On the Link between Gaussian Homotopy Continuation and Convex Envelopes

Hossein Mobahi and John W. Fisher III

Computer Science and Artificial Intelligence Lab. (CSAIL)
Massachusetts Institute of Technology (MIT)
{hmobahi,fisher}@csail.mit.edu

Abstract. The continuation method is a popular heuristic in computer vision for nonconvex optimization. The idea is to start from a simplified problem and gradually deform it to the actual task while tracking the solution. It was first used in computer vision under the name of graduated nonconvexity. Since then, it has been utilized explicitly or implicitly in various applications. In fact, state-of-the-art optical flow and shape estimation rely on a form of continuation. Despite its empirical success, there is little theoretical understanding of this method. This work provides some novel insights into this technique. Specifically, there are many ways to choose the initial problem and many ways to progressively deform it to the original task. However, here we show that when this process is constructed by Gaussian smoothing, it is optimal in a specific sense. In fact, we prove that Gaussian smoothing emerges from the best affine approximation to Vese's nonlinear PDE. The latter PDE evolves any function to its convex envelope, hence providing the optimal convexification.

Keywords: Continuation Method, Diffusion Equation, Nonconvex Optimization, Vese's PDE.

1 Introduction

Minimization of nonconvex energy functions arises frequently in computer vision. Examples include image segmentation [49], image alignment [67], image completion [46], dictionary learning [44], part-based models [25], and optical flow [62]. Unfortunately, a severe limitation of nonconvex problems is that finding their global minimum is generally intractable.

Some possible options for handling nonconvex tasks include[1] local optimization methods (e.g. gradient descent), convex surrogates, and the continuation method. Each of these ideas has its own merit and is preferred in certain settings. For example, local methods are useful when most local minima produce reasonably good solutions; otherwise the algorithm may get stuck in poor local minima. Convex surrogates are helpful when the nonconvexity of the task is mild, so that little structure is lost by the convex approximation. For example, it has been observed

[1] In this paper we only discuss deterministic schemes.

X.-C. Tai et al. (Eds.): EMMCVPR 2015, LNCS 8932, pp. 43–56, 2015.

that for face recognition problem, the nonconvex sparsity encouraging ℓ_0 norm can be replaced by the convex ℓ_1 and yet produce impressive result [69]. Recently [23] proposed an a surrogate construction with bounded discrepancy between the solution of the convexified and original task.

The third idea is to utilize the continuation method. It solves a sequence of subproblems, starting from a convex (hence easy) task and progressively changing it to the actual problem while tracing the solution. Such complexity progression is in contrast to convex surrogates that produce a one-shot relaxation. Here, the solution of each subproblem guides solving the next one. This approach is often useful when the nonconvexity of the problem is so severe that convex surrogates cannot provide any meaningful approximation.

In this paper, we focus on optimization by the continuation method. The idea has been known to the computer vision community for at least three decades. This dates back to the works of Terzopoulos [63], Blake and Zisserman [6], and Yuille [72,73,74,75,76]. Since then, this technique has been used with growing interest to solve some difficult optimization problems. In particular, it is a key component in several state-of-the-art solutions for computer vision and machine learning problems as we discuss in Section 2.

Despite its long history and empirical success, there is little understanding about the fundamental aspects of this method. For example, it is known that the continuation method cannot always find the global minimizer of all nonconvex tasks. In fact, the quality of the solution attained by this approach heavily depends on the choice of the subproblems. However, there are endless choices for the initial convex problem, and endless ways to progressively change it to the original nonconvex task. Obviously, some of these choices should work better than the others. However, to date, there is no known principle for preferring one construction versus another.

For example, a possible way to construct the subproblem sequence is by Gaussian smoothing [50,47]. The idea is to convolve the original nonconvex function with an isotropic Gaussian kernel at various bandwidth values. This generates a sequence of functions varying from a highly smoothed (large bandwidth) to the actual nonconvex function (zero bandwidth). In fact, it can be proved that under certain conditions, enough smoothing can lead to a convex function [43]. The convexity implies that finding the minimizer of the smoothed function is easy. This minimizer is used to initialize the next subproblem, with slightly smaller bandwidth. The process repeats until reaching the last subproblem, which is the actual task. Since this type of progression goes from low-frequency toward fully detailed, it is also called *coarse-to-fine optimization*.

In this paper, we provide original insights into the choice of subproblems for the continuation method. Specifically, we prove that constructing the subproblems by Gaussian smoothing of the nonconvex function is optimal in a specific sense. Recall that the continuation method starts from an already convex objective and progressively maps it to the actual nonconvex function. Among infinite choices for the initial convex task, the *convex envelope* of the nonconvex problem is (in many senses) the best choice. Unfortunately, the convex envelope of an arbitrary function

is nontrivial and generally expensive to compute. Vese has shown that the convex envelope of a function can be generated by an evolutionary PDE [66]. However, this PDE does not have an analytical solution. Our contribution is to prove that the best affine approximation to Vese's PDE results in the *heat equation*. The solution of the latter is known; it is the Gaussian convolution of the nonconvex function. Hence, Gaussian smoothing is the outcome of the best affine approximation of the (nonlinear) convex envelope generating PDE.

2 Related Works

Here we review some remarkable works that rely on the concept of optimization by the continuation method.

In computer vision, the early works around this concept were Blake and Zisserman's *Graduated Non-Convexity* (GNC) [6] as well as works by Terzopoulos' [63], both on surface reconstruction problems. Shortly afterward, Geiger and Girosi [29] as well as Yuille [72] used similar concepts from a statistical physics viewpoint. The latter method is known as *Mean Field Annealing* (MFA). Motivated by problems in stereo and template matching, Yuille popularized MFA in a series of works [30,72,73,74,75,76]. MFA is a *deterministic* variant of simulated annealing[2], where the stochastic behavior is approximated by the mean state. This model starts from high temperature (smoother energy and hence fewer extrema) and gradually cools down toward the desired optimization task.

Since then, the concept of optimization by the continuation method has been successfully utilized in various vision applications such as image segmentation [9], shape matching [64], image deblurring [8], image denoising [54,51], template matching [22], pixel correspondence [40], active contours [18], Hough transform [39], edge detection [78], early vision [5], robot navigation [52], and image matting [53]. In fact, many computer vision methods that rely on multiscale image representation within the optimization loop are implicitly performing the continuation method, e.g. for image alignment [47].

The growing interest in this method within computer vision community has made it one of the most popular solutions for the contemporary nonconvex minimization problems. Just within the past few years, it has been utilized for low-rank matrix recovery [45], error correction [48], super resolution [19], photometric stereo [70], image segmentation [35], face alignment [57], 3D surface estimation [1], motion estimation in videos [61], optical flow [10,62], shape and illumination recovery [2], and dense correspondence of images [36]. The last three are in fact *state of the art* solutions for their associated problems.

Independently, the machine learning community has been using similar ideas for optimization. Notably, Rose popularized the method of *Deterministic Annealing* (DA) for clustering problems [55]. This method starts from the *maximum entropy* solution (the simple task), and gradually reduces the entropy

[2] There is some conceptual similarity between simulated annealing (SA) and some of the continuation methods. However, SA is an MCMC method and is known for its very slow convergence. The continuation methods studied here are deterministic and converge much faster [7,40].

to only leave the actual objective function. Variants of DA have been recently used for learning occluding objects [20], object tracking [33], image deblurring [41], clustering boolean data [26], graph clustering [56], unsupervised language learning [60]. Chapelle has utilized continuation in various applications such as semi-supervised learning [12,13,59], semi-supervised structured output [21], multiple instance learning [28], and ranking [14]. Bengio argues that some recent breakthroughs in the training of deep architectures [34,24], has been made by algorithms that use some form of continuation for learning [4].

Other examples that utilize continuation for optimization are clustering [32], graph matching [31,77,42], multiple instance learning [37], language modeling [3], manifold sampling [58], and ℓ_0 norm minimization [65]. One of the most interesting applications, however, has been recently introduced by [16,17]. The goal is to find optimal parameters in computer programs. The authors define a smoothing operator acting on programs to construct smooth interpretations. They then seek the optimal parameters by starting from highly smoothed interpretations and gradually reducing the smoothing level. The idea is further extended to smoothing the space of proofs and seeking the optimal proof to a problem by the continuation method [15].

Throughout this paper, we use x for scalars, $\boldsymbol{x}$ for vectors, $\boldsymbol{X}$ for matrices, and $\mathcal{X}$ for sets. Here $\|\boldsymbol{x}\|$ means $\|\boldsymbol{x}\|_2$ and $\triangleq$ means equality by definition. When a function is denoted as $g(\boldsymbol{x};t)$, the gradient ∇, Hessian ∇^2 and Laplacian Δ operators are only applied to the vector $\boldsymbol{x}$ and not t. The convolution operator is denoted by $\star$. The isotropic Gaussian kernel with standard deviation σ is shown by k_σ,

$$k_\sigma(\boldsymbol{x}) \triangleq \frac{1}{(\sqrt{2\pi}\sigma)^{\dim(\boldsymbol{x})}} e^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}} .$$

3 Optimization by Continuation

Given an (possible nonconvex) objective function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} = \mathbb{R}^n$. Consider an embedding of f into a family of functions $g : \mathcal{X} \times \mathcal{T}$, where $\mathcal{T} \triangleq [0, \infty)$, with the following properties. First, $g(\boldsymbol{x}, 0) = f(\boldsymbol{x})$. Second, when $t \to \infty$, then $g(\boldsymbol{x}, t)$ is strictly convex and has a unique minimizer (denoted by $\boldsymbol{x}_\infty$). Third, $g(\boldsymbol{x}, t)$ is continuously differentiable in $\boldsymbol{x}$ and t. Such embedding g is sometimes called a *homotopy*, as it continuously transforms one function to another.

Define the curve $\boldsymbol{x}(t)$ for $t \geq 0$ as one with the following properties. First, $\lim_{t\to\infty} \boldsymbol{x}(t) = \boldsymbol{x}_\infty$. Second, $\forall t \geq 0 \quad ; \quad \nabla g\big(\boldsymbol{x}(t), t\big) = \boldsymbol{0}$. Third, $\boldsymbol{x}(t)$ is *continuous* in t. This curve simply sweeps a specific stationary path of g originated at $\boldsymbol{x}_\infty$, as the parameter t progresses backward (See Figure 1). In general, such curve neither needs to exist, nor needs to be unique. However, with some additional assumptions on g, it is possible to guarantee *existence* and *uniqueness* of $\boldsymbol{x}(t)$, e.g. by Theorem 3 of [71].

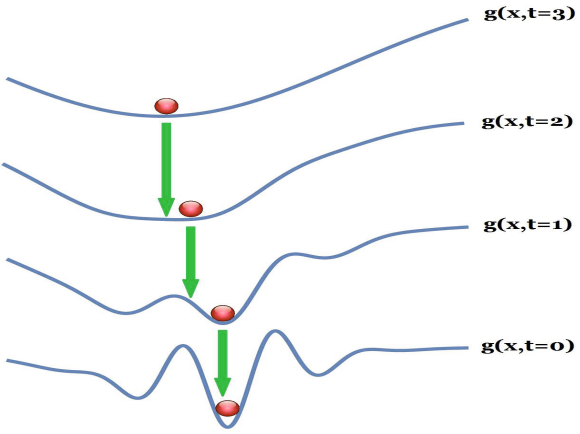

Fig. 1. Plots show g versus x for each fixed time t. The marble indicates the location for $x(t)$.

Algorithm 1. Algorithm for Optimization by the Continuation Method

1: Input: $f : \mathcal{X} \to \mathbb{R}$, Sequence $t_0 > t_1 > \cdots > t_m = 0$.
2: $\boldsymbol{x}_0 =$ global minimizer of $g(\boldsymbol{x}; t_0)$.
3: **for** $k = 1$ **to** m **do**
4: $\quad \boldsymbol{x}_k =$ Local minimizer of $g(\boldsymbol{x}; t_k)$, initialized at $\boldsymbol{x}_{k-1}$.
5: **end for**
6: Output: $\boldsymbol{x}_m$

In practice, the continuation method is realized as follows. First, $\boldsymbol{x}_\infty$ is either derived analytically[3] or approximated numerically as $\arg\min_{\boldsymbol{x}} g(\boldsymbol{x}; t)$ for large enough t. The latter can use standard convex optimization tools as $g(\boldsymbol{x}; t)$ approaches a convex function in $\boldsymbol{x}$ for large t. Then, the stationary path $\boldsymbol{x}(t)$ is numerically tracked until $t = 0$ (See Algorithm 1). As discussed in Section 2, for a wide range of applications, the continuation solution $\boldsymbol{x}(0)$ often provides a good local minimizer of $f(\boldsymbol{x})$, if not the global minimizer.

4 Motivation for Gaussian Homotopy

There are some limited number of studies on very specific problems which guarantee the continuation method can discover the global minimum of the problem. An example of this kind is the work by Yuille and Kosowsky [38] on assignment problem. However, in general, there is no guarantee for the continuation method to reach the global minimizer of $f(\boldsymbol{x})$.

In fact, the quality of the solution attained by the continuation method depends on the choice of the homotopy map $g(\boldsymbol{x}; t)$. It is therefore crucial to choose

[3] For functions whose tails vanish fast enough, this point is simply the center of mass of the function [43].

$g(\boldsymbol{x};t)$ in the most sensible way. Currently, there is no pointer in the literature to justify one homotopy versus others. For example, Fua and Leclerc [27] use $g(\boldsymbol{x},t) = f(\boldsymbol{x}) + t\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A} \succ \boldsymbol{O}$. Blake and Zisserman [6] utilize a task-tailored polynomial map. Methods based on deterministic annealing use negative entropy $g(\boldsymbol{x},t) = f(\boldsymbol{x}) + t\,\boldsymbol{x}^T \log(\boldsymbol{x})$ (applicable only to nonnegative variables $\boldsymbol{x}$) [55,56]. Nielson [50] and Mobahi [47] use Gaussian homotopy by convolving f with the Gaussian kernel, i.e. $g(\,.\,,t) = f \star k_t$. When Gaussian homotopy is used for optimization, it is sometimes called *coarse-to-fine optimization*[4].

In this section, we claim that Gaussian homotopy is optimal in a specific sense; it solves the best affine approximation (around the origin of the function space, i.e. the function $f(\boldsymbol{x}) = 0$) to a nonlinear PDE that generates convex envelopes. We will postpone the proof to the next section.

By definition, a homotopy for optimizing $f(\boldsymbol{x}) = g(\boldsymbol{x};0)$ must continuously convexify it to $g(\boldsymbol{x},\infty)$. Among all convex choices $g(\boldsymbol{x},\infty)$, the *convex envelope* is the optimal convexifier of f in many senses. For example, it provides the best (largest) possible convex underestimator of the f. Furthermore, geometrically, the convex envelope is precisely the function whose epigraph coincides with the convex hull of the epigraph of f.

The convex envelope, however, is often unknown itself and its computation is generally very expensive. Interestingly, Vese [66] has characterized an elegant PDE that if its initial condition is set to $f(\boldsymbol{x})$, it evolves toward the convex envelope of f and reaches there in the limit $t \to \infty$. More precisely, this is a nonlinear PDE that evolves a function $v(\boldsymbol{x};t)$ for $v : \mathcal{X} \times \mathcal{T}$ as the following,

$$\frac{\partial}{\partial t} v = \sqrt{1 + \|\nabla v\|^2}\, \min\{0, \lambda_{\min}(\nabla^2 v)\} \quad , \quad \text{s.t. } v(\,.\,;0) = f(\,.\,), \tag{1}$$

where $\lambda_{\min}(\nabla^2 v)$ is the smallest (sign considered) eigenvalue of the Hessian of v. Intuitively, this PDE acts like a conditional diffusion process. At any evolution moment t, $v(\boldsymbol{x};t)$ is spatially diffused at points $\boldsymbol{x}$ where $v(\boldsymbol{x};t)$ is nonconvex and is left as is at points $\boldsymbol{x}$ where $v(\boldsymbol{x};t)$ is convex (nonconvexity and convexity of v here are w.r.t. to the variable $\boldsymbol{x}$). Consequently, throughout the evolution, nonconvex structures diminish by diffusion while convex structures sustain.

Vese's PDE involves the nonsmooth function min, which complicates its treatment for the purpose of this paper[5]. Hence, we introduce the *modified Vese's PDE* by replacing min with a smooth approximation,

$$\frac{\partial}{\partial t} u = \sqrt{1 + \|\nabla u\|^2}\, m\big(\boldsymbol{\lambda}(\nabla^2 u)\big) \quad , \quad \text{s.t. } u(\,.\,;0) = f(\,.\,) \tag{2}$$

$$m(\boldsymbol{\lambda}) \triangleq \frac{\sum_{k=1}^{n} \lambda_k e^{-\frac{\lambda_k}{\delta}}}{1 + \sum_{k=1}^{n} e^{-\frac{\lambda_k}{\delta}}},$$

[4] This is because moving from large to small t reveals coarse to fine structure of the optimization landscape.

[5] The difficulty arises later in Section 5, where we need to differentiate the r.h.s. of (1), but min is not differentiable.

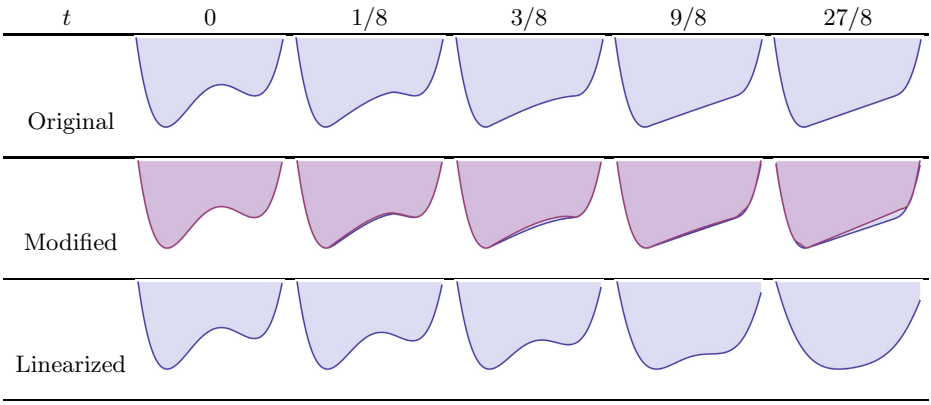

Fig. 2. Evolution of the function $x^4+2x^3-12x^2-2x$ by Vese's original PDE (1) (top), versus its modified (middle) and linearized (3) (bottom) forms. Since the difference between the original and modified version of Vese's PDE is very subtle, in the middle row the modified solution (magenta) is superimposed on the original solution (blue). The modified version uses $\delta = 10$ to make the the two visually distinct (with $\delta = 1$ these plots already become indistinguishable). While all three evolutions convexify the initial function, the original and modified Vese's equations respectively generate the perfect and close approximate to the convex envelope.

where $\delta > 0$, and $\boldsymbol{\lambda} \triangleq (\lambda_1, \ldots, \lambda_n)$ is a $n \times 1$ vector. Observe that $\lim_{\delta \to 0^+} m(\boldsymbol{\lambda}) = \min\{0, \lambda_1, \ldots, \lambda_n\}$. Hence, we can construct an arbitrarily close approximation to $\min\{0, \lambda_1, \ldots \lambda_n\}$ by choosing a small enough $\delta > 0$. Although Vese's PDE and its modified form are not identical, from practical viewpoint their difference is often negligible (See Figure 2, the top and middle rows). Hence, we proceed with the modified Vese's PDE for our analysis in Section 5, solely for technical reasons.

Neither the original nor the modified versions of Vese's PDE can be solved analytically due to their highly *nonlinear* nature. However, in Section 5 we will prove that the best affine approximation of the modified Vese's operator around the origin of the function space (i.e. the function $f(\boldsymbol{x}) = 0$) is the *Laplace* operator, hence the following approximation (See Figure 2 for an illustrative example),

$$\frac{\partial}{\partial t}\hat{u} = \frac{1}{n+1}\Delta\hat{u} \quad , \quad \hat{u}(\,.\,;0) = f(\,.\,)\,. \tag{3}$$

The resulted PDE (3) is essentially the *heat equation* [68] on the domain $\mathcal{X} = \mathbb{R}^n$ with the initial condition $\hat{u}(\boldsymbol{x}, 0) = f(\boldsymbol{x})$. The solution of the heat equation in (3) is known to have the following form,

$$\hat{u}(\boldsymbol{x}; t) = \left(\frac{n+1}{4\pi t}\right)^{\frac{n}{2}} \left[f(.) \star e^{-\frac{\|\,.\,\|^2\,(n+1)}{4t}}\right](\boldsymbol{x})\,.$$

The function $\hat{u}$ can be reparameterized in its scale parameter via $\sigma^2 = \frac{2t}{n+1}$. This only changes the speed of progression, which is not crucial for our conclusion here. Hence, the homotopy can be expressed as the convolution of f with the Gaussian kernel k_σ as below,

$$\hat{h}(\boldsymbol{x};\sigma) = \left[f \star k_\sigma\right](\boldsymbol{x})\,.$$

This approximation buys us a significant benefit in practice, for the following reason. While the nonlinear operator appearing in the original PDE (1) or its modified version (2) does not allow for a closed form solution, the linear PDE (3) makes this possible, provided that the integral for the Gaussian convolution of f in (4) has a closed form expression. The latter is true for some important classes of functions including polynomials and Gaussian bumps. Both of these classes are rich enough to represent almost any function, respectively through Taylor series and Gaussian Radial-Basis-Functions (RBF). For example, [47] uses these function spaces in order to formulate the image alignment problem and then solves it by Gaussian homotopy continuation.

Note that unlike Vese's equation that always evolves the nonconvex function to a convex one (in fact, to its convex envelope), heat equation does not necessarily produce a convex function. However, it does so for functions that on average (across all points) are convex[6]. There exist sufficient conditions[7] on the nonconvex functions to guarantee their convexity after enough smoothing [43].

5 Affine Approximation of Modified Vese's Operator

Here we prove our earlier claim that the best affine approximation to the modified Vese's PDE around the origin of function space (i.e. the function $f(\boldsymbol{x}) = 0$) leads to the Laplace operator. We first need a few definitions. In the sequel, let $\mathcal{H}$ be the space of twice differentiable functions $h : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \triangleq \mathbb{R}^n$. We consider linear and nonlinear operators that have the form $\mathcal{H} \to \mathcal{H}$ and denote them by $\mathscr{L}$ and $\mathscr{N}$ respectively. We say an operator is linear if and only if it obeys $\forall h_1 \in \mathcal{H}\,,\, h_2 \in \mathcal{H}\,,\, a \in \mathbb{R}\,,\, b \in \mathbb{R}\,;\, \mathscr{L}\{ah_1 + bh_2\} = a\,\mathscr{L}\{h_1\} + b\,\mathscr{L}\{h_2\}$.

Definition 1 (Affine Operator). *An affine operator is the form $\mathcal{L}\{h\} + c$ where $\mathcal{L}$ is a linear operator in h and c is constant in h.*

Definition 2 (Modified Vese's Operator). *The modified Vese's operator is defined as the operator acting on the function $h \in \mathcal{H}$ to return $\sqrt{1 + \|\nabla h\|^2}\, m\big(\boldsymbol{\lambda}(\nabla^2 h)\big)$, where $m\big(\boldsymbol{\lambda}\big) \triangleq \frac{\sum_{k=1}^n \lambda_k e^{-\frac{\lambda_k}{\delta}}}{1+\sum_{k=1}^n e^{-\frac{\lambda_k}{\delta}}}$.*

[6] For example, in univariate functions $f(x)$, the Gaussian smoothed function is $g(x;\sigma) \triangleq [f \star k_\sigma](x)$ and hence $g''(x;\sigma) \triangleq [f'' \star k_\sigma](x)$. When $\sigma \to \infty$, convolution with $k_\sigma(x)$ acts an averaging operator. Hence if $\int_{\mathcal{X}} f''(x)\,dx > 0$, i.e. f is on average convex, then $\forall x\,;\, g''(x;\sigma) > 0$ (i.e. $g(x;\sigma)$ is convex everywhere) a large enough σ.

[7] For example, if the tails vanish fast enough and the $-\infty < \int_{\mathcal{X}} f(\boldsymbol{x})\,d\boldsymbol{x} < 0$, the it is guaranteed that for a large enough σ, $g(\boldsymbol{x};\sigma)$ is convex. See [43] for details.

Definition 3 (Best Affine Approximation of a Nonlinear Operator). *Consider a $h \in \mathcal{H}$ and suppose it is resulted by perturbing some function $h^* \in \mathcal{H}$ by the function $\epsilon\phi$, that is,*

$$h = h^* + \epsilon\phi\,, \tag{4}$$

where $\phi \in \mathcal{H}$ and $\epsilon \in \mathbb{R}$. Suppose $\mathscr{N}\{h^* + \epsilon\phi\}$ is differentiable in ϵ around $\epsilon = 0$ so that its first order expansion w.r.t. ϵ obeys $\mathscr{N}\{h\} = \mathscr{N}\{h^* + \epsilon\phi\} = \mathscr{N}\{h^*\} + \epsilon(\frac{d}{d\epsilon}\mathscr{N}\{h^* + \epsilon\phi\})_{|\epsilon=0} + o(\epsilon)$. The "best affine approximation" to $\mathscr{N}\{h\}$ around the fixed function h^* is defined as discarding the term $o(\epsilon)$ from the above, so that,

$$c_{opt} \triangleq \mathscr{N}\{h^*\} \quad , \quad \mathcal{L}_{opt}\{h\} \triangleq \epsilon(\frac{d}{d\epsilon}\mathscr{N}\{h^* + \epsilon\phi\})_{|\epsilon=0}\,. \tag{5}$$

Theorem 1. *The best affine approximation of the modified Vese's operator, acting on functions close to the zero function ($h^*(\boldsymbol{x}) = 0$) and with bounded zeroth, first and second order derivatives, is equal to $\frac{1}{n+1}\,\Delta$.*

Proof. The nonlinear operator of interest here is the modified Vese's operator,

$$\mathscr{N}\{h\} \triangleq \sqrt{1 + \|\nabla h\|^2}\, m\big(\boldsymbol{\lambda}(\nabla^2 h)\big)\,. \tag{6}$$

Observe that for this operator, $\mathscr{N}\{h^* + \epsilon\phi\}$ is differentiable in ϵ. Since $h^*(\boldsymbol{x}) = 0$, (5) implies that $c_{\text{opt}} = \mathscr{N}(0) = 0$. Note that we exploited the fact that ϕ, $\nabla\phi$ and $\boldsymbol{\lambda}(\nabla^2\phi)$ are bounded at any $\boldsymbol{x} \in \mathcal{X}$ so that by $\epsilon = 0$ one can conclude $h = \epsilon\phi = 0$, $\|\nabla h\|^2 = \epsilon^2\|\nabla\phi\|^2 = 0$, and $\boldsymbol{\lambda}(\nabla^2 h) = \epsilon\boldsymbol{\lambda}(\nabla^2\phi) = 0$.

We now focus on computing $\mathcal{L}_{\text{opt}}\{h\}$ using (5), which amounts to finding,

$$\epsilon\,\big(\frac{d}{d\epsilon}\sqrt{1 + \|\nabla\epsilon\phi\|^2}\, m(\boldsymbol{\lambda}(\nabla^2\epsilon\phi))\big)_{|\epsilon=0}\,. \tag{7}$$

We proceed by first computing $\big(\frac{d}{d\epsilon}\sqrt{1 + \epsilon^2\|\nabla\phi\|^2}\, m(\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\epsilon=0}$. By chain rule, this is equivalent to,

$$\begin{aligned}&\big(\frac{d}{d\epsilon}\sqrt{1 + \epsilon^2\|\nabla\phi\|^2}\big)_{|\epsilon=0}\,\big(m(\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\epsilon=0}\\ +\,&\big(\frac{d}{d\epsilon}m(\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\epsilon=0}\,\big(\sqrt{1 + \epsilon^2\|\nabla\phi\|^2}\big)_{|\epsilon=0}\,.\end{aligned}$$

Since $\nabla\phi$ and $\boldsymbol{\lambda}(\nabla^2\phi)$ are assumed to be bounded, at $\epsilon = 0$, the above expression can be written as

$$\big(\frac{d}{d\epsilon}\sqrt{1 + \epsilon^2\|\nabla\phi\|^2}\big)_{|\epsilon=0}\, m(\mathbf{0}) + \big(\frac{d}{d\epsilon}m(\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\epsilon=0}\,\sqrt{1+0}\,. \tag{8}$$

Hence the above sum simplifies to $\big(\frac{d}{d\epsilon}m(\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\epsilon=0}$. Applying chain rule again, this becomes $\big(\nabla m(\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\,\epsilon=0} \bullet \big(\frac{d}{d\epsilon}\epsilon\,\boldsymbol{\lambda}(\nabla^2\phi)\big)_{|\,\epsilon=0}$, where $\bullet$ denotes the inner product between two $n \times 1$ vectors. Evaluating it at $\epsilon = 0$ yields

$\nabla m(\mathbf{0}) \bullet \boldsymbol{\lambda}(\nabla^2 \phi)$. Since $\nabla m(\mathbf{0}) = \frac{1}{n+1}\mathbf{1}$, where $\mathbf{1}$ is a $n \times 1$ vector with all entries equal to 1, the expression becomes $\frac{1}{n+1}\,\mathbf{1} \bullet \boldsymbol{\lambda}(\nabla^2 \phi)$. However, $\mathbf{1} \bullet \boldsymbol{\lambda}(\nabla^2 \phi)$ is simply the sum of the eigenvalues, thus it is Trace$(\nabla^2 \phi)$. Finally, since Trace$(\nabla^2 \phi)$ is sum of the diagonals of the Hessian matrix for ϕ, it is equivalent to the Laplacian $\Delta\phi$. In summary, we just derived that,

$$\big(\frac{d}{d\epsilon}\sqrt{1+\epsilon^2\|\nabla\phi\|^2}\, m(\epsilon\, \boldsymbol{\lambda}(\nabla^2\phi))\big)_{|\epsilon=0} = \frac{1}{n+1}\Delta\phi\,, \tag{9}$$

Going back to the definition of $\mathcal{L}_{\text{opt}}\{h\}$ in (7), it follows that,

$$\mathcal{L}_{\text{opt}}\{h\} \triangleq \epsilon\,\big(\frac{d}{d\epsilon}\sqrt{1+\|\nabla\epsilon\phi\|^2}\, m(\boldsymbol{\lambda}(\nabla^2\epsilon\phi))\big)_{|\epsilon=0} \tag{10}$$

$$= \epsilon\,\frac{1}{n+1}\,\Delta\phi\,. \tag{11}$$

We now manipulate $\epsilon\,\frac{1}{n+1}\,\Delta\phi$. Moving ϵ inside, it can be equivalently be written as $\frac{1}{n+1}\,\Delta(\epsilon\,\phi)$. However, by (4), $\epsilon\,\phi$ is just the definition of $h - h^*$. Using that fact that $h^* = 0$, we obtain,

$$\mathcal{L}_{\text{opt}}\{h\} = \frac{1}{n+1}\,\Delta h\,. \tag{12}$$

□

6 Discussion and Future Works

This work provided new insights into the optimization by homotopy continuation. We showed that constructing the homotopy by Gaussian convolution is optimal in a specific sense. That is, the Gaussian homotopy is the result of the *best affine approximation* to the modified Vese's PDE. Vese's PDE is interesting for homotopy construction because it evolves the nonconvex function to its convex envelope. The convex envelope provides optimal convexification for nonconvex functions. However, Vese's PDE does not have any closed form solution due to its nonlinearity, hence cannot be used in practice. In contrast, Gaussian smoothing can be computed in closed form for a large family of functions, including those represented by polynomials or Gaussian radial basis functions.

Recall that the optimality of the Gaussian homotopy is proved here in a certain setting; when the modified Vese's PDE is *linearized* around the *origin* of the function space $h^*(\boldsymbol{x}) = 0$. Such linearization severely degrades the fidelity of the approximation. An important question is whether linearity or working around the origin could be relaxed without losing the advantage of closed form solution to the PDE. Such exntension is a clear direction for future studies.

A possibility might be exploiting the conditional diffusion property of Vese's PDE. Remember this PDE only diffuses nonconvex regions throughout the evolution, and is insensitive to convex regions. If the nonconvex and convex parts of an objective function could be separated, applying Gaussian smoothing only

to the nonconvex part might produce a better approximation to Vese's PDE, as opposed to smoothing the entire objective function. This is obviously a nonlinear evolution because it requires a switching behavior between convex and nonconvex regions.

Another direction for improving the approximation quality is to manipulate the objective function. For example, transforming the objective function $f(\boldsymbol{x})$ to $-\exp(-M\, f(\boldsymbol{x}))$, where $M > 0$ is a large constant, does not alter the global minimizers. However, the latter form may lead to a better agreement between the linearized and original PDE, when used as their initial condition. The intuition is that, the transformed function is very close to zero almost everywhere (recall that our linearization is around $h^*(\boldsymbol{x}) = 0$). Smoothing the exponentially transformed function is also pursued by [23], but for one-shot convexification. Note that the exponential transform followed by the diffusion process is related to the *Burgers'* PDE [11]. This connection might be of value, but does not completely answer all questions. That is because while the solution of Burgers' equation has a known form, it involves Gaussian convolution of $\exp(-M\, f(\boldsymbol{x}))$, which may not have an analytical form for interesting choices of $f(\boldsymbol{x})$, e.g. polynomials. This integral also arises in [23] and is approximated by sampling based methods.

Acknowledgment. This work is partially funded by the Shell Research. First author is thankful to Alan L. Yuille (UCLA) and Steven G. Johnson (MIT) and Vadim Zharnitsky (UIUC) for discussions, and grateful to William T. Freeman (MIT) for supporting this work.

References

1. Balzer, J., Mörwald, T.: Isogeometric finite-elements methods and variational reconstruction tasks in vision - a perfect match. In: CVPR (2012)
2. Barron, J.: Shapes, Paint, and Light. Ph.D. thesis, EECS Department, University of California, Berkeley (August 2013)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
4. Bengio, Y.: Learning Deep Architectures for AI. Now Publishers Inc. (2009)
5. Black, M.J., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. International Journal of Computer Vision 19(1), 57–91 (1996)
6. Blake, A., Zisserman, A.: Visual Reconstruction. MIT Press (1987)
7. Blake, A.: Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. IEEE PAMI 11(1), 2–12 (1989)
8. Boccuto, A., Discepoli, M., Gerace, I., Pucci, P.: A gnc algorithm for deblurring images with interacting discontinuities (2002)
9. Brox, T.: From pixels to regions: partial differential equations in image analysis. Ph.D. thesis, Saarland University, Germany (April 2005)
10. Brox, T., Malik, J.: Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. 33(3), 500–513 (2011)

11. Burgers, J.M.: The nonlinear diffusion equation: asymptotic solutions and statistical problems. D. Reidel Pub. Co. (1974)
12. Chapelle, O., Chi, M., Zien, A.: A continuation method for semi-supervised svms. pp. 185–192. ICML 2006 (2006)
13. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. J. Mach. Learn. Res. 9, 203–233 (2008)
14. Chapelle, O., Wu, M.: Gradient descent optimization of smoothed information retrieval metrics. Inf. Retr. 13(3), 216–235 (2010)
15. Chaudhuri, S., Clochard, M., Solar-Lezama, A.: Bridging boolean and quantitative synthesis using smoothed proof search. SIGPLAN Not 49(1), 207–220 (2014)
16. Chaudhuri, S., Solar-Lezama, A.: Smooth interpretation. In: PLDI. pp. 279–291. ACM (2010)
17. Chaudhuri, S., Solar-Lezama, A.: Smoothing a program soundly and robustly. In: Gopalakrishnan, G., Qadeer, S. (eds.) CAV 2011. LNCS, vol. 6806, pp. 277–292. Springer, Heidelberg (2011)
18. Cohen, L.D., Gorre, A.: Snakes: Sur la convexite de la fonctionnelle denergie (1995)
19. Coupé, P., Manjón, J.V., Chamberland, M., Descoteaux, M., Hiba, B.: Collaborative patch-based super-resolution for diffusion-weighted images. NeuroImage 83, 245–261 (2013)
20. Dai, Z., Lücke, J.: Unsupervised learning of translation invariant occlusive components. In: CVPR, pp. 2400–2407 (2012)
21. Dhillon, P.S., Keerthi, S.S., Bellare, K., Chapelle, O., Sundararajan, S.: Deterministic annealing for semi-supervised structured output learning. In: AISTATS 2012, vol. 15 (2012)
22. Dufour, R.M., Miller, E.L., Galatsanos, N.P.: Template matching based object recognition with unknown geometric parameters. IEEE Transactions on Image Processing 11(12), 1385–1396 (2002)
23. Dvijotham, K., Fazel, M., Todorov, E.: Universal convexification via risk-aversion. In: UAI (2014)
24. Erhan, D., Manzagol, P.A., Bengio, Y., Bengio, S., Vincent, P.: The difficulty of training deep architectures and the effect of unsupervised pre-training. In: AISTATS, pp. 153–160 (2009)
25. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010)
26. Frank, M., Streich, A.P., Basin, D., Buhmann, J.M.: Multi-assignment clustering for boolean data. J. Mach. Learn. Res. 13(1), 459–489 (2012)
27. Fua, P., Leclerc, Y.: Object-centered surface reconstruction: combining multi-image stereo shading. International Journal on Computer Vision 16(1), 35–56 (1995)
28. Gehler, P., Chapelle, O.: Deterministic annealing for multiple-instance learning. In: AISTATS 2007, pp. 123–130. Microtome, Brookline (2007)
29. Geiger, D., Girosi, F.: Coupled markov random fields and mean field theory. In: NIPS, pp. 660–667. Morgan Kaufmann (1989)
30. Geiger, D., Yuille, A.L.: A common framework for image segmentation. International Journal of Computer Vision 6(3), 227–243 (1991)
31. Gold, S., Rangarajan, A.: A graduated assignment algorithm for graph matching. IEEE PAMI 18, 377–388 (1996)
32. Gold, S., Rangarajan, A., Mjolsness, E.: Learning with preknowledge: Clustering with point and graph matching distance measures. In: NIPS, pp. 713–720 (1994)
33. Held, D., Levinson, J., Thrun, S., Savarese, S.: Combining 3d shape, color, and motion for robust anytime tracking. In: RSS, Berkeley, USA (July 2014)

34. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief networks. Neural Computation 18(7), 1527–1554 (2006)
35. Hong, B.W., Lu, Z., Sundaramoorthi, G.: A new model and simple algorithms for multi-label mumford-shah problems. In: CVPR (June 2013)
36. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: CVPR, pp. 2307–2314. IEEE (2013)
37. Kim, M., Torre, F.D.: Gaussian processes multiple instance learning, pp. 535–542 (2010)
38. Kosowsky, J.J., Yuille, A.L.: The invisible hand algorithm: Solving the assignment problem with statistical physics. Neural Networks 7(3), 477–490 (1994)
39. Leich, A., Junghans, M., Jentschel, H.J.: Hough transform with GNC. 12th European Signal Processing Conference (EUSIPCO, 2004)
40. Leordeanu, M., Hebert, M.: Smoothing-based optimization. In: CVPR (2008)
41. Li, X.: Fine-granularity and spatially-adaptive regularization for projection-based image deblurring. IEEE Transactions on Image Processing 20(4), 971–983 (2011)
42. Liu, Z., Qiao, H., Xu, L.: An extended path following algorithm for graph-matching problem. IEEE Trans. Pattern Anal. Mach. Intell. 34(7), 1451–1456 (2012)
43. Loog, M., Duistermaat, J.J., Florack, L.M.J.: On the behavior of spatial critical points under gaussian blurring (A folklore theorem and scale-space constraints). In: Kerckhove, M. (ed.) Scale-Space 2001. LNCS, vol. 2106, pp. 183–192. Springer, Heidelberg (2001)
44. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: ICML, vol. 382, p. 87. ACM (2009)
45. Malek-Mohammadi, M., Babaie-Zadeh, M., Amini, A., Jutten, C.: Recovery of low-rank matrices under affine constraints via a smoothed rank function. IEEE Transactions on Signal Processing 62(4), 981–992 (2014)
46. Mobahi, H., Rao, S., Ma, Y.: Data-driven image completion by image patch subspaces. In: Picture Coding Symposium (2009)
47. Mobahi, H., Ma, Y., Zitnick, L.: Seeing through the Blur. In: CVPR (2012)
48. Mohimani, G.H., Babaie-Zadeh, M., Gorodnitsky, I., Jutten, C.: Sparse recovery using smoothed ℓ^0 (sl0): Convergence analysis. CoRR abs/1001.5073 (2010)
49. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. 42(5), 577–685 (1989)
50. Nielsen, M.: Graduated non-convexity by smoothness focusing. In: Proceedings of the British Machine Vision Conference, pp. 60.1–60.10. BMVA Press (1993)
51. Nikolova, M., Ng, M.K., Tam, C.P.: Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. Trans. Img. Proc. 19(12), 3073–3088 (2010)
52. Pretto, A., Soatto, S., Menegatti, E.: Scalable dense large-scale mapping and navigation. In: Proc. of: Workshop on Omnidirectional Robot Vision, ICRA (2010)
53. Price, B.L., Morse, B.S., Cohen, S.: Simultaneous foreground, background, and alpha estimation for image matting. In: CVPR, pp. 2157–2164. IEEE (2010)
54. Rangarajan, A., Chellappa, R.: Generalized graduated nonconvexity algorithm for maximum a posteriori image estimation, pp. II:127–II:133 (1990)
55. Rose, K., Gurewitz, E., Fox, G.: A deterministic annealing approach to clustering. Pattern Recognition Letters 11(9), 589–594 (1990)
56. Rossi, F., Villa-Vialaneix, N.: Optimizing an organized modularity measure for topographic graph clustering: A deterministic annealing approach. Neurocomputing 73(7-9), 1142–1163 (2010)
57. Saragih, J.: Deformable face alignment via local measurements and global constraints, pp. 187–207. Springer, Heidelberg (2013)

58. Shroff, N., Turaga, P.K., Chellappa, R.: Manifold precis: An annealing technique for diverse sampling of manifolds. In: NIPS, pp. 154–162 (2011)
59. Sindhwani, V., Keerthi, S.S., Chapelle, O.: Deterministic annealing for semi-supervised kernel machines. In: ICML 2006, pp. 841–848. ACM, New York (2006)
60. Smith, N.A., Eisner, J.: Annealing techniques for unsupervised statistical language learning. In: ACL, Barcelona, Spain, pp. 486–493 (July 2004)
61. Stoll, M., Volz, S., Bruhn, A.: Joint trilateral filtering for multiframe optical flow. In: ICIP, pp. 3845–3849 (2013)
62. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: CVPR, pp. 2432–2439. IEEE (2010)
63. Terzopoulos, D.: The computation of visible-surface representations. IEEE Trans. Pattern Anal. Mach. Intell. 10(4), 417–438 (1988)
64. Tirthapura, S., Sharvit, D., Klein, P., Kimia, B.: Indexing based on edit-distance matching of shape graphs. In: SPIE International Symposium on Voice, Video, and Data Communications, pp. 25–36 (1998)
65. Trzasko, J., Manduca, A.: Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0 -minimization. IEEE Trans. Med. Imaging 28(1), 106–121 (2009)
66. Vese, L.: A method to convexify functions via curve evolution. Commun. Partial Differ. Equations 24(9-10), 1573–1591 (1999)
67. Vural, E., Frossard, P.: Analysis of descent-based image registration. SIAM J. Imaging Sciences 6(4), 2310–2349 (2013)
68. Widder, D.V.: The Heat Equation. Academic Press (1975)
69. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31(2), 210–227 (2009)
70. Wu, Z., Tan, P.: Calibrating photometric stereo by holistic reflectance symmetry analysis, pp. 1498–1505. IEEE (2013)
71. Wu, Z.: The Effective Energy Transformation Scheme as a Special Continuation Approach to Global Optimization with Application to Molecular Conformation. SIAM J. on Optimization 6, 748–768 (1996)
72. Yuille, A.: Energy Functions for Early Vision and Analog Networks. A.I. memo, Defense Technical Information Center (1987)
73. Yuille, A.L.: Generalized deformable models, statistical physics, and matching problems. Neural Computation 2, 1–24 (1990)
74. Yuille, A., Geiger, D., Bulthoff, H.: Stereo integration, mean field theory and psychophysics. In: Faugeras, O. (ed.) ECCV 1990. LNCS, vol. 427, pp. 71–82. Springer, Heidelberg (1990)
75. Yuille, A.L., Peterson, C., Honda, K.: Deformable templates, robust statistics, and hough transforms, San Diego, CA, pp. 166–174. International Society for Optics and Photonics (1991)
76. Yuille, A.L., Stolorz, P.E., Utans, J.: Statistical physics, mixtures of distributions, and the em algorithm. Neural Computation 6(2), 334–340 (1994)
77. Zaslavskiy, M., Bach, F., Vert, J.P.: A path following algorithm for the graph matching problem. IEEE PAMI (2009)
78. Zerubia, J., Chellappa, R.: Mean field annealing using compound gauss-markov random fields for edge detection and image estimation. IEEE Transactions on Neural Networks 4(4), 703–709 (1993)