PGMO Lecture: Vision, Learning and Optimization

6. Non-convex optimization

Thomas Pock

Institute of Computer Graphics and Vision, TU Graz

February 19, 2020

Overview

Non-convex optimization

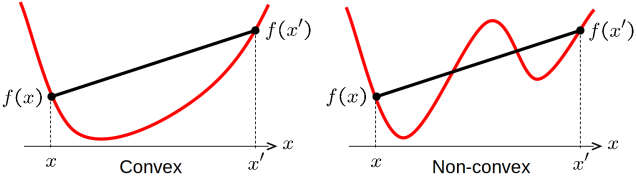
Non-convex proximal gradient method

Non-convex accelerated proximal gradient method

Applications

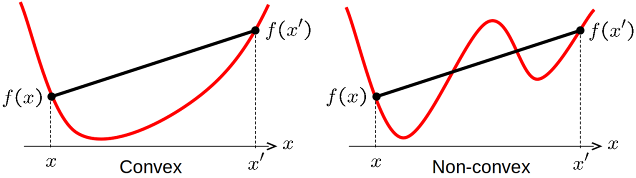
Proximal alternating linearization method

Convex versus non-convex



"The great watershed in optimization is not between linearity and non-linearity, but convexity and non-convexity." R. Rockafellar, 1993

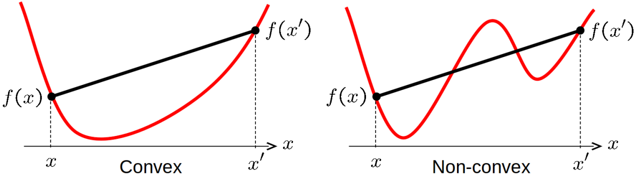
Convex versus non-convex



"The great watershed in optimization is not between linearity and non-linearity, but convexity and non-convexity." *R. Rockafellar, 1993*

- Convex problems
  - Any local minimizer is a global minimizer
  - Result is independent of the initialization
  - Convex models often inferior

Convex versus non-convex



"The great watershed in optimization is not between linearity and non-linearity, but convexity and non-convexity." R. Rockafellar, 1993

- Convex problems
  - Any local minimizer is a global minimizer
  - Result is independent of the initialization
  - Convex models often inferior
- Non-convex problems
  - In general no chance to find the global minimizer
  - Result strongly depends on the initialization
  - Often gives more accurate models (prior modeling)

Non-convex optimization problems

- Smooth non-convex problems can be solved via generic nonlinear numerical optimization algorithms (SD, CG, BFGS, Newton, ...)
- Hard to generalize to constraints, or non-differentiable functions
- Line-search procedure can be time intensive

Non-convex optimization problems

- Smooth non-convex problems can be solved via generic nonlinear numerical optimization algorithms (SD, CG, BFGS, Newton, ...)
- Hard to generalize to constraints, or non-differentiable functions
- Line-search procedure can be time intensive

- A reasonable idea is to develop algorithms for special classes of structured non-convex problems
- A promising class of problems that has a moderate degree of non-convexity is given by the sum of a smooth non-convex function and a non-smooth convex function [Sra '12], [Chouzenoux, Pesquet, Repetti '13]

Problem definition

- We consider the problem of minimizing a function $F\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$

$$\min_{x \in \mathcal{X}} F(x) := f(x) + g(x)\,,$$

  where $\mathcal{X}$ is a finite dimensional real vector space.
- We assume that F is coercive, i.e. $\|x\|_2 \to +\infty \quad \Rightarrow \quad F(x) \to +\infty$ and bounded from below by some value $\underline{F} > -\infty$

Problem definition

- We consider the problem of minimizing a function $F\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$

$$\min_{x \in \mathcal{X}} F(x) := f(x) + g(x)\,,$$

  where $\mathcal{X}$ is a finite dimensional real vector space.
- We assume that F is coercive, i.e. $\|x\|_2 \to +\infty \quad \Rightarrow \quad F(x) \to +\infty$ and bounded from below by some value $\underline{F} > -\infty$
- The function f is possibly non-convex but has a Lipschitz continuous gradient, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\,\|x - y\|_2$$

Problem definition

- We consider the problem of minimizing a function $F\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$

$$\min_{x \in \mathcal{X}} F(x) := f(x) + g(x)\,,$$

where $\mathcal{X}$ is a finite dimensional real vector space.

- We assume that F is coercive, i.e. $\|x\|_2 \to +\infty \quad \Rightarrow \quad F(x) \to +\infty$ and bounded from below by some value $\underline{F} > -\infty$

- The function f is possibly non-convex but has a Lipschitz continuous gradient, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \|x - y\|_2$$

- The function g is a proper lower semi-continuous convex function with an efficient to compute proximal map

$$\operatorname{prox}_{\tau g}(\bar{x}) := \arg\min_{x \in \mathcal{X}} \frac{\|x - \bar{x}\|_2^2}{2} + \tau g(x)\,,$$

where $\tau > 0$.

Overview

Non-convex optimization

Non-convex proximal gradient method

Non-convex accelerated proximal gradient method

Applications

Proximal alternating linearization method

Proximal gradient method

- We aim at seeking a critical point x^*, i.e. a point satisfying $0 \in \partial F(x^*)$ which in our case becomes
$$-\nabla f(x^*) \in \partial g(x^*) \,.$$
- A critical point can also be characterized via the *proximal residual*
$$r(x) := x - \mathrm{prox}_{\tau g}(x - \tau \nabla f(x)) \,,$$
where I is the identity map.
- Clearly $r(x^*) = 0$ implies that x^* is a critical point.
- The norm of the proximal residual can be used as a (bad) measure of optimality

Proximal gradient method

- We aim at seeking a critical point x^*, i.e. a point satisfying $0 \in \partial F(x^*)$ which in our case becomes
$$-\nabla f(x^*) \in \partial g(x^*)\,.$$
- A critical point can also be characterized via the *proximal residual*
$$r(x) := x - \mathrm{prox}_{\tau g}(x - \tau \nabla f(x))\,,$$
where I is the identity map.
- Clearly $r(x^*) = 0$ implies that x^* is a critical point.
- The norm of the proximal residual can be used as a (bad) measure of optimality
- The proximal residual already suggests an iterative method of the form
$$x^{k+1} = \mathrm{prox}_{\tau g}(x^k - \tau \nabla f(x^k))$$
- For f convex, this algorithm is well studied [Lions, Mercier '79], [Tseng '91], [Daubechie et al. '04], [Combettes, Wajs '05], [Raguet, Fadili, Peyré '13]

Basic descent rule

- We can derive a basic descent rule by noting that the proximal gradient step

$$x \mapsto g(x) + f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2\tau} \|x - \bar{x}\|^2,$$

  is $1/\tau$ strongly convex, that is for the unique minimizer $\hat{x}$ of the proximal map one has for all x

$$g(x) + f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \tfrac{1}{2\tau} \|x - \bar{x}\|^2 \geq$$
$$g(\hat{x}) + \underbrace{f(\bar{x}) + \langle \nabla f(\bar{x}), \hat{x} - \bar{x} \rangle}_{\geq f(\hat{x}) - \frac{L}{2}\|\hat{x} - \bar{x}\|^2} + \tfrac{1}{2\tau} \|\hat{x} - \bar{x}\|^2 + \tfrac{1}{2\tau} \|x - \hat{x}\|^2.$$

- Choosing $x = \bar{x} = x^k$ and $\hat{x} = x^{k+1}$, one directly obtains

$$F(x^{k+1}) \leq F(x^k) - \left(\frac{1}{\tau} - \frac{L}{2}\right) \left\|x^k - x^{k+1}\right\|^2,$$

  and hence the proximal gradient method is a descent method as long as $\tau \leq 2/L$.
- Observe that in case g is non-convex one merely looses the red term:

$$F(x^{k+1}) \leq F(x^k) - \left(\frac{1}{2\tau} - \frac{L}{2}\right) \left\|x^k - x^{k+1}\right\|^2,$$

  which is still a descent method but for smaller $\tau \leq 1/L$.

Convergence

- The previous descent rule easily implies that the sequence of objective values $(F(x^k))_{k\in\mathbb{N}}$ is non-increasing and converging and that the proximal residual

$$r(x^k) \to 0 \text{ as } k \to \infty.$$

- Moreover, assuming that the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded, there exists a subsequence $(x^k)_{k\in K\subset\mathbb{N}}$, converging to a critical x^*, i.e. a point satisfying

$$-\nabla f(x^*) \in \partial g(x^*).$$

- Under the additional assumption that $F(x)$ satisfies the Kurdyka-Łojasiewicz inequality, one can show finite length of $(x^k)_{k\in\mathbb{N}}$.

The Kurdyka-Łojasiewicz property

Definition

The function $F\colon \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ has the Kurdyka-Łojasiewicz property at $x^* \in \operatorname{dom} \partial F$, if there exist $\eta \in (0, \infty]$, a neighborhood U of x^* and a continuous concave function $\phi\colon [0, \eta) \to \mathbb{R}_+$ such that $\phi(0) = 0$, $\phi \in C^1((0, \eta))$, for all $s \in (0, \eta)$ it is $\phi'(s) > 0$, and for all $x \in U \cap [F(x^*) < F < F(x^*) + \eta]$ the Kurdyka-Łojasiewicz inequality holds, i.e.,

$$\phi'(F(x) - F(x^*)) \operatorname{dist}(0, \partial F(x)) \geq 1 \,.$$

- Intuitively, we can bound the subgradients from below by a re-parametrization of the function values
- The Kurdyka-Łojasiewicz property holds for real, semi-algebraic functions
- The Kurdyka-Łojasiewicz property attracted a lot of attention for proving convergence of descent methods [Attouch, Bolte et al. '10-'13], [Chouzenoux, Pesquet, Repetti '13], [Bolte, J., Sabach, S. and Teboulle '13], ...

Overview

Non-convex FISTA

- A natural question is whether accelerated proximal gradient methods a'la FISTA can also be applied in the non-convex setting.
- An important remark is, that the descent lemma provides an upper and lower bound:

$$f(x) + \langle \nabla f(x), y - x \rangle - \frac{\underline{L}}{2} \|x - y\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\overline{L}}{2} \|x - y\|^2,$$

  where the parameter $\overline{L}$ defines the upper and $\underline{L}$ defines the lower bound.
- Consider the steps of the standard FISTA algorithm:

$$\begin{cases} y^k = x^k + \beta(x^k - x^{k-1}) \\ x^{k+1} = \text{prox}_{\tau g}(y^k - \tau \nabla f(y^k)) \end{cases}$$

- Choosing $\bar{x} = y^k$, $x = x^k$, $\hat{x} = x^{k+1}$, and both the upper and lower bound, the basic descent rule becomes

$$g(x^k) + \underbrace{f(y^k) + \left\langle \nabla f(y^k), x^k - y^k \right\rangle}_{\leq f(x^k) + \frac{\underline{L}}{2} \|x^k - y^k\|^2} + \tfrac{1}{2\tau} \left\| x^k - y^k \right\|^2 \geq$$

$$g(x^{k+1}) + \underbrace{f(y^k) + \left\langle \nabla f(y^k), x^{k+1} - y^k \right\rangle}_{\geq f(x^{k+1}) - \frac{\overline{L}}{2} \|x^{k+1} - y^k\|^2} + \tfrac{1}{2\tau} \left\| x^{k+1} - y^k \right\|^2 + \tfrac{1}{2\tau} \left\| x^k - x^{k+1} \right\|^2.$$

Lyapunov function

- The previous descent rule becomes

$$F(x^k) + \left(\frac{1}{2\tau} + \frac{\underline{L}}{2}\right)\left\|x^k - y^k\right\|^2 \geq$$
$$F(x^{k+1}) + \left(\frac{1}{2\tau} - \frac{\overline{L}}{2}\right)\left\|x^{k+1} - y^k\right\|^2 + \frac{1}{2\tau}\left\|x^k - x^{k+1}\right\|^2$$

- Choosing $\tau = 1/\overline{L}$ and $y^k - x^k = \beta(x^k - x^{k-1})$,

$$F(x^{k+1}) + \frac{\overline{L}}{2}\left\|x^k - x^{k+1}\right\|^2 \leq F(x^k) + \beta^2\frac{\overline{L} + \underline{L}}{2}\left\|x^k - x^{k-1}\right\|^2.$$

- This defines a Lyapunov function, which decreases as long as

$$\beta^2(\overline{L} + \underline{L}) \leq \overline{L} \iff \beta \leq \sqrt{\frac{\overline{L}}{\overline{L} + \underline{L}}}.$$

- Observe that if f is convex such that $\underline{L} = 0$ one has $\beta \in [0, 1]$
- Moreover, if $\underline{L} = \overline{L} = L$ then $\beta \leq \frac{1}{\sqrt{2}}$.

Convex-Concave Inertial (CoCaIn) proximal gradient method

In [Mukkamala, Ochs, P. Sabach '19], we proposed a FISTA algorithm for non-convex optimization that makes use of a double convex-concave backtracking procedure to determine $\underline{L}$ and $\overline{L}$.

Algorithm 1 CoCaIn

Choose $x^0, x^{-1} \in \mathcal{X}$, and for all k, parameters $\overline{L}_0 > 0$:

for all $k \geq 0$ **do**

Compute

$$y^k = x^k + \sqrt{\frac{\overline{L}_k}{\overline{L}_k + \underline{L}_k}}(x^k - x^{k-1}),$$

where $\underline{L}_k$ satisfies

$$f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle \leq f(x^k) + \frac{\underline{L}_k}{2} \left\| x^k - y^k \right\|^2 .$$

Compute

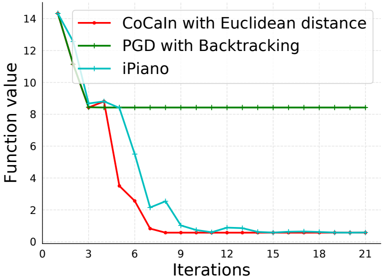
$$x^{k+1} = \operatorname{prox}_{\frac{1}{\overline{L}_k} g}(y^k - \frac{1}{\overline{L}_k} \nabla f(y^k)),$$

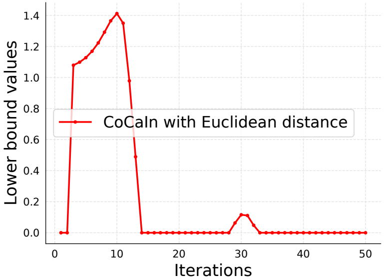
where $\overline{L}_k \geq \overline{L}_{k-1}$ satisfies

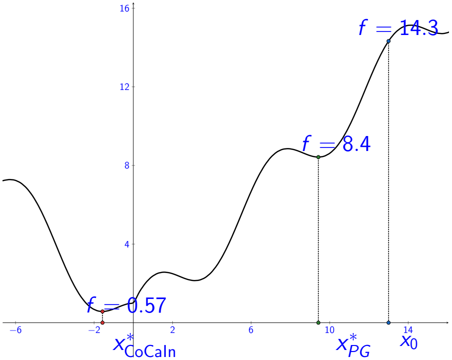
$$f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle \geq f(x^{k+1}) - \frac{\overline{L}_k}{2} \left\| x^{k+1} - y^k \right\|^2$$

end for

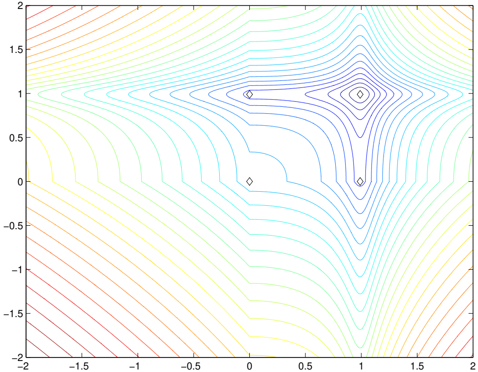
Simple Function: adapt to “local convexity”



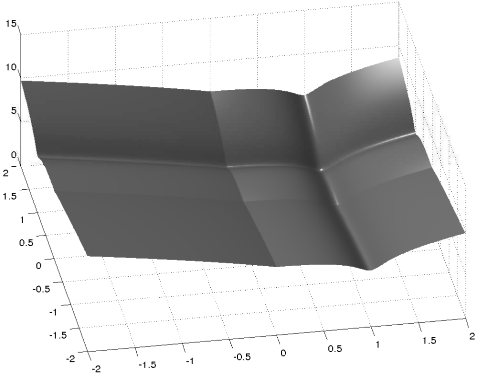




Ability to overcome spurious stationary solutions



(a) Contour plot of $h(x)$

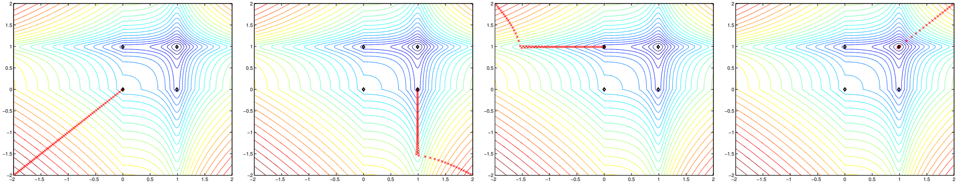


(b) Energy landscape of $h(x)$

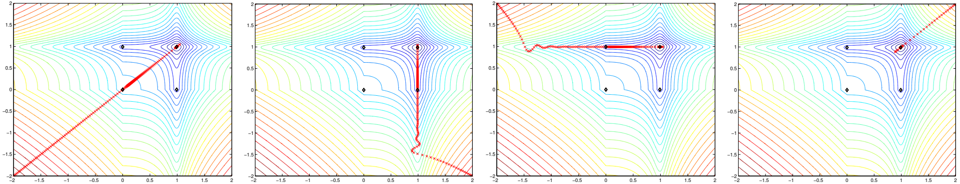
$$\min_{x \in \mathbb{R}^n} h(x) := f(x) + g(x)\,, \quad f(x) = \frac{1}{2} \sum_{i=1}^{n} \log(1 + l(x_i - y_i)^2)\,, \quad g(x) = \lambda \|x\|_1\,,$$

Effect of the inertial force

PG:



CoCaIn:



The inertial force helps to overcome spurious stationary solutions

Overview

Non-convex optimization

Non-convex proximal gradient method

Non-convex accelerated proximal gradient method

Applications

Proximal alternating linearization method

Application to image compression based on linear diffusion

- A new image compression methodology introduced in [Galic, Weickert, Welk, Bruhn, Belyaev, Seidel '08]
- The idea is to select a subset of image pixels such that the reconstruction of the whole image via linear diffusion yields the best reconstruction [Hoeltgen, Setzer, Weickert '13]

original image d | sampling mask c | compressed image u

Application to image compression based on linear diffusion

- Is written as the following constrained optimization problem

$$\begin{array}{rl} \min_{u,c} & \frac{1}{2}\|u-d\|_2^2 + \lambda\|c\|_1 \\ \text{s.t.} & C(u-d)-(I-C)Lu = 0\,, \end{array}$$

  where $C = \text{diag}(c) \in \mathbb{R}^{N\times N}$ and L is the Laplace or biharmonic operator.
- The ℓ_1 norm is used to induce sparsity in the selection mask c, $\lambda > 0$ can be used to control the amount of sparsity.
- From the (linear) side constraint we can compute

$$u = A(c)^{-1}Cd, \quad A(c) = C + (C-I)L.$$

- Hence, we can transform the original problem into an non-convex LASSO problem of the form

$$\min_c \frac{1}{2}\|A(c)^{-1}Cd - d\|_2^2 + \lambda\|c\|_1.$$

- Perfectly fits to the framework of iPiano
- We choose $f = \frac{1}{2}\|A^{-1}Cd - d\|_2^2$ and $g = \lambda\|c\|_1$
- The gradient of f is given by

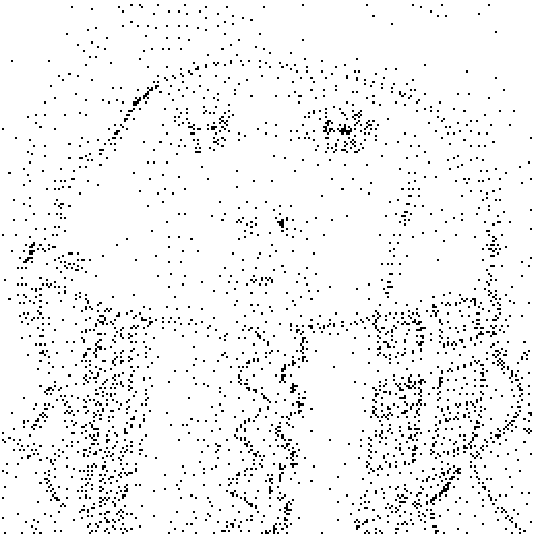
$$\nabla f(c) = \operatorname{diag}(-(I+L)u + d)(A^\top)^{-1}(u - d)\,, \quad u = A^{-1}Cd$$

- Lipschitz, if at least one entry of c is non-zero
- One evaluation of the gradient requires to solve two linear systems
- Proximal map with respect to g is standard

Results for Trui

Input

Results for Trui



5% of the pixels

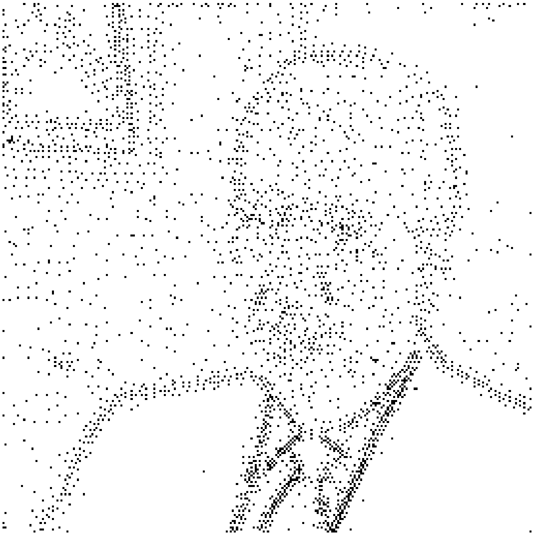
Results for Trui

Reconstruction

Results for Walter

Input

Results for Walter



5% of the pixels

Results for Walter

Reconstruction

Phase field models

- Mathematical model for solving interfacial problems
- Approximation of the interface length via the Mordica-Mortola phase field energy [Modica, Mortola, '77]

$$\mathrm{Per}(\mathbf{1}_u) \approx \int_\Omega \frac{\varepsilon}{2}|\nabla u|^2 + \frac{1}{\varepsilon} W(u)\,\mathrm{d}x\,,$$

  where u is a smooth interfacial image and $W(t) = \frac{1}{2}t^2(1-t)^2$ is a double-well potential.
- Non-convex, but Lipschitz continuous gradient.
- Note that the total variation is actually a convex functional to measure the length of the interface.

Mean curvature motion

Computes the mean curvature motion starting from a binary image:

$\beta = 0$ $\beta = 0.9$

Multi-phase-field model

Let us consider a multi-phase-field approximation of the Potts model

$$\min_{(u_i)_{i=1}^K} \frac{1}{2}\sum_{i=1}^{K}\int_\Omega \frac{\varepsilon}{2}|\nabla u_i|^2 + \frac{1}{\varepsilon}W(u_i)\ \mathrm{d}x + \int_\Omega u_i(x)f_i(x)\ \mathrm{d}x,$$

$$\text{s.t. } u_i(x) \geq 0,\ \sum_{i=1}^{K} u_i(x) = 1$$

Curvature

- Phase-fields are close to distance functions around the interface and hence they allow to reliably estimate the curvature of the interface
- Approximation of the Willmore energy

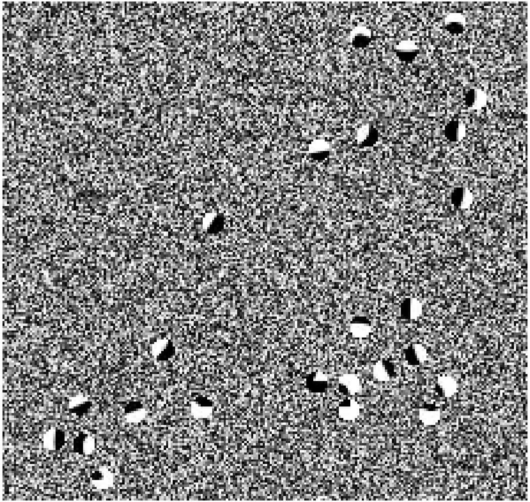
$$\frac{1}{2}\int_\Gamma h^2 \,\mathrm{d}\gamma \approx \frac{1}{2\varepsilon}\int_\Omega (\Delta u - \frac{1}{\varepsilon}W'(u))^2 \,\mathrm{d}x$$

- De Giorgi conjecture: Γ-convergence as $\varepsilon \to 0$
- Length vs. curvature regularization

Length

Curvature

Shape inpainting



Shape inpainting

Overview

Non-convex optimization

Non-convex proximal gradient method

Non-convex accelerated proximal gradient method

Applications

Proximal alternating linearization method

A more general class of non-convex problems

- Let us finally consider the following more general class of non-convex problems

$$\min_{x,y} F(x,y) := f(x,y) + g_1(x) + g_2(y),$$

  where f is smooth non-convex, $g_{1,2}$ non-smooth non-convex, simple
- In [Bolte, Sabach, Teboulle '14] the authors proposed a proximal alternating linearization method (PALM)
- Convergence of the whole sequence in case the KL property holds

---

Algorithm 2 PALM

Choose $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$.
for all $k \geq 0$ **do**
  Choose $\tau_1^k = 1/L_1(y^k)$ and let

$$x^{k+1} = \text{prox}_{\tau_1^k g_1}(x^k - \tau_1^k \nabla_x f(x^k, y^k)).$$

  Choose $\tau_2^k = 1/L_2(x^{k+1})$ and let

$$y^{k+1} = \text{prox}_{\tau_2^k g_2}(y^k - \tau_2^k \nabla_y f(x^{k+1}, y^k)).$$

end for

---

inertial PALM

- In [P., Sabach '16], we have developed an inertial variant of the PALM algorithm

Algorithm 3 iPALM

Choose $(x^{-1}, y^{-1}) = (x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$.
for all $k \geq 0$ **do**
Choose $\beta_1^k, \tau_1^k > 0$ and let

$$\begin{aligned} \hat{x}^k &= x^k + \beta_1^k (x^k - x^{k-1}) \\ x^{k+1} &= \mathrm{prox}_{\tau_1^k g_1}(\hat{x}^k - \tau_1^k \nabla_x f(\hat{x}^k, y^k)). \end{aligned}$$

Choose $\beta_2^k, \tau_2^k > 0$ and let

$$\begin{aligned} \hat{y}^k &= y^k + \beta_2^k (y^k - y^{k-1}) \\ y^{k+1} &= \mathrm{prox}_{\tau_2^k g_2}(\hat{y}^k - \tau_2^k \nabla_y f(x^{k+1}, \hat{y}^k)). \end{aligned}$$

end for

Convergence

- We again define a suitable Lyapunov function for which we can guaranteed monotone descent
- For a completely non-convex block (f and g)

$$\tau^k \leq \frac{1 - 2\beta^k}{L(x^k)(1 + 2\beta^k)}$$

- For a block with the non-smooth function g being convex

$$\tau^k \leq \frac{2(1 - \beta^k)}{L(x^k)(1 + 2\beta^k)}$$

- In case the involved functions satisfy the Kurdyka-Łojasiewicz property, we can show convergence of the whole sequence of iterates

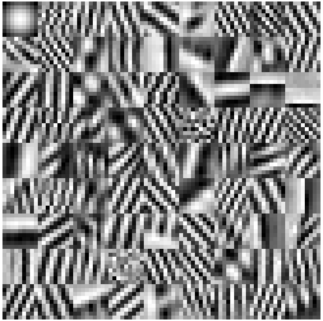
Example: Convolutional Lasso

- Let us consider the following convolutional LASSO problem [Zeiler et al.'10]

$$\min_{(d_j)_{j=1}^p,(v_j)_{j=1}^p} \sum_{j=1}^p \lambda \|v_j\|_1 + \frac{1}{2}\left\|\sum_{j=1}^p d_j *_{m,n} v_j - f\right\|^2,$$

$$\text{s.t. } d_1 = g,\ v_1 = f \ \sum_{a,b=1}^l (d_j)_{a,b} = 0,\ \|d_j\|_2 \leq 1,\ j = 2,3,\ldots,p$$

- Optimizing for the convolution kernels d_j and the coefficients v_j exactly matches the structure of the iPALM algorithm



Performance evaluation

	$K = 100$	$K = 200$	$K = 500$	$K = 1000$	time (s)
$\alpha_{1,2} = \beta_{1,2} = 0.0$	336.13	328.21	322.91	321.12	3274.97
$\alpha_{1,2} = \beta_{1,2} = 0.4$	329.20	324.62	321.51	319.85	3185.04
$\alpha_{1,2} = \beta_{1,2} = 0.8$	325.19	321.38	319.79	319.54	3137.09
$\alpha_{1,2}^k = \beta_{1,2}^k = \frac{k-1}{k+2}$	323.23	319.88	318.64	318.44	3325.37

Table: Values of the objective function for the convolutional LASSO model using different settings of the inertial parameters.

- The inertial parameter consistently improves the performance
- The dynamic step size works best, but is not supported by our convergence theory