Combinatorial Optimization in Computer Vision (IN2245)

Frank R. Schmidt
Csaba Domokos

Winter Semester 2015/2016

20. Fast Trust Region

Trust Region

Continuous Functions

The **trust region framework** is a popular framework for **continuous energy minimization**. After revising it, we will see how it can be adapted for **combinatorial optimization**.

In the continuous setup, we assume that a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$ is given, *i.e.*, f and all its derivatives are continuous and themselves differentiable.

The general derivatives turn out to be high-order tensors. For that reason, we will focus just on the 1^{st} and 2^{nd} order derivatives.

The **gradient** ∇f and the **Hesse matrix** Hf (cf. **Analysis I/II**) are mappings

$$\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n \qquad\qquad Hf\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}.$$

Taylor Approximations

Given a specific point $x_0 \in \mathbb{R}^n$, we can define the Taylor sequence (cf. **Analysis I / II**). Of practical use are the 1st and 2nd order **Taylor approximations**

$$\begin{aligned} T^1_{x_0} f(x) &= f(x_0) + \langle \nabla f(x_0), x - x_0\rangle \\ T^2_{x_0} f(x) &= f(x_0) + \langle \nabla f(x_0), x - x_0\rangle \\ &\quad + \frac{1}{2}\langle Hf(x_0)\cdot(x-x_0), x - x_0\rangle\,. \end{aligned}$$

Since the minimization of linear and quadratic functions is easy to perform, most local minimization approaches focus on the linear Taylor approximation $T^1 f$ or the quadratic Taylor approximation $T^2 f$.

Approximation Reliability

All Taylor approximations are accurate for $x = x_0$, but the further we move away from this **development point** x_0, the less reliable the approximation becomes.

Thus, we want to define a **region** around x_0 in which we **trust** a specific approximation. This region can be defined by a single parameter, its distance $d \in \mathbb{R}^+$ from the development point x_0, *i.e.*,

$$B_d(x_0) := \{x \in \mathbb{R}^n \,|\, \|x - x_0\| \leqslant d\}$$

During the **trust region optimization**, we will adapt this distance d. It is therefore not a global parameter, but a parameter that changes due to the local behavior of the function f.

Trust Region Method

Let us assume that we have an energy $f\colon \mathbb{R}^n \to \mathbb{R}$ and for each $x_0 \in \mathbb{R}^n$ an approximative energy $\tilde{f}_{x_0}\colon \mathbb{R}^n \to \mathbb{R}$ such that $\tilde{f}_{x_0}(x_0) = f(x_0)$. Then, the trust region method works as follows:

1. Let $x_0 \in \mathbb{R}^n$ be an arbitrary initial solution and $d > 0$ be an arbitrary initial distance that defines the trust region $B_d(x_0)$ around x_0.
2. Let x^* be a global optimum of the **trust region subproblem**:
$$\min_{x \in B_d(x_0)} \tilde{f}_{x_0}(x)$$
3. Set $\rho := \frac{f(x_0) - f(x^*)}{\tilde{f}_{x_0}(x_0) - \tilde{f}_{x_0}(x^*)}$.
4. If $\rho > \tau_1$, set $x_0 := x^*$.
5. If $\rho > \tau_2$, set $d := d \cdot \alpha$, otherwise $d := d/\alpha$.
6. If not converged, go to Step 2.

Trust Region Parameters

There are three different design parameters for the trust region method, namely $0 < \tau_1 < \tau_2 < 1$ and $\alpha > 1$.

The α-parmeter describes how the trust region changes during the method. It is common to use $\alpha = 2$.

The τ_1-parmeter controls when to accept x^* as a good candidate. In the continuous framework, $\tau_1 > 0$ has to be chosen in order to guarantee convergence. In the combinatorial framework, we can also use $\tau_1 = 0$, but this may result in a higher running time.

The τ_2-parmeter controls when to expand or shrink the trust region.
A common choice is $\tau_2 = 0.25$.

First Order Trust Region

If we use $T^1_{x_0} f$ as the approximative energy $\tilde{f}_{x_0}$, the trust region subproblem

$$\min_{\|x-x_0\|\leqslant d} f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$$

is uniquely minimized by

$$x^* = x_0 - d \cdot \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}$$

In the case of the first order Taylor approximation, the trust region method is a **gradient descent** variant.

It is called the **normalized gradient descent method**.

Second Order Trust Region

If we use $T^2_{x_0} f$ as the approximative energy $\tilde{f}_{x_0}$, the trust region subproblem

$$\min_{\|x-x_0\|\leqslant d} f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2} \langle Hf(x_0)(x - x_0), x - x_0 \rangle$$

is uniquely minimized by $x^* \in B_d(x_0)$. This method is called the **trust region Newton method**.

For big d, we obtain

$$x^* = x_0 - Hf(x_0)^{-1}(\nabla f(x_0)),$$

which is the **Newton iteration**.

Summary and Outlook: Trust Region

The trust region framework can be seen as a generalization of different, continuous optimization techniques.

In particular, it includes the *normalized gradient descent approach* and the *trust region Newton method*.

Nonetheless, it is not restricted to continuous energy minimization. In fact, it can be easily extended to arbitrary combinatorial problems.

In order to do this, we have to make specific decisons: Firstly, we have to define the approximation $\tilde{E}$ of function E given a specific discrete solution S_0.

Secondly, we have to define a trust region with respect to S_0.

Thirdly, we have to find the global optimum of $\tilde{E}$ within the trust region.

Shape Distances

Trust Region of Segmentations

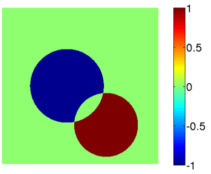
Given a specific segmentation $S_0 \subset \Omega$, we want to define a region $B_d(S_0) \subset \mathcal{P}(\Omega)$ of segmentations for which we like to trust an approximation $\tilde{E}$ of an energy E.

This can be done by introducing a positive definite distance function $\text{dist}\colon \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}^+_0$ on the set $\mathcal{P}(\Omega)$ of possible segmentations.

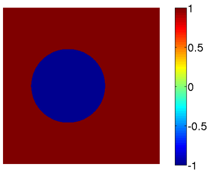
A commonly used distances between sets is the **Hamming distance**:

$$\begin{aligned}\text{dist}_H(S,T) =& \text{area}(S \triangle T)\\ =& \int_{S\setminus T} 1\text{dx} + \int_{T\setminus S} 1\text{dx}\\ =& \int_{S-(S\cap T)} 1\text{dx} + \int_{T-(S\cap T)} 1\text{dx}\end{aligned}$$

Hamming Distance



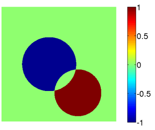
Hamming



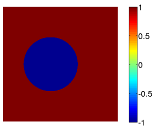
Indicator ϕ_S

$$\begin{aligned}\text{dist}_H(S,T) &= \int_{T-(S\cap T)} 1\text{dx} + \int_{S-(S\cap T)} 1\text{dx}\\ &= \int_{T-(S\cap T)} \phi_S\text{dx} - \int_{S-(S\cap T)} \phi_S\text{dx}\\ &= \int_T \phi_S\text{dx} - \int_S \phi_S\text{dx}\end{aligned}$$

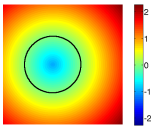
Modular Distances



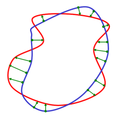
Hamming



Indicator ϕ_S



Signed Distance



Shape Distance

The Hamming distance $\text{dist}_H(S,T)$ is modular in T and can therefore be easily optimized. Similarily, we can define different modular distance function like dist_2 that is driven by the signed distance function.

$$\begin{aligned}\text{dist}_H(S,T) &= \int_T \phi_S(x)\text{dx} - \int_S \phi_S(x)\text{dx}\\ \text{dist}_2(S,T) &= \int_T \text{sdf}_{S_0}(x)\text{dx} - \int_S \text{sdf}_S(x)\text{dx}\end{aligned}$$

Quadratic Pseudo-Boolean Optimization

Local Submodular Approx. of QPBO

Given a quadratic pseudo-Boolean funtion $E\colon \mathbb{B}^n \to \mathbb{R}$

$$E(x) = C + \sum_{i=1}^{n} f_i x_i + \sum_{\substack{i,j=1\\ f_{i,j}<0}}^{n} f_{i,j} x_i x_j + \sum_{\substack{i,j=1\\ f_{i,j}>0}}^{n} f_{i,j} x_i x_j$$

and an initial segmentation $x^{(0)} \in \mathbb{B}^n$, we define the following submodular approximation of E:

$$\tilde{E}(x) = C + \sum_{i=1}^{n} f_i x_i + \sum_{\substack{i,j=1\\ f_{i,j}<0}}^{n} f_{i,j} x_i x_j + \sum_{\substack{i,j=1\\ f_{i,j}>0}}^{n} f_{i,j} x_i^{(0)} x_j + f_{i,j} x_i x_j^{(0)}$$

This approximation only computes the first order Taylor approximation for the supermodular term. It is therefore a submodular (not a linear) approximation.

Combinatorial Constrained Optimization

Given the submodular energy $\tilde{E}$, we like to solve

$$\min_{\mathrm{dist}(x_0,x)\leqslant d} \tilde{E}(x). \tag{1}$$

We can relax this problem by minimizing

$$\tilde{E}(x) + \lambda \cdot \mathrm{dist}(x_0, x) \tag{2}$$

instead.

Note that $\lambda = 0$ ($\lambda = \infty$) in (2) leads to the same solution as $d = \infty$ ($\lambda = 0$) in (1). Thus, we can assume a relationship of $d \approx \frac{1}{\lambda}$.

This leads to a different method, that we called **Fast Trust Region**.

Fast Trust Region Method

Let us assume that we have an energy $E\colon \mathbb{B}^n \to \mathbb{R}$ and for each $x_0 \in \mathbb{B}^n$ an approximative energy $\tilde{E}_{x_0}\colon \mathbb{B}^n \to \mathbb{R}$ such that $\tilde{E}_{x_0}(x_0) = E(x_0)$. Then, the fast trust region method works as follows:

1. Let $x_0 \in \mathbb{B}^n$ be an arbitrary initial solution and $\lambda > 0$.
2. Let x^* be a global optimum of the **fast trust region subproblem**:

$$\min_{x\in\mathbb{B}^n} \tilde{E}_{x_0}(x) + \lambda \cdot \mathrm{dist}(x_0, x)$$

3. Set $\rho := \frac{E(x_0)-E(x^*)}{\tilde{E}_{x_0}(x_0)-\tilde{E}_{x_0}(x^*)}$.
4. If $\rho > \tau_1$, set $x_0 := x^*$.
5. If $\rho > \tau_2$, set $\lambda := \lambda/\alpha$, otherwise $\lambda := \lambda \cdot \alpha$.
6. If not converged, go to Step 2.

In practice, we choose $\tau_1 = 0$, $\tau_2 = 0.25$ and $\alpha = 2$.

Segmentation with Repulsion

The classical image segmentation framework combines a **data term** with a **length term**

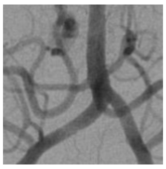
$$E(x) = \sum_{i=1}^{n} f_i x_i + \sum_{i=1}^{n} \sum_{j\in\mathcal{N}(i)} f_{ij} x_i x_j \qquad f_{ij} \leqslant 0.$$

The length term can also be seen as an **attraction force**, since label clusters attract each other. In other words, it is cheaper to combine two clusters into one cluster if they are close to one another.

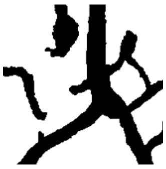
If we want to model a **repulsion force**, we need to use supermodular terms:

$$f_{ij} = \frac{|I(i) - I(j)| - c}{\mathrm{dist}(i,j)}$$

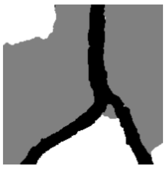
Image Segmentation with Repulsion



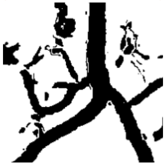
Image



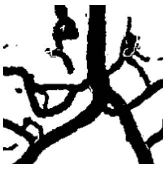
Graph Cut (without Repulsion)



QPBO



TRWS



FTR

Regional Functions

Outlook: Regional Functions

In the next lecture we will address a special class of **higher-order pseudo-Boolean energies**

$$E(S) = E_0(S) + R^F_{\{f_i\}_{i<k}}(S)$$

- E_0 is a submodular function
- $R^F_{\{f_i\}_{i<k}}$ is a **regional function**, *i.e.*,

$$R^F_{\{f_i\}_{i<k}}(S) = F\left(\langle f_0, S\rangle, \ldots, \langle f_{k-1}, S\rangle\right)$$

where

$$f_i : \Omega \to \mathbb{R} \qquad \text{“indicator” function}$$
$$F : \mathbb{R}^k \to \mathbb{R} \qquad \text{smooth composition}$$

Literature *

Trust Region

- Levenberg, *A Method for the Solution of Certain Non-Linear Problems in Least Squares*, 1944, Quarterly of Applied Mathematics 2, 164–168.
- Marquardt, *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, 1963, SIAM J. on Applied Mathematics 11(2), 431–441.
- Sorenson, *Newton's Method with a Model Trust Region Modification*, 1987, SIAM J. Numerical Analysis 24, 1152–1170.

Fast Trust Region in Computer Vision

- Gorelick, Schmidt, Boykov, Delong, Ward. *Segmentation with non-linear regional constraints via line-search cuts*, 2012, ECCV, 583–597.
- Gorelick, Schmidt, Boykov. *Fast Trust Region for Segmentation*, 2013, IEEE CVPR.
- Gorelick, Boykov, Veksler, Ben Ayed, Delong. *Submodularization for Binary Pairwise Energies*, 2014, IEEE CVPR.