat{\boldsymbol\theta}[k])$                 ▷ assuming descent steps
    until a stopping criterion or iteration limit is reached
end for
return $\hat{\boldsymbol\theta}[0]$
```

Choice of Ψ^k: For robust costs the natural approach to construct the sequence $(\Psi^0, \dots, \Psi^{k_{\max}})$ is by appropriate scaling of the kernels. Let $(s_k)_{k=0}^{k_{\max}}$ be a sequence of scaling parameters with $s_0 = 1$ and $s_k < s_{k+1}$. Define

$$\psi^k(r) := s_k^2 \psi(r/s_k) \qquad \text{and} \qquad \Psi^k(\boldsymbol\theta) := \sum\nolimits_i \psi^k(\|\mathbf{f}_i(\boldsymbol\theta)\|). \tag{16}$$

In most cases one will choose $s_k = \tau^k$ for a user-specified value τ (a typical choice also used in our experiments is $\tau = 2$). Due to the following lemma this construction of $(\Psi^k)_{k=0}^{k_{\max}}$ is not only natural, but also has a solid justification:

Lemma 1. *Let ψ be a robust kernel and $s \geq 1$. The following statements hold:*

1. *$\psi(r/s) \leq \psi(r) \leq s^2\psi(r/s)$ for all r.*
2. *Let $0 \leq r' \leq r$. Then we have inequality $\psi(r) - \psi(r') \leq s^2\big(\psi(r/s) - \psi(r'/s)\big)$.*

Proof. $\psi(r/s) \leq \psi(r)$ follows from monotonicity of ψ and that $r/s \leq r$ for $s \geq 1$, yielding one part of the first claim. Since ψ is a robust kernel, then the associated mapping $\phi(z) = \psi(\sqrt{2z})$ is concave and monotonically increasing in its domain $\mathbb{R}_0^+$. Further, ψ is normalized such that $\psi(0) = \phi(0) = 0$. From the concavity of ϕ we deduce that

$$\phi(\alpha z) = \phi\big(\alpha z + (1-\alpha)\cdot 0\big) \geq \alpha\phi(z) + (1-\alpha)\phi(0) = \alpha\phi(z)$$

for all $\alpha \in [0,1]$. Now set $\alpha = 1/s^2$ for $s \geq 1$, and we obtain

$$\phi(z/s^2) = \psi(\sqrt{2z}/s) \geq \phi(z)/s^2 = \psi(\sqrt{2z})/s^2.$$

Substituting $z = r^2/2$ (or $r = \sqrt{2z}$) yields $\psi(r/s) \geq \psi(r)/s^2$ or equivalently $s^2\psi(r/s) \geq \psi(r)$. This proves the first claim.

The inequality in the second claim is equivalent to $s^2\psi(r'/s) - \psi(r') \leq s^2\psi(r/s) - \psi(r)$. The function $d(r) := s^2\psi(r/s) - \psi(r) \geq 0$ is monotonically increasing, since $d'(r) = s\psi'(r/s) - \psi'(r) = r(\omega(r/s) - \omega(r)) \geq 0$ (as ω is monotonically decreasing). This verifies the second claim.

The first statement implies that $\Psi^k(\boldsymbol\theta) \leq \Psi^{k+1}(\boldsymbol\theta)$ for all $\boldsymbol\theta$, and optimizing Ψ^k means that an upper bound of $\Psi^0 = \Psi$ is minimized.[9] The second statement in

[9] Note that Ψ^k is upper bounding Ψ, but generally it is not a majorizer of Ψ (which would additionally require $\Psi^k(\bar{\boldsymbol\theta}) = \Psi(\bar{\boldsymbol\theta})$ at the current solution $\bar{\boldsymbol\theta}$).

the lemma shows that Ψ^k is in a certain sense easier than Ψ^ℓ for $\ell < k$: if $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^+$ are solutions such that $\|\mathbf{f}_i(\boldsymbol{\theta}^+)\| \leq \|\mathbf{f}_i(\bar{\boldsymbol{\theta}})\|$ for all i (i.e. going from $\bar{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}^+$ decreases all residuals), then $\Psi^\ell(\bar{\boldsymbol{\theta}}) - \Psi^\ell(\boldsymbol{\theta}^+) \leq \Psi^k(\bar{\boldsymbol{\theta}}) - \Psi^k(\boldsymbol{\theta}^+)$. Thus, Ψ^k is not only an upper bound of Ψ^{k-1}, but also tends to be steeper.

Stopping criterion: We propose to utilize a relative stopping criterion. Let $\bar{\boldsymbol{\theta}}$ be the current solution and $\boldsymbol{\theta}^+ := \bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}$ be a new solution. Define

$$\mathcal{I}_> := \left\{ i : \mathbf{f}_i(\boldsymbol{\theta}^+) > \mathbf{f}_i(\bar{\boldsymbol{\theta}}) \right\}, \tag{17}$$

i.e. $\mathcal{I}_>$ indexes the strictly increasing residuals after updating the solution. Further, let

$$\Psi^k_>(\boldsymbol{\theta}) := \sum\nolimits_{i \in \mathcal{I}_>} \psi^k(\mathbf{f}_i(\boldsymbol{\theta})) \qquad \Psi^k_\leq(\boldsymbol{\theta}) := \sum\nolimits_{i \notin \mathcal{I}_>} \psi^k(\mathbf{f}_i(\boldsymbol{\theta})) \tag{18}$$

(analogously we introduce $\Psi^{k-1}_\leq(\boldsymbol{\theta})$ and $\Psi^{k-1}_>(\boldsymbol{\theta})$). We have $\Psi^k(\boldsymbol{\theta}) = \Psi^k_\leq(\boldsymbol{\theta}) + \Psi^k_>(\boldsymbol{\theta})$ by construction, and $\Psi^\ell_>(\bar{\boldsymbol{\theta}}) \leq \Psi^\ell_>(\boldsymbol{\theta}^+)$ and $\Psi^\ell_\leq(\boldsymbol{\theta}^+) \leq \Psi^\ell_\leq(\bar{\boldsymbol{\theta}})$ for all $\ell \in \{0, \dots, k_{\max}\}$. We also introduce

$$\Delta^\ell_\leq := \Psi^\ell_\leq(\bar{\boldsymbol{\theta}}) - \Psi^\ell_\leq(\boldsymbol{\theta}^+) \geq 0 \qquad \text{and} \qquad \Delta^\ell_> := \Psi^\ell_>(\boldsymbol{\theta}^+) - \Psi^\ell_>(\bar{\boldsymbol{\theta}}) \geq 0 \tag{19}$$

for all $\ell \in \{0, \dots, k_{\max}\}$ (note the different positions of $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^+$ in $\Delta^\ell_\leq$ and $\Delta^\ell_>$). Now if $\bar{\boldsymbol{\theta}}$ is close to a stationary point of Ψ^k, then $\Delta^k_\leq \approx \Delta^k_>$. Since $\boldsymbol{\theta}^+$ is assumed to improve Ψ^k, we read $\Delta^k_\leq \geq \Delta^k_> 0$ and therefore $\Delta^k_\leq - \Delta^k_> \leq \bar{\eta}$ (for a small value $\bar{\eta} > 0$) indicates that $\bar{\boldsymbol{\theta}}$ is close to a stationary point. Since the functions Ψ^k are scaled differently across the hierarchy, we suggest to use a relative stopping criterion,

$$\rho^k_\Delta := \frac{\Delta^k_\leq - \Delta^k_>}{\Delta^k_\leq + \Delta^k_>} = \frac{\Psi^k(\bar{\boldsymbol{\theta}}) - \Psi^k(\boldsymbol{\theta}^+)}{\Delta^k_\leq + \Delta^k_>} \leq \eta \tag{20}$$

for a user-specified value of η. Due to Lemma 1 the denominator monotonically increases with k, hence the criterion becomes looser for larger k.

Contribution: The novel stopping criterion we derived (Eq. 20) allows to speed up particularly the early stages of graduated optimization. Interestingly, there is a connection between the above stopping criterion and the gain ratio

$$\rho^k_\Psi := \frac{\Psi^{k-1}(\bar{\boldsymbol{\theta}}) - \Psi^{k-1}(\boldsymbol{\theta}^+)}{\Psi^k(\bar{\boldsymbol{\theta}}) - \Psi^k(\boldsymbol{\theta}^+)}, \tag{21}$$

that is commonly used in trust region methods (e.g. [25]) to evaluate the quality of a surrogate model (here Ψ^k) w.r.t. a target cost (Ψ^{k-1}):

Lemma 2. *Let $\eta \in (0,1)$. If $\rho^k_\Psi \geq \frac{\eta+1}{2\eta} > 0$ or $\rho^k_\Psi \leq \frac{\eta-1}{2\eta} < 0$ then $\rho^k_\Delta \leq \eta$.*

The lemma asserts that if Ψ^{k-1} either increases or decreases sufficiently faster than Ψ^k, then we are near a stationary points of Ψ^k (according to the stopping criterion Eq. 20). It is less relevant in practice, but tells us that Ψ^k and Ψ^{k-1} (or Ψ^ℓ for any $\ell < k$) cannot behave too different when far from a local minimum. The proof uses Lemma 1 and is given in [27].

7 Numerical Results

In this section we compare the performance of the different approaches for robust cost optimization, and we are mostly interested in the quality (i.e. achieved objective value) that is reached after a sensible amount of run-time.

Implementation remarks: The core of our implementations is a sparse but direct Cholesky solver from the SuiteSparse libraries [6]. We apply Levenberg-type damping $\mathtt{J}^\top \mathtt{J} + \lambda \mathtt{I}$ to (i) ensure the system matrix is sufficiently positive definite for a direct solver and (ii) to obtain a dampled Newton/Gauss-Newton method for non-linear problems. The damping parameter is adjusted using the classical $\times 10/\div 10$ rule. In the graduated optimization method we used 6 scale levels (i.e. $k_{\max} = 5$), where the scale parameter is doubled at each level. The r.h.s. η in the stopping criterion Eq. 20 is set to $\eta = 1/5$. In the figures we abbreviate lifted Gauss-Newton and lifted Newton by l-G-N and l-Newton, respectively. GOM refers to graduated optimization with an uniform allocation of iterations at each level, and GOM+ leverages Eq. 20 as stopping criterion. We use IRLS as direct method inside GOM. We allow 100 iterations (i.e. 100 times solving the underlying system equation for the update $\varDelta\boldsymbol{\theta}$) for each method, which results in rather similar wall-clock runtimes for all methods.

7.1 Synthetic data: Robust mean and Image smoothing

Estimating the mode (i.e. robust mean) of data points is arguably the simplest robust optimization problem. We follow [26] and create Gaussian distributed inliers and uniformly distributed outliers in a $[-20, 20]^D$ domain. The mean of the Gaussian inlier distribution is also uniformly sampled from the same domain, hence in most cases the outliers will not be symmetrically distributed around the inlier points. Let $(\mathbf{y}_1, \dots, \mathbf{y}_N)$ be the entire set of data points, then the task is to estimate $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \varPsi^{\text{mean}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_i \psi(\|\boldsymbol{\theta} - \mathbf{y}_i\|)$, where our choice of ψ is the Welsch kernel, $\psi(r) = \frac{1}{2}(1 - e^{-r^2})$. The initial value $\boldsymbol{\theta}^0$ provided to the optimization methods is uniformly sampled as well. We depict in Fig. 2 the average objective values (and corresponding standard deviation using 100 runs) reached by several methods for different choices of inlier ratios and $D = 3$. The included methods are standard IRLS, the accelerated graduated optimization method (GOM+), the lifted Gauss-Newton and Newton methods parametrizing either $w(u) = u^2$ (l-G-N^a, l-Newton^a) or $w(u) = \text{sigmoid}(u)$ (l-G-N^b, l-Newton^b). Graduated optimization (GOM+) is a clear winner, and the lifted Newton method dominates the corresponding lifted Gauss-Newton version. Using the sigmoid parametrization is clearly beneficial, and we will use this parametrization from now on in the lifting-based methods.

Since $\boldsymbol{\theta}$ has very small dimension in the robust mean example, these types of low-parametric robust estimation problems are easily solved by random sampling methods such as RANSAC and variants (e.g. [10, 22]). Therefore we now consider a problem with a high dimensional vector of unknowns. We selected the weak membrane energy for image smoothing (e.g. [5]), which is a prototypical instance

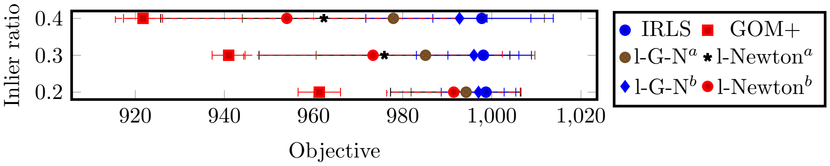

Fig. 2. Final objective values for robust mean instances at varying inlier ratios.

Method	IRLS	l-G-N	l-Newton	GOM+
Objective	231.8133±1.9040	45.0811±0.0861	45.0496±0.0446	45.0463±3.66e-13

Table 1. Final objective values for the weak membrane energy.

of a difficult low-level image processing problem. Given an observed image $\mathbf{u}$ the weak membrane energy is given by $\Psi^{\text{Membrane}}(\boldsymbol{\theta};\mathbf{u}) = \sum_{i\in\mathcal{V}} \psi^{\text{data}}(\theta_i - u_i) + \sum_{(i,j)\in\mathcal{E}} \psi^{\text{smooth}}(\theta_i - \theta_j)$. The node set $\mathcal{V}$ corresponds to pixels, and the edge set $\mathcal{E}$ is induced by the 4-neighborhood. ψ^{data} and ψ^{smooth} are based on the smooth truncated kernel (see [27]). Table 1 lists the reached average objectives (and standard deviation over 25 runs) for the different methods for the 256 × 256 "Lena" image. The initial guesses $\boldsymbol{\theta}^0$ are uniformly sampled images from $[0,1]^{|\mathcal{V}|}$. Only IRLS falls clearly behind in terms of reported optimal value. More interesting is the evolution of objective values shown in Fig. 3, that allows to make two observations: the lifted Gauss-Newton method is the fastest to achieve a near optimal value, and the stopping criterion leveraged in GOM+ significantly accelerates convergence of graduated optimization. Further (also visual) results are provided in [27].

7.2 Real data: Robust bundle adjustment

One of the main applications of robust cost minimization in computer vision is bundle adjustment (BA). We took 10 problem instances (the list is provided in [27]) from the "bundle adjustment in the large collection" [2]. The robust

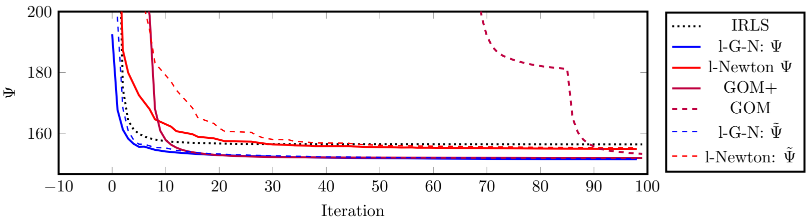

Fig. 3. Evolution of Ψ^{Membrane} w.r.t. the number of iterations. For the lifting based methods we plot the original cost Ψ and lifted one $\tilde{\Psi}$ (Eq. 7).

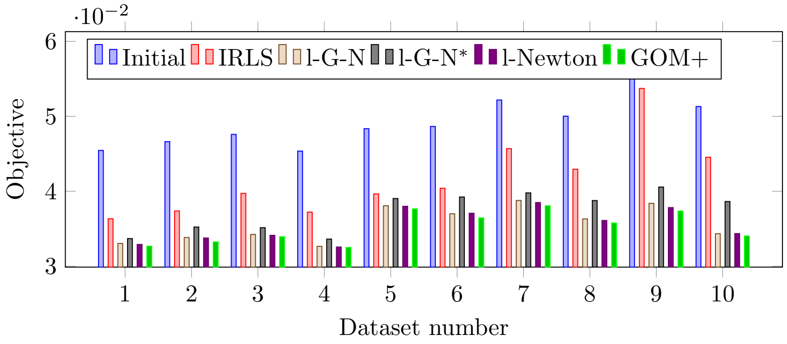

Fig. 4. Objective values (normalized w.r.t. the number of image measurements) reached by the different methods for *linearized* BA.

bundle objective is given by

$$\Psi^{\mathrm{BA}}(\{\mathtt{R}_i\},\{\mathbf{t}_i\},\{\mathbf{X}_j\}) := \sum\nolimits_{i,j} \psi\big(\|\pi(\mathtt{R}_i\mathbf{X}_j+\mathbf{t}_i)-\mathbf{q}_{ij}\|\big), \tag{22}$$

where $\mathbf{q}_{ij} \in \mathbb{R}^2$ is the observed image observation of the j-th 3D point $\mathbf{X}_j \in \mathbb{R}^3$ point in the i-th image (which has associated parameters $\mathtt{R}_i \in SO(3)$ and $\mathbf{t}_i \in \mathbb{R}^3$). $\pi(X) = X/X_3$ is the projection function of a pinhole camera model. $\mathbf{q}_{ij}$ is measured on the image plane, i.e. the original pixel coordinates are premultiplied by the (provided) inverse calibration matrix. ψ is chosen to be the smooth truncated kernel with parameter $\frac{1}{2}$, i.e. $\psi(r) = \frac{1}{16}\left(1 - [1-4r^2]_+^2\right)$. This choice makes the problem instances sufficiently difficult, as the initial inlier ratio of image observations ranges between 14% and 50% (depending on the dataset). The inlier ratios obtained after robust cost minimization cluster around 60% for the best obtained local minima.

First, we focus on a *linearized* version of bundle adjustment, where the residuals $\mathbf{f}_{ij} = \pi(\mathtt{R}_i\mathbf{X}_j + \mathbf{t}_i) - \mathbf{q}_{ij}$ are replaced by their linearized versions w.r.t. the provided initial values. The non-robust objective is therefore convex, and the performance differences depicted in Fig. 4 indicate how well each method escapes poor local minima. In order to obtain similar objective values regardless of the dataset size, the objective values are normalized w.r.t. the number of image

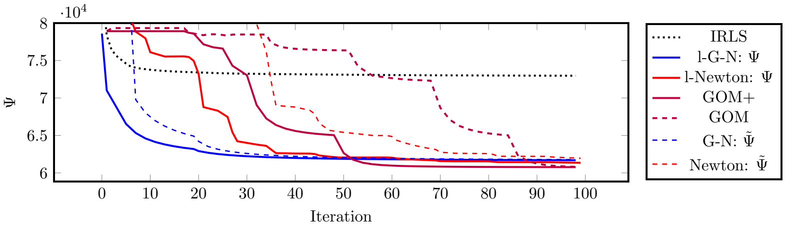

Fig. 5. Evolution of Ψ^{BA} w.r.t. the number of iterations for the Venice-427 instance.

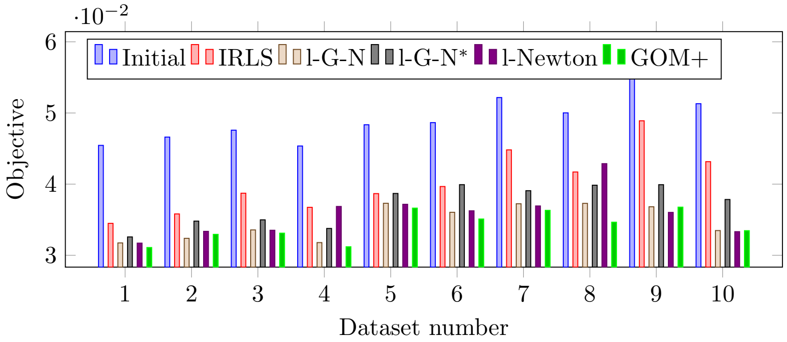

Fig. 6. Objective values (normalized w.r.t. the number of image measurements) reached by the different methods for *metric* BA.

measurements. The unnormalized BA objective values Eq. 22 are approximately between 6000 and 60000 (depending on the dataset and method). Thus, none of the methods is in its respective comfort zone. IRLS is clearly inferior to the other methods, and GOM+ is slightly ahead of the lifted formulations. l-G-N* is the lifted Gauss-Newton method, but the lifted parameters u_i are initialized to their optimal value (given the initial values $\boldsymbol{\theta}^0$). The resulting performance is between IRLS and l-G-N. If we take a closer look on the evolution of objectives (Fig. 5), then the lifted Gauss-Newton method reduces the actual cost Ψ^{BA} very quickly, although graduated optimization eventually reaches a better minimum.

Fig. 6 illustrates the reached objectives (normalized w.r.t. the number of image measurements) by the different methods for non-linear metric bundle adjustment. Due to the additional non-linearity introduced by the non-robust objective, the results are more diverse than the ones in Fig. 4. In particular, the lifted Newton method shows an unstable behavior. Details and results for dense disparity estimation are provided in [27].

8 Conclusion and future work

In this work we first unified several direct and lifting-based methods for robust cost minimization. We also demonstrated that a graduated optimization method has very competitive performance in terms of the reached objective values and in terms of speed of convergence. Hence, our recommendation is as follows: a lifted Gauss-Newton method is a very strong candidate when very fast decrease of objectives is desired, and the proposed graduated optimization approach is the method of choice when reaching the best objective is the main interest—especially when the quality of the initial solution is unknown.

The fact that the best performing methods "forget" to a large extend the given initial solution is not very satisfactory. Future work will investigate whether methods adapting to the quality of the provided starting point result in faster overall convergence.

References

1. Agarwal, S., Mierle, K., Others: Ceres solver. https://code.google.com/p/ceres-solver/
2. Agarwal, S., Snavely, N., Seitz, S.M., Szeliski, R.: Bundle adjustment in the large. In: Proc. ECCV, pp. 29–42. Springer (2010)
3. Albersmeyer, J., Diehl, M.: The lifted newton method and its application in optimization. SIAM Journal on Optimization **20**(3), 1655–1684 (2010)
4. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer vision and image understanding **63**(1), 75–104 (1996)
5. Blake, A., Zisserman, A.: Visual reconstruction (1987)
6. Davis, T.A., Hu, Y.: The university of florida sparse matrix collection. ACM Transactions on Mathematical Software (TOMS) **38**(1), 1 (2011)
7. Dunlavy, D.M., O'Leary, D.P.: Homotopy optimization methods for global optimization. Tech. rep., Sandia National Laboratories (2005)
8. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular slam. In: European Conference on Computer Vision. pp. 834–849. Springer (2014)
9. Engels, C., Stewénius, H., Nistér, D.: Bundle adjustment rules. In: Photogrammetric Computer Vision (PCV) (2006)
10. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
11. Geman, D., Reynolds, G.: Constrained restoration and the recovery of discontinuities. IEEE Trans. Pattern Anal. Mach. Intell. **14**(3), 367–383 (1992)
12. Geman, D., Yang, C.: Nonlinear image recovery with half-quadratic regularization. IEEE Transactions on Image Processing **4**(7), 932–946 (1995)
13. Green, P.J.: Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society. Series B (Methodological) pp. 149–192 (1984)
14. Hong, J.H., Fitzgibbon, A.: Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4130–4138 (2015)
15. Huber, P.J.: Robust statistics. Wiley (1981)
16. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. science **220**(4598), 671–680 (1983)
17. Lange, K., Hunter, D.R., Yang, I.: Optimization transfer using surrogate objective functions. Journal of computational and graphical statistics **9**(1), 1–20 (2000)
18. Liwicki, S., Zach, C., Miksik, O., Torr, P.H.: Coarse-to-fine planar regularization for dense monocular depth estimation. In: European Conference on Computer Vision. pp. 458–474. Springer (2016)
19. Mobahi, H., Fisher, J.W.: On the link between gaussian homotopy continuation and convex envelopes. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. pp. 43–56. Springer (2015)
20. Mobahi, H., Fisher III, J.W.: A theoretical analysis of optimization by gaussian continuation. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
21. Rose, K.: Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE **86**(11), 2210–2239 (1998)

22. Torr, P.H., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding **78**(1), 138–156 (2000)
23. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment – A modern synthesis. In: Vision Algorithms: Theory and Practice. LNCS, vol. 1883, pp. 298–372 (2000)
24. Ye, M., Haralick, R.M., Shapiro, L.G.: Estimating piecewise-smooth optical flow with global matching and graduated optimization. IEEE transactions on pattern analysis and machine intelligence **25**(12), 1625–1630 (2003)
25. Yuan, Y.: A review of trust region algorithms for optimization. In: ICM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics (1999)
26. Zach, C., Bourmaud, G.: Iterated lifting for robust cost optimization. In: Proc. BMVC (2017)
27. Zach, C., Bourmaud, G.: Descending, lifting or smoothing: Secrets of robust cost optimization (supplementary material). In: Proc. ECCV (2018)
28. Zach, C., Bourmaud, G.: Multiplicative vs. additive half-quadratic minimization for robust cost optimization. In: Proc. BMVC (2018)
29. Zach, C.: Robust bundle adjustment revisited. In: Proc. ECCV. pp. 772–787 (2014)
30. Zhou, Q.Y., Park, J., Koltun, V.: Fast global registration. In: Proc. ECCV. pp. 766–782. Springer (2016)
31. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., Stamminger, M.: Real-time non-rigid reconstruction using an RGB-D camera. In: SIGGRAPH (2014)