Accelerations of Forward–Backward Splitting
— Part 1: Gradient Descent —

Peter Ochs
Saarland University
ochs@math.uni-sb.de

— June 11th – 13th, 2018 —

www.mop.uni-saarland.de

Table of Contents:

Gradient Descent Method:

- Solve an **unconstrained smooth optimization problem**:

$$\min_{x\in\mathbb{R}^N} f(x)\,, \quad \text{where } f\in C^1(\mathbb{R}^N)$$

- **Update Equation**:

$$x^{(k+1)} = x^{(k)} - \tau_k \nabla f(x^{(k)})\,.$$

- Contribution **historically** assigned to Cauchy in 1847:

[A.L. Cauchy: *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes rendus, Ac. Sci. Paris 25, 536–538 (1847).]

- He was motivated by **calculations in astronomy**.

- He wants to solve **non-linear equations**.

Augustin Louis Cauchy

[Augustin Louis Cauchy, 1789–1857
(Wikimedia, Cauchy Dibner-Collection Smithsonian Inst.)]

Facts about Gradient Descent

Gradient Descent is also known as Steepest Descent:.

- Objective has steepest descent along $d = -\nabla f(\bar{x})$.
- W.l.o.g., we can assume that $|d| = 1$ (the scaling of d can be absorbed by τ).
- For sufficiently small $\tau > 0$, the direction d **is optimal** with respect to:

$$\min_{d \in \mathbb{R}^N} \frac{f(\bar{x} + \tau d) - f(\bar{x})}{\tau} \quad s.t.\ |d| = 1\,.$$

- Consider the first order Taylor expansion:

$$f(\bar{x} + \tau d) = f(\bar{x}) + \tau \langle \nabla f(\bar{x}), d \rangle + o(\tau |d|)\,.$$

  (Note that for $\tau \to 0$, the term $o(\tau)$ vanishes faster than $\tau \langle \nabla f(\bar{x}), d \rangle$.)

- The direction d solves the following problem

$$\min_{d \in \mathbb{R}^N} \langle \nabla f(\bar{x}), d \rangle \quad s.t.\ |d| = 1\,.$$

Facts about Gradient Descent

- **Problem**:

$$\min_{d \in \mathbb{R}^N} \langle \nabla f(\bar{x}), d \rangle \quad s.t. \ |d| = 1 \,.$$

- Denote by θ the angle between $\nabla f(\bar{x})$ and d and write:

$$\langle \nabla f(\bar{x}), d \rangle = |\nabla f(\bar{x})||d| \cos \theta \,,$$

- Therefore, problem is **solved by**

$$d = -\frac{\nabla f(\bar{x})}{|\nabla f(\bar{x})|} \,.$$

- **Negative gradient** $-\nabla f(\bar{x})$ points in the **direction of steepest descent**.

Descent Direction

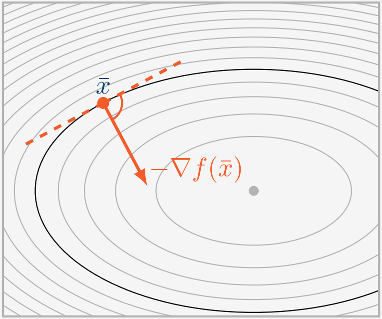
Definition: (Descent Direction)

A vector $0 \neq d \in \mathbb{R}^N$ is a *descent direction* for the function f at the point $\bar{x}$, if $\langle \nabla f(\bar{x}), d \rangle < 0$ holds, i.e. the angle between d and $\nabla f(\bar{x})$ is larger than 90 degree (obtuse angle).

- For descent direction d:

$$f(\bar{x} + \tau d) = f(\bar{x}) + \tau \underbrace{\langle \nabla f(\bar{x}), d \rangle}_{<0} + o(\tau |d|)$$

$$\underset{\tau \text{ small}}{<} f(\bar{x})$$



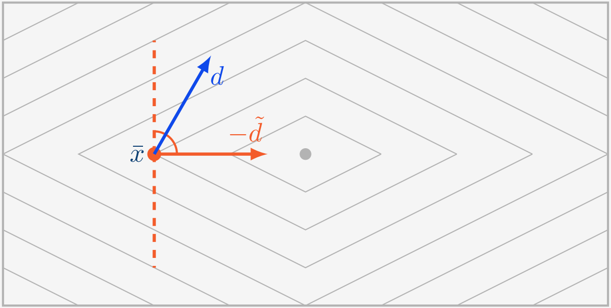
Example:

- B positive definite, $d = -B\nabla f(\bar{x}) \neq 0$:

$$\langle \nabla f(\bar{x}), d \rangle \leq -\lambda_{\text{min}}(B) |\nabla f(\bar{x})|^2 < 0\,.$$

Descent Direction for Non-smooth Functions?

Remark: This definition is not true for non-smooth functions:



- $-\tilde{d}$ steepest descent direction.
- d satisfies $\left\langle d, \tilde{d} \right\rangle < 0$.
- **However**, $f(\bar{x} + \tau d) > f(\bar{x})$ for any $\tau > 0$.

Sufficient Descent Condition is Required

Sufficient Descent Condition:

- Is $f(x^{(k+1)}) < f(x^{(k)})$ "**sufficient**" to find a **minimizer** or a **stationary point**

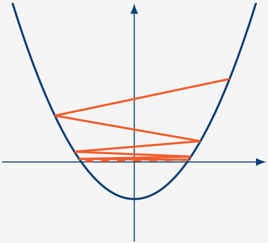
$$\nabla f(x^\star) = 0\,? \qquad (x^\star \text{ is called } \textit{stationary} \text{ or } \textit{critical point})$$

Example:

$f(x) = x^2 - 1$. Start at $x^{(0)} = 2$; descent direction $d^{(k)} = -x^{(k)}/|x^{(k)}|$ and $\tau^{(k)}$ such that $f(x^{(k)}) = 1/(k+1)$. Then, obviously,

$$f(x^{(k+1)}) = \frac{1}{k+2} < \frac{1}{k+1} = f(x^{(k)})\,,$$

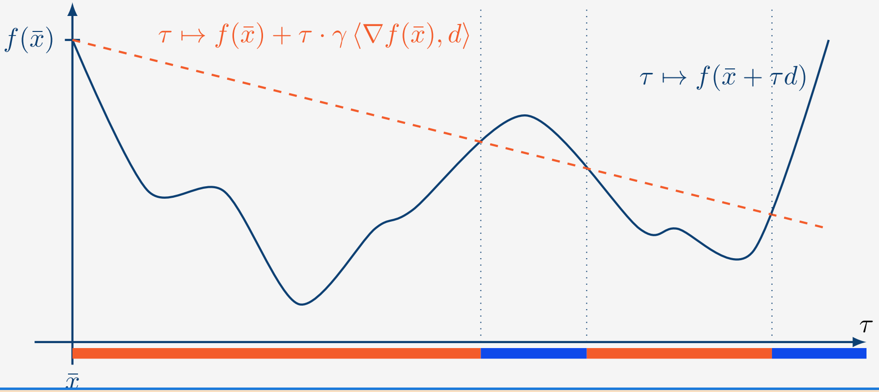
however $f(x^{(k)}) \to 0$ for $k \to \infty$ and $\min f = -1$. This algorithm does not converge to the minimum.



Definition (Armijo condition):

The step size $\tau > 0$ is said to satisfy the *Armijo condition* for $\gamma \in (0, 1)$ and the descent direction $d \in \mathbb{R}^N$ at the point $\bar{x} \in \mathbb{R}^N$, if the following holds:

$$f(\bar{x} + \tau d) \leq f(\bar{x}) + \gamma\tau \langle \nabla f(\bar{x}), d \rangle$$



Armijo condition

Example: (Armijo condition)

- Let $d = -\nabla f(\bar{x})$. Then, the Armijo condition reads

$$f(\bar{x} + \tau d) \leq f(\bar{x}) - \gamma\tau|\nabla f(\bar{x})|^2 .$$

- Descent achieved whenever $\tau|\nabla f(\bar{x})|^2 > 0$ (i.e. $\bar{x}$ is not a stationary point).

- A small descent of the objective values means that τ is small or $|\nabla f(\bar{x})|^2$ is small:

$$\gamma\tau|\nabla f(\bar{x})|^2 \leq f(\bar{x}) - f(\bar{x} + \tau d)$$

- The difference between successive objective values is a **measure for the stationarity** of the iterates (scaled by τ).

Algorithm (Backtracking Line Search Method):

- **Prerequisites:** Descent direction $d \in \mathbb{R}^N$ at $\bar{x} \in \mathbb{R}^N$ for $f \in C^1(\mathbb{R}^N)$.
- **Goal:** Find a step size τ that satisfies the Armijo condition.
- **Procedure:**
  - **Initialize:** Let $\bar{\tau} > 0$, $\gamma, \rho \in (0, 1)$ and set $\tau^{(0)} = \bar{\tau}$.
  - **For** $j = 0, 1, 2, \dots$**:** If the condition

$$f(\bar{x} + \tau^{(j)} d) \leq f(\bar{x}) + \gamma \tau^{(j)} \langle \nabla f(\bar{x}), d \rangle$$

is satisfied, then stop the algorithm and return $\tau^{(j)}$, otherwise

set $\tau^{(j+1)} = \rho \tau^{(j)}$.

Convergence of Gradient Descent

Proposition (Stationarity of Limit Points):

Let

- $f \in C^1(\mathbb{R}^N)$
- $(x^{(k)})_{k\in\mathbb{N}}$ be generated by Gradient Descent $d^{(k)} = -\nabla f(x^{(k)})$
- $(\tau_k)_{k\in\mathbb{N}}$ selected by backtracking line search satisfies the Armijo condition.

Then

- every limit point of $(x^{(k)})_{k\in\mathbb{N}}$ is a stationary point of f.

Convergence of Gradient Descent

Proposition (Constant Step Size Rule):

Let

- $f \in C^1(\mathbb{R}^N)$ with L-Lipschitz continuous gradient ∇f:

$$|\nabla f(x) - \nabla f(y)| \leq L|x - y|\,, \quad \forall x, y \in \mathbb{R}^N$$

- $(x^{(k)})_{k\in\mathbb{N}}$ be generated by Gradient Descent $d^{(k)} = -\nabla f(x^{(k)})$
- for some $\varepsilon > 0$, the step sizes $(\tau_k)_{k\in\mathbb{N}}$ satisfy

$$\varepsilon \leq \tau_k \leq \frac{2-\varepsilon}{L}\,.$$

Then

- every limit point of $(x^{(k)})_{k\in\mathbb{N}}$ is a stationary point of f.

Discussion Convergence

Discussion: (Convergence of Gradient Descent):

- $(f(x^{(k)}))_{k\in\mathbb{N}}$ **converges** to $f^* > -\infty$.
- Every **limit point** x^* satisfies

$$\nabla f(x^*) = 0\,, \quad \text{i.e. it is a } \textbf{stationary} \text{ point.}$$

- x^* is not necessarily a local minimizer.
- Possibly: Convergence to a saddle point or local maximum.
- The sequence $(x^{(k)})_{k\in\mathbb{N}}$ **does not necessarily converge**, although

$$|\nabla f(x^{(k)})| \to 0 \quad \stackrel{\tau_k=\tau\neq 0}{\Longrightarrow} \quad |x^{(k+1)} - x^{(k)}| \to 0\,.$$

Counterexample for Convergence

Counterexample:

- Gradient Descent with line minimization does not converge to a single point.

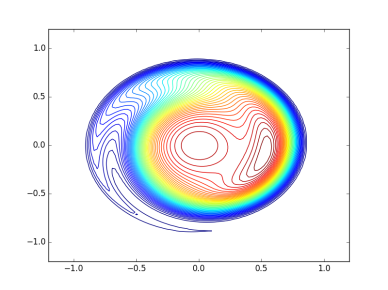
- [H. B. Curry: *The method of steepest descent for non-linear minimization problems*, Quart. Appl. Math., 2 (1944), pp. 258–261.]:

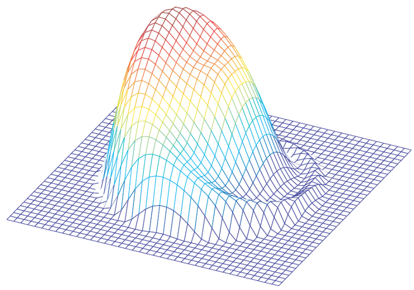
  Let $f(x_1, x_2) = 0$ on the unit circle and $f(x_1, x_2) > 0$ for any other point. Outside the unit circle let the surface have a spiral gully making infinitely many turns about the circle. The iterates will follow the gully and have all points of the circle as limit points.

- Counterexample given by a C^∞-function. (See next slide.)

Counterexample for Convergence

Counterexample:





From [Absil, Mahony, Andrews 2005]

- Defined in polar coordinates (r, θ):

$$f(r,\theta) := \begin{cases} e^{-\frac{1}{1-r^2}} \left(1 - \frac{4r^4}{4r^4+(1-r^2)^4} \sin\left(\theta - \frac{1}{1-r^2}\right)\right), & \text{if } r < 1\,; \\ 0\,, & \text{if } r \geq 1\,; \end{cases}$$

Convergence to a Single Stationary Point

Convergence to a Single Point: (Requires additional assumptions)

- Critical points isolated or Hessian non-degenerate [Helmke, Moore 1994].
- Strictly convex functions: Global minimum is unique isolated critical point.
- Objective differentiable quasi-convex [Kiwiel, Murty 1996].
- Convergence to isolated local minimum [Bertsekas 1995].
  (Capture Theorem)
- Pseudo-convexity conditions and growth conditions [Dunn 1981, 1987].
- f convex, ∇f Lipschitz, const. step size, e.g. [Bauschke, Combettes 2011].
  (using Fejér Monoticity)
- Real analytic functions [Absil, Mahony, Andrews 2005].
  (using Łojasiewicz inequality)
- Tame functions [Bolte, Daniilidis, Ley, Mazet 2010].

Part 4:
Single Point Convergence

1. Łojasiewicz Inequality
2. Kurdyka-Łojasiewicz Inequality
3. Abstract Convergence Theorem
4. Convergence of Non-convex Forward–Backward Splitting
5. A Generalized Abstract Convergence Theorem
6. Convergence of iPiano
7. Local Convergence of iPiano

Convergence Speed of Gradient Descent

Convergence Rate for Smooth Strongly Convex Functions:

- $f \in \mathscr{S}^{1,1}_{\mu,L}$ (smooth strongly convex), i.e. $f(x) - \frac{\mu}{2}|x|^2$ convex.
- For $\tau \in (0, 2/(\mu+L)]$

$$|x^{(k+1)} - x^\star|^2 \leq \Big(1 - \frac{2\tau\mu L}{\mu + L}\Big)^k |x^{(0)} - x^\star|^2 \,.$$

If $\tau = 2/(\mu+L)$, then

$$|x^{(k+1)} - x^\star|^2 \leq \Big(\frac{L-\mu}{L+\mu}\Big)^{2k} |x^{(0)} - x^\star|^2 \,.$$

Linear convergence rate [Nesterov 2004].

Convergence Speed of Gradient Descent

Convergence Rate for Smooth Convex Functions:

- $f \in \mathcal{F}_L^{1,1}$ (smooth convex).
- For $\tau \in (0, 2/L)$

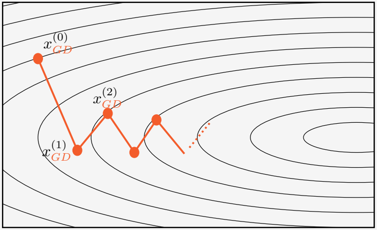
$$f(x^{(k)}) - f^\star \leq \frac{2(f(x^{(0)}) - f^\star)\|x^{(0)} - x^*\|^2}{2\|x^{(0)} - x^*\|^2 + k\tau(2 - \tau L)(f(x^{(0)}) - f^\star)} = \mathcal{O}(1/k)\,.$$

Sub-Linear convergence rate [Nesterov 2004].

Convergence Speed of Gradient Descent

Convergence Speed of Gradient Descent: (Discussion)

- We have upper complexity bounds for Gradient Descent.
- Still **unclear**, how good Gradient Descent is.
- For irregularly scaled level sets, Gradient Descent is bad.



- For some classes of problems, we have **lower complexity bounds**.
  [Nesterov 2004], [Nemirovski, Yudin 1983].

Lower complexity bound for $\mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^\infty)$, [Nesterov 2004]

Theorem: (Lower Bound for Smooth Strongly Convex Functions)

For any $x^{(0)} \in \mathbb{R}^\infty$ and any constants $\mu > 0$, $L > \mu$ there exists a function $f \in \mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^\infty)$ such that for any first-order method $\mathcal{M}$ satisfying our assumptions, we have

$$\|x^{(k)} - x^\star\|^2 \geq q^{2k} \|x^{(0)} - x^\star\|^2 \,, \qquad q := \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

$$f(x^{(k)}) - f^\star \geq \tfrac{\mu}{2} q^{2k} \|x^{(0)} - x^\star\|^2 \,.$$

Discussion:

- The “worst function” depends on μ and L, but not on k.
- The bound is uniform in the dimension.
- Turns out to be tight for quadratic functions (e.g. Conjugate Gradient Method).
- The rate is “much” worse for Gradient Descent:

$$q_{\text{GD}} := \frac{L-\mu}{L+\mu} \quad \text{vs} \quad q_{\text{opt}} := \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

Lower complexity bound for $\mathcal{F}_L^{\infty,1}(\mathbb{R}^N)$, [Nesterov 2004]

Theorem: (Lower Bound for Smooth Convex Functions)

For any k with $1 \leq k \leq \frac{1}{2}(N-1)$ and any $x^{(0)} \in \mathbb{R}^N$, there exists at least one function $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^N)$ such that for any first order method $\mathcal{M}$ satisfying our assumption, we have that

$$f(x^{(k)}) - f^\star \geq \frac{3L\|x^{(0)} - x^\star\|^2}{32(k+1)^2}\,, \quad \text{i.e. } f(x^{(k)}) - f^\star \in \mathcal{O}(1/k^2)$$

Discussion:

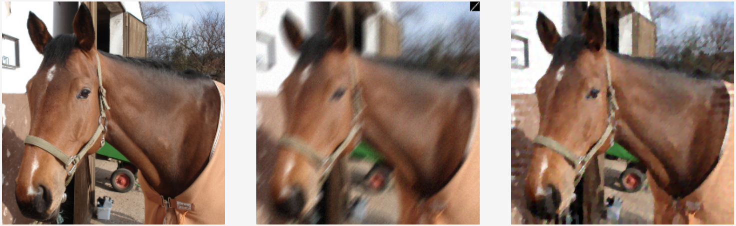
- The estimates are valid for large scale problems ($N > 10^5$), or for the first iterates of small problems ($N < 10^4$).
- The complexity bound is **uniform in the dimension** of the problem.
- Unclear whether the estimation of the lower complexity bound is tight.
- After $k = 100$ iterations we can decrease our initial residual by a factor of 10^4.
- In order to improve the situation, we have to find another problem class.
- Obviously, **Gradient Descent is not optimal** $\mathcal{O}(1/k)$.

Part 2:
Acceleration Strategies

1. Time Continuous Setting
2. Heavy-ball Method
3. Nesterov's Acceleration
4. Quasi-Newton Methods
5. Subspace Acceleration

Image Processing: (Image Denoising, Deblurring)

- $\mathbf{f} \in \mathbb{R}^N$: degraded (grey-value) image



clean image $\mathbf{g}$ noisy image $\mathbf{f}$ reconstruction $\mathbf{u}$

- Suppose degradation process is known $\mathcal{A}\colon \mathbb{R}^N \to \mathbb{R}^N$ (linear):

$$\mathbf{f} = \mathcal{A}(\mathbf{g}) + \mathbf{n}$$

- $\mathbf{g} \in \mathbb{R}^N$: ground truth/clean image.
- $\mathbf{n} \in \mathbb{R}^N$: noise (e.g. Gaussian or Impulse noise)
- We also consider (non-additive) Poisson noise. (*different formula*)

Image Processing: (Image Denoising, Deblurring)

Reconstruction by Variational Methods:

$$\min_{\mathbf{u}\in\mathbb{R}^N} \underbrace{D(\mathbf{u})}_{\text{data term}} + \lambda \underbrace{R(\mathbf{u})}_{\text{regularization term}}$$

- **Data term:** Reconstruction/solution $\mathbf{u}$ should be similar to $\mathbf{f}$.
  - $D(\mathbf{u}) = \|\mathcal{A}(\mathbf{u}) - \mathbf{f}\|_2^2$: good for removing Gaussian noise.
  - $D(\mathbf{u}) = \|\mathcal{A}(\mathbf{u}) - \mathbf{f}\|_1$: good for removing impulse noise.

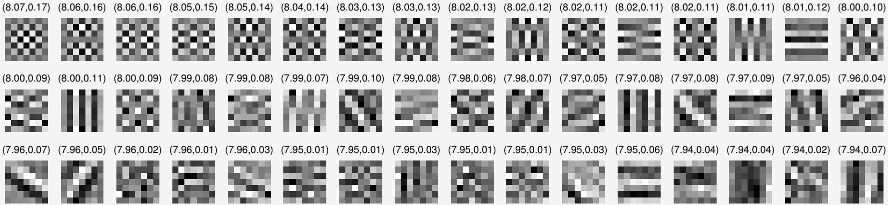
- **Regularization term:** $\mathbf{u}$ should not contain noise, i.e. it should be smooth:
  - Define **finite-difference operator** $\mathcal{D}\colon \mathbb{R}^N \to \mathbb{R}^{2N}$ for $\mathbf{u} \in \mathbb{R}^{n_x \times n_y} \simeq \mathbb{R}^N$ by

$$\mathcal{D} = (\mathcal{D}^x, \mathcal{D}^y)\,, \quad (\mathcal{D}\mathbf{u})^x_{i,j} = \begin{cases} \mathbf{u}_{i+1,j} - \mathbf{u}_{i,j}\,, & \text{if } i < n_x \\ 0\,, & \text{otherwise.} \end{cases}$$

  - $R(\mathbf{u}) = \|\mathcal{D}\mathbf{u}\|_2^2$ (Tikhonov regularization)
  - $R(\mathbf{u}) = \|\mathcal{D}\mathbf{u}\|_{2,1} = \sum_{i,j} ((\mathcal{D}^x\mathbf{u})^2_{i,j} + (\mathcal{D}^y\mathbf{u})^2_{i,j})^{1/2}$ ((isotropic) Total Variation)
  - $R(\mathbf{u}) = \|\mathcal{D}\mathbf{u}\|_1 = \sum_{i,j} |(\mathcal{D}^x\mathbf{u})_{i,j}| + |(\mathcal{D}^y\mathbf{u})_{i,j}|$ ((anisotropic) Total Variation)
  - $R(\mathbf{u}) = \sum_{i,j} \varphi((\mathcal{D}\mathbf{u})_{i,j})$ with $\varphi(p) = \log(1 + \nu|p|)$ (non-convex) ...

Regularization term:

- Also known as **prior assumption**.
- **Natural image statistics** motivate the use of **non-convex** regularizers.
- **Learned regularization filters:**



Applications: LASSO

Least Absolute Shrinkage and Selection Operator: [Tibshirani 1994]

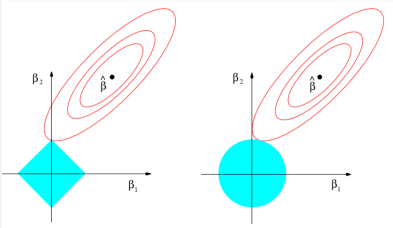
$$\min_{x\in\mathbb{R}^N} \frac{1}{2}|Ax-b|^2 + \lambda\|x\|_1 \quad \text{or} \quad \min_{x\in\mathbb{R}^N} \frac{1}{2}|Ax-b|^2 \quad s.t.\ \|x\|_1 \leq \lambda\,.$$

- Sparse **linear regression**: ($A_i \in \mathbb{R}^M$ is a feature for describing b)

$$b \approx \sum_{i=1}^N A_i x_i\,, \quad A = (A_1,\dots,A_N) \in \mathbb{R}^{M\times N}\,,\ x = (x_1,\dots,x_N)^\top\,.$$

- $\|x\|_1$ used as a convex approximation to $\#\{i\,:\,x_i \neq 0\}$.
- **Motivation**: Many zero-coordinates yield an interpretable model

$$b \approx \sum_{i=1}^N A_i x_i = \sum_{j\in\{i\,:\,x_i\neq 0\}} A_j x_j\,.$$



Applications

Similar problems:

- **Group Lasso**, **Fused Lasso**, ...

- **Logistic Regression:** $(x_i, y_i) \in X \times \{-1, 1\}$ given "training data":

$$\min_{w \in \mathbb{R}^N} \sum_i \log(1 + \exp(-y_i \langle w, x_i \rangle)) + \lambda \|w\|_1 \, .$$

- **Non-negative Least Squares:**

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} |Ax - b|^2 \quad s.t.\ x_i \geq 0 \ \forall i = 1, \ldots, N \, .$$

- **Elastic Net Regularization:**

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} |Ax - b|^2 + \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2 \, .$$

- **Low Rank Approximation:** (e.g. Matrix completion)

$$\min_{X \in \mathbb{R}^{M \times N}} \frac{1}{2} \|A - X\|_F^2 + \lambda \|X\|_* \, .$$

Neural Networks:

- **Non-linear Regression Problem:** (or interpolation)
- Given training data $(x_i, y_i) \in X \times Y$, $i = 1, \dots, M$.
- **Training**: Find $w \in \mathbb{R}^P$ such that

$$\mathcal{N}_w(x_i) \approx y_i \qquad i = 1, \dots, M$$

- The **non-linear prediction function** has a composition structure (L layer):

$$\mathcal{N}_w(x) = w_L \sigma(\dots \sigma(w_2 \sigma(w_1 x + b_1) + b_2) \dots) + b_L$$

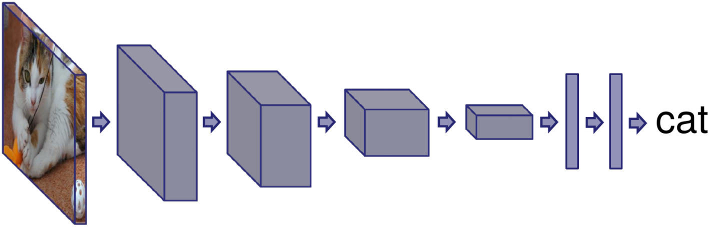
with “activation functions” σ (coordinate-wise non-linear functions) and

$$w = (w_1, \dots, w_L, b_1, \dots, b_L)\,.$$

Neural Networks

- **Optimization Problem/Training**: (e.g. Empirical risk)

$$\min_{w\in\mathbb{R}^N} \frac{1}{2}\sum_{i=1}^{M} |\mathcal{N}_w(x_i) - y_i|^2 \quad \text{or} \quad \min_{w\in\mathbb{R}^N} \frac{1}{2}\sum_{i=1}^{M} \max(0, 1 - y_i\mathcal{N}_w(x_i))\,.$$



- Can also be complemented with sparsity or other priors for w.
- Use **robust non-linear regression**, when outliers are expected:

$$\min_{w\in\mathbb{R}^N} \frac{1}{2}\sum_{i=1}^{M} \|\mathcal{N}_w(x_i) - y_i\|_1\,.$$

Part 3:
Non-smooth Optimization

1. Basic Definitions
2. Infimal Convoution
3. Proximal Mapping
4. Subdifferential
5. Optimality Condition (Fermat's Rule)
6. Proximal Point Algorithm
7. Forward–Backward Splitting

Non-smooth Optimization

Structured Optimization Problems:

- Most of the applications yield **structured non-smoothness**:

$$\min_{x \in \mathbb{R}^N} \; f(x) + g(x)$$

- f is a smooth function.
- g is a non-smooth function with “nice properties”.
- **Forward–Backward Splitting** is designed for such problems.

Part 3: Non-smooth Optimization

6. Proximal Point Algorithm
7. Forward–Backward Splitting

Part 4: Single Point Convergence

4. Convergence of Non-convex Forward–Backward Splitting

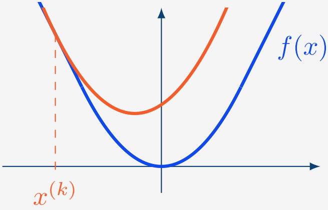
Part 5: Variants and Acceleration of Forward–Backward Splitting

1. FISTA
2. Adaptive FISTA
3. Proximal Quasi-Newton Methods
4. Efficient Solution for Rank-1 Perturbed Proximal Mapping
5. Forward–Backward Envelope
6. Generalized Forward–Backward Splitting

Interpretation of Gradient Descent: (Relations to other Algorithms)

- Gradient Descent step equivalent to **minimizing a quadratic function**:

$$x^{(k+1)} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\ f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle + \frac{1}{2\tau} |x - x^{(k)}|^2 \,.$$



- Optimality condition:

$$\nabla f(x^{(k)}) + \frac{1}{\tau}(x - x^{(k)}) = 0$$
$$\Leftrightarrow x = x^{(k)} - \tau \nabla f(x^{(k)})$$

Interpretation of Gradient Descent

Another point of view:

- Minimization of a **linear function**

$$f_{x^{(k)}}(x) = f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle$$

with **quadratic penalty on the distance** to $x^{(k)}$:

$$D_h(x, x^{(k)}) = \frac{1}{2\tau} |x - x^{(k)}|^2 \, .$$

- Update step:

$$x^{(k+1)} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \; f_{x^{(k)}}(x) + D_h(x, x^{(k)})$$

Interpretation of Gradient Descent

Generalization to non-smooth functions f:

- Minimization of a convex **model function**

$$f_{x^{(k)}}(x) \quad \text{with} \quad |f(x) - f_{x^{(k)}}(x)| \le \underbrace{\omega(|x - x^{(k)}|)}_{\text{growth function}}$$

with **quadratic penalty on the distance** to $x^{(k)}$:

$$D_h(x, x^{(k)}) = \frac{1}{2\tau}|x - x^{(k)}|^2 \,.$$

- Update step:

$$x^{(k+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} \; f_{x^{(k)}}(x) + D_h(x, x^{(k)})$$

Interpretation of Gradient Descent

Generalization to non-smooth functions f:

- Minimization of a convex **model function**

$$f_{x^{(k)}}(x) \quad \text{with} \quad |f(x) - f_{x^{(k)}}(x)| \leq \underbrace{\omega(|x - x^{(k)}|)}_{\textbf{growth function}}$$

  with **penalty on the distance** to $x^{(k)}$:

$$D_h(x, x^{(k)})\,.$$

- Update step:

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} f_{x^{(k)}}(x) + D_h(x, x^{(k)})$$

Part 6:
Bregman Proximal Minimization

1. Model Function Framework
2. Examples of Model Functions
3. Examples of Bregman Functions
4. Convergence Results
5. Applications

Example for Unification: (Convergence Rate for the Gradient Method)

- Set the **model**: $f_{\bar{x}}(x) = f(\bar{x}) + \left\langle \nabla f(\bar{x}), x - x^k \right\rangle$ (**Gradient Descent**).

- $f_{\bar{x}}$ satisfies the **model assumption**:

$$0 \leq f(x) - f_{\bar{x}}(x) \leq \frac{L}{2} \|x - \bar{x}\|^2 \,.$$

- Define:

$$f_{\bar{x}}^{\tau}(x) := f_{\bar{x}}(x) + \frac{1}{2\tau} \|x - \bar{x}\|^2 \,,$$

  i.e.

$$\hat{x} = \arg\min_{x \in \mathbb{R}^N} \; f_{\bar{x}}^{\tau}(x) \,.$$

- $f_{\bar{x}}^{\tau}$ is τ^{-1}-strongly convex, i.e.

$$f_{\bar{x}}^{\tau}(\hat{x}) + \frac{1}{2\tau} \|\hat{x} - x\|^2 \leq f_{\bar{x}}^{\tau}(x) \,.$$

Convergence Rate for the Gradient Method

- $f_{\bar{x}}^{\tau}$ is τ^{-1}-strongly convex, i.e.

$$f_{\bar{x}}^{\tau}(\hat{x}) + \frac{1}{2\tau}\|\hat{x} - x\|^2 \leq f_{\bar{x}}^{\tau}(x)\,.$$

- Using the model assumption, we obtain:

$$f(\hat{x}) + \left(\frac{1}{2\tau} - \frac{L}{2}\right)\|\hat{x} - \bar{x}\|^2 + \frac{1}{2\tau}\|\hat{x} - x\|^2 \leq f(x) + \frac{1}{2\tau}\|x - \bar{x}\|^2\,.$$

- Using $x = \bar{x}$ and $0 < \tau < \frac{2}{L}$, we obtain a **descent algorithm**.
- Restricting to $0 < \tau \leq \frac{1}{L}$, we obtain

$$f(\hat{x}) - f(x) \leq \frac{1}{2\tau}\left(\|x - \bar{x}\|^2 - \|x - \hat{x}\|^2\right)\,.$$

- Set $x = x^\star$, $\hat{x} = x^{(k+1)}$ and $\bar{x} = x^{(k)}$, and sum both sides

$$f(x^{(k+1)}) - f(x^\star) \leq \frac{\|x^\star - x^{(0)}\|^2}{2\tau k} \overset{\tau = \frac{1}{L}}{=} \frac{L\|x^\star - x^{(0)}\|^2}{2k}\,.$$

Accelerations of Forward–Backward Splitting
— Part 2: Acceleration Strategies —

Peter Ochs
Saarland University
ochs@math.uni-sb.de

— June 11th – 13th, 2018 —

www.mop.uni-saarland.de

Table of Contents

Time Continuous Interpretation of Gradient Descent:

- Let $(x^{(k)})_{k\in\mathbb{N}}$ be generated by Gradient Descent.
- Then
$$x^{(k+1)} = x^{(k)} - \tau\nabla f(x^{(k)}) \quad \Leftrightarrow \quad \frac{x^{(k+1)} - x^{(k)}}{\tau} = -\nabla f(x^{(k)})\,.$$
- Consider as discretization of a curve $X\colon [0,+\infty) \to \mathbb{R}^N$, $t \mapsto X(t)$.
- Set
$$t_k := k\tau \quad \text{and} \quad X(t_k) = x^{(k)}\,.$$
- Taylor expansion:
$$\begin{aligned} X(t_{k+1}) &= X(t_k) + \dot X(t_k)(t_{k+1} - t_k) + \mathcal{O}(\tau^2) \\ &= X(t_k) + \tau\dot X(t_k) + \mathcal{O}(\tau^2) \end{aligned}$$
- Therefore
$$\frac{X(t_{k+1}) - X(t_k)}{\tau} = \dot X(t_k) + \mathcal{O}(\tau) = -\nabla f(X(t_k))\,.$$

Gradient descent dynamical system

Gradient descent dynamical system:

- Also known as **gradient descent dynamical system**.

- Given by the differential equation:

$$\dot{X}(t) + \nabla f(X(t)) = 0$$

- $X : [0, +\infty) \to \mathbb{R}^N$ curve with time derivative $\dot{X}$.

- $X \in C^1$ is a *solution (curve)*, when it satisfies the differential equation.

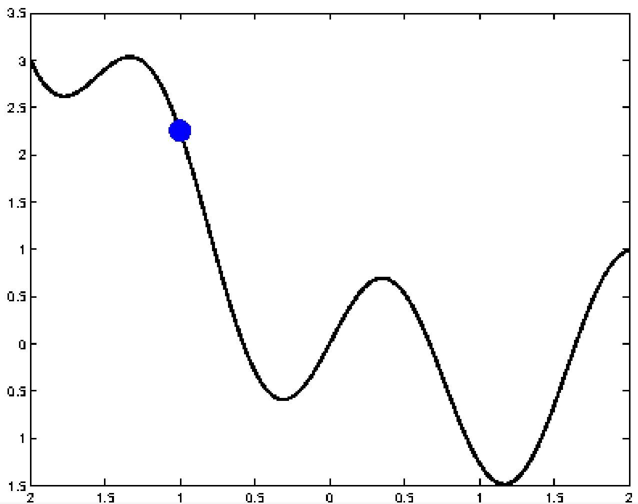
- If we fix $X(0) = X_0 \in \mathbb{R}^N$, existence and uniqueness is a classical result in the theory of Ordinary Differential Equations.

- f is a Lyapunov function, i.e. it decreases along the solution curve:

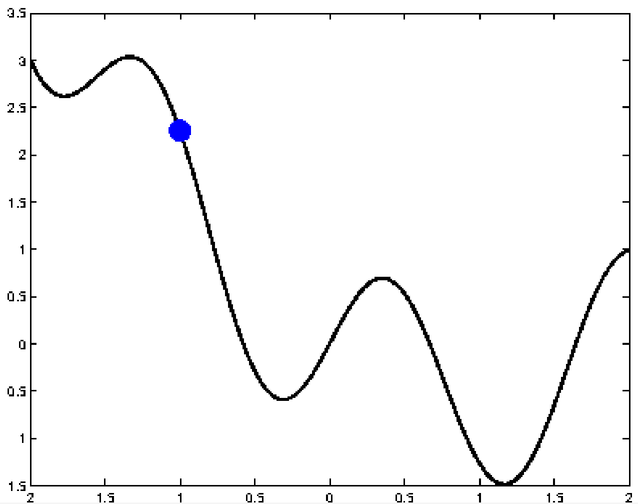
$$\frac{d}{dt}(f \circ X)(t) = \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle = -|\nabla f(X(t))|^2 \overset{\nabla f(X(t)) \neq 0}{<} 0\,.$$

Gradient descent dynamical system

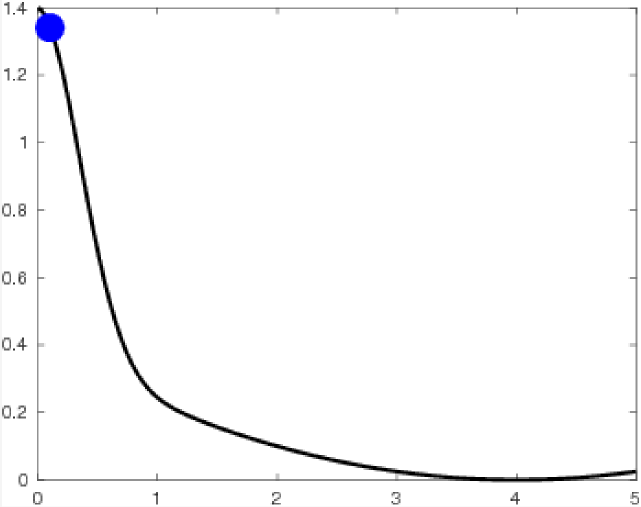
Gradient descent dynamical system:



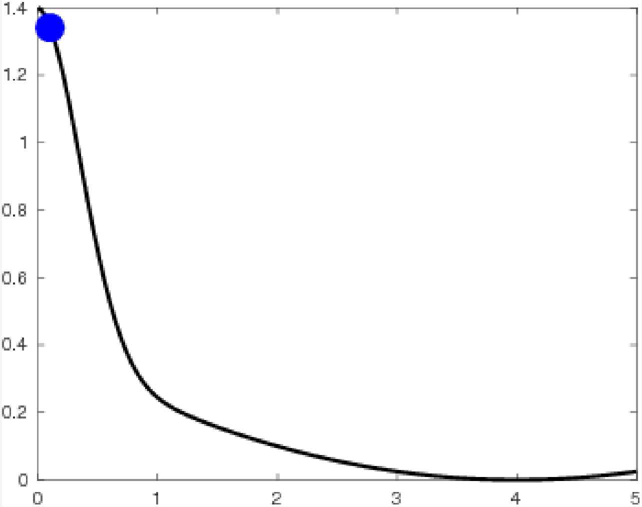
Gradient descent dynamical system:



Gradient descent dynamical system:



Gradient descent dynamical system:



Heavy-ball Dynamical System with Friction

Heavy-ball Dynamical System with Friction:

- Differential equation:

$$\ddot{X}(t) = -\gamma \dot{X}(t) - \nabla f(X(t))$$

- Describes the motion of a ball on the graph of the objective function f.

- $\ddot{X}(t)$ is the second derivative ($\sim$ acceleration).
  $\rightsquigarrow$ models **inertia** / **momentum**.

- $-\gamma \dot{X}$ is a viscous friction force ($\gamma > 0$).

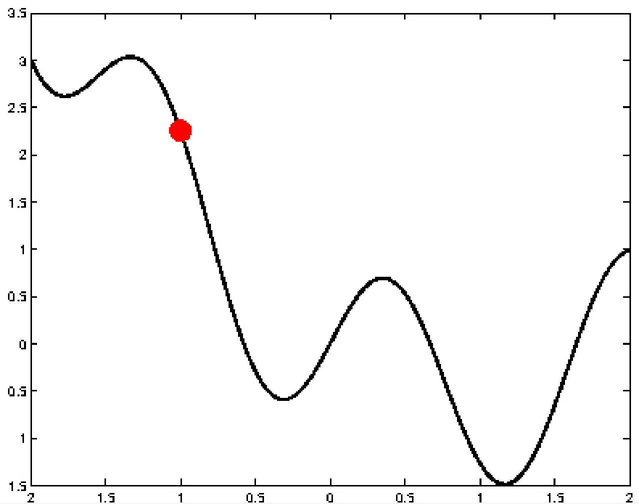
- **Lyapunov function**: $F(t) := f(X(t)) + \frac{1}{2}|\dot{X}(t)|^2$

$$\frac{d}{dt}(F \circ X)(t) = \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle + \left\langle \dot{X}(t), \ddot{X}(t) \right\rangle = -\gamma |\dot{X}(t)|^2 \overset{\dot{X}(t) \neq 0}{<} 0\,.$$

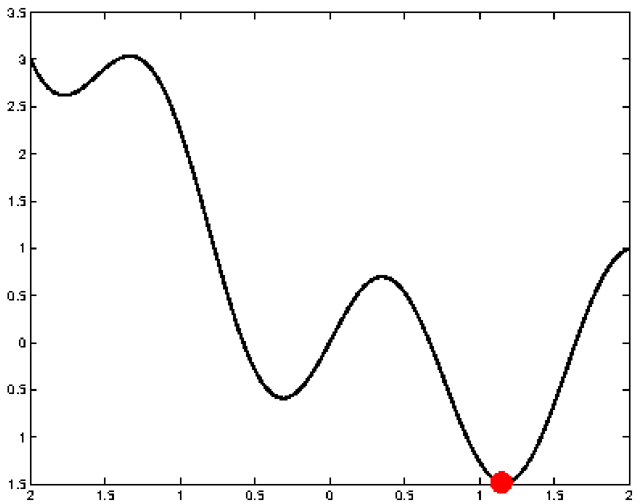
- [Attouch, Goudou, Redont 2000]:

$$\lim_{t \to \infty} \dot{X}(t) = \lim_{t \to \infty} \ddot{X}(t) = \lim_{t \to \infty} \nabla f(X(t)) = 0\,.$$

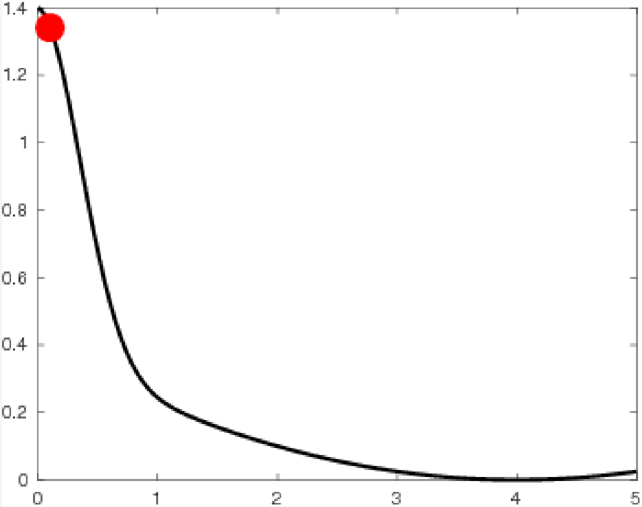
Inertial methods can speed up convergence



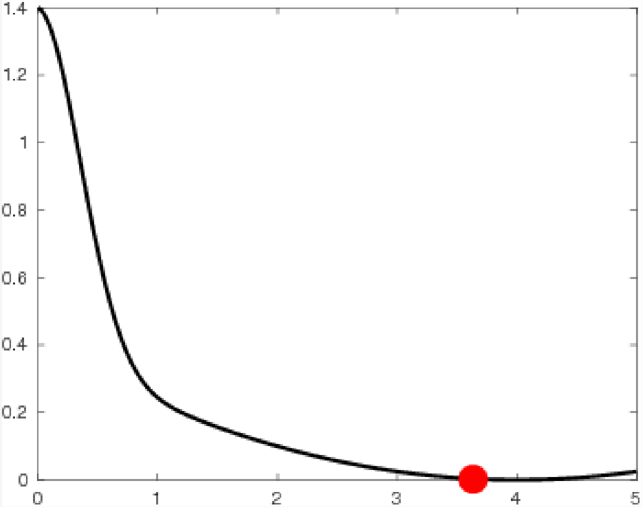
Inertial methods can speed up convergence



Inertial methods can speed up convergence



Inertial methods can speed up convergence



Inertial methods can speed up convergence

Inertial methods can speed up convergence:

- Polyak investigates **multi-step methods** in the paper:

  [*Some methods for speeding up the convergence of iteration methods*. Polyak, 1964].

- A m-step method constructs $x^{(k+1)}$ using the previous m iterations $x^{(k)}, \ldots, x^{(k-m+1)}$.

- Gradient descent method is a single-step method.

- Inertial methods are multi-step methods.

- **Heavy-ball method is a 2-step method**.

Heavy-ball method

(Time-discrete) Heavy-ball method:

- Time-continuous dynamical system:

$$\ddot{X}(t) + \gamma\dot{X}(t) + \nabla f(X(t)) = 0\,.$$

- Discretization yields:

$$\begin{aligned} 0 &= \frac{x^{(k+1)} - 2x^{(k)} + x^{(k-1)}}{\tau^2} + \gamma\frac{x^{(k+1)} - x^{(k)}}{\tau} + \nabla f(x^{(k)}) \\ \Leftrightarrow 0 &= (1+\tau\gamma)x^{(k+1)} - (\tau\gamma + 2)x^{(k)} + x^{(k-1)} + \tau^2\nabla f(x^{(k)}) \\ \Leftrightarrow 0 &= (1+\tau\gamma)x^{(k+1)} - (\tau\gamma + 1)x^{(k)} - (x^{(k)} - x^{(k-1)}) + \tau^2\nabla f(x^{(k)}) \\ \Leftrightarrow 0 &= x^{(k+1)} - x^{(k)} - \frac{1}{1+\tau\gamma}(x^{(k)} - x^{(k-1)}) + \frac{\tau^2}{1+\tau\gamma}\nabla f(x^{(k)}) \end{aligned}$$

- Set $\alpha = \frac{\tau^2}{1+\tau\gamma}$ and $\beta = \frac{1}{1+\tau\gamma}$: (momentum β vs. friction γ)

$$x^{(k+1)} = x^{(k)} - \alpha\nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\,.$$

(Time-discrete) Heavy-ball method:

- **Update rule:**

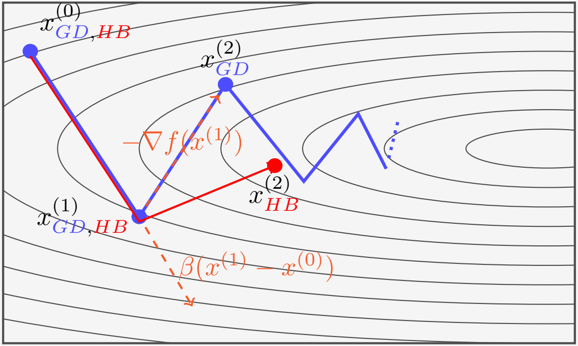
$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\,.$$

- $(x^{(k)})_{k\in\mathbb{N}}$: sequence of iterates.
- $\alpha > 0$: step size parameter.
- $\beta \in [0,1)$: inertial parameter.
- For $\beta = 0$, we recover the gradient descent method.
- **Optimal** for strongly convex functions [Polyak 1964]

$$|x^{(k+1)} - x^\star|^2 \leq c q^{2k} |x^{(0)} - x^\star|^2\,, \quad q_{\text{HB}} := \frac{\sqrt{L} - \sqrt{l}}{\sqrt{L} + \sqrt{l}}\,.$$

Some properties:

- It is not a classical descent method.
- It avoids zick-zacking.
- Similarity to conjugate gradient method.



Accelerated Gradient Descent

Nesterov's Accelerated Gradient Method: f convex

- **A differential equations:**

$$\ddot{X}(t) + \frac{\rho}{t}\dot{X}(t) + \nabla f(X(t)) = 0\,.$$

[Su, Boyd, Candès, 2015] [Attouch, Peypouquet, Redont 2015]

- For $\rho > 3$: any trajectory converges weakly to a minimizer.
- Convergence rate: $\mathcal{O}(1/t^2)$. (actually $\mathrm{o}(1/t^2)$ [Attouch, Peypouquet 2016].)
- From overdamping to underdamping.
- Studied before in the following context: [Cabot, Engler, Gadat 2009]

$$\ddot{X}(t) + g(t)\dot{X}(t) + \nabla f(X(t)) = 0\,.$$

Accelerated Gradient Descent

Nesterov's Accelerated Gradient Method:

- **Update step:**

$$x^{(k+1)} = y^{(k)} - \tau \nabla f(y^{(k)})$$
$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$y^{(k+1)} = x^{(k+1)} + \frac{t_k - 1}{t_{k+1}}(x^{(k+1)} - x^{(k)})$$

- [Nesterov, 1983]: $f \in C_L^{1,1}$ convex, **optimal method**

$$f(x^{(k)}) - f^\star \leq \frac{4L|y^{(0)} - x^\star|^2}{(k+2)^2}$$

- In the setting of Forward–Backward Splitting: **FISTA** [Beck, Teboulle 2009].

Optimized Accelerated Gradient Descent

Adaptive FISTA: [O., Pock, 2017]

- **Update step:**

$$y^{(k)}(\beta) = x^{(k)} + \beta(x^{(k)} - x^{(k-1)})$$
$$x^{(k+1)} = \operatorname*{argmin}_x \min_\beta f^L(x; y^{(k)}(\beta))$$

- $f^L(x; y^{(k)}(\beta))$: quadratic approximation of f around $y^{(k)}(\beta)$.

- If f is quadratic, equivalent to (*details later*)

$$x^{(k+1)} = x^{(k)} - M^{-1}\nabla f(x^{(k)}) \qquad (\textbf{Quasi-Newton step})$$

  with positive definite M (rank-1 modification of a diagonal matrix)

- **Quasi-Newton Methods** are also accelerations of Gradient Descent.
  - For example: BFGS, DFP, SR1, ...
  - try to approximate Newton's method (quadratic convergence).
  - Some Quasi-Newton Methods converge superlinearly.

Subspace Acceleration Methods:

- **Update step:**

$$x^{(k+1)} = x^{(k)} + D^{(k)}s^{(k)}\,, \quad D^{(k)} = (d_1^{(k)}, \dots, d_M^{(k)})\,,\; d_i^{(k)} \in \mathbb{R}^N\,.$$

- $s^{(k)} \in \mathbb{R}^M$ is a multi-dimensional step size that aims at minimizing

$$s \mapsto f(x^{(k)} + D^{(k)}s)\,.$$

- First such algorithm: **Memory Gradient Method** [Miele, Cantrell 1960's]

$$D^{(k)} = (-\nabla f(x^{(k)}), d^{(k-1)})\,, \qquad s^{(k)} \text{ by exact minimization}\,.$$

- L-BFGS quasi-Newton method: subspace of size $2m+1$, where m is the limited memory parameter.

- Adaptive FISTA tries to minimize w.r.t. the overrelaxation parameter β.

Construction of Subspaces

Acronym	Algorithm	Set of directions $\boldsymbol{D}_k$	Subspace size
MG	Memory gradient [23, 31]	$[-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}]$	2
SMG	Supermemory gradient [24]	$[-\boldsymbol{g}_k, \boldsymbol{d}_{k-1}, \dots, \boldsymbol{d}_{k-m}]$	$m+1$
SMD	Supermemory descent [32]	$[\boldsymbol{p}_k, \boldsymbol{d}_{k-1}, \dots, \boldsymbol{d}_{k-m}]$	$m+1$
GS	Gradient subspace [33, 34, 37]	$[-\boldsymbol{g}_k, -\boldsymbol{g}_{k-1}, \dots, -\boldsymbol{g}_{k-m}]$	$m+1$
ORTH	Orthogonal subspace [36]	$[-\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \sum_{i=0}^k w_i \boldsymbol{g}_i]$	3
SESOP	Sequential Subspace Optimization [26]	$[-\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_0, \sum_{i=0}^k w_i \boldsymbol{g}_i, \boldsymbol{d}_{k-1}, \dots, \boldsymbol{d}_{k-m}]$	$m+3$
QNS	Quasi-Newton subspace [20, 25, 38]	$[-\boldsymbol{g}_k, \boldsymbol{\delta}_{k-1}, \dots, \boldsymbol{\delta}_{k-m}, \boldsymbol{d}_{k-1}, \dots, \boldsymbol{d}_{k-m}]$	$2m+1$
SESOP-TN	Truncated Newton subspace [27]	$[\boldsymbol{d}_k^\ell, \boldsymbol{G}_k(\boldsymbol{d}_k^\ell), \boldsymbol{d}_k^\ell - \boldsymbol{d}_k^{\ell-1}, \boldsymbol{d}_{k-1}, \dots, \boldsymbol{d}_{k-m}]$	$m+3$

from [Chouzenoux, Idier, Moussaoui 2011]

Subspace Acceleration Methods

Multi-dimensional step size search via Majorization–Minimization:

- [Chouzenoux, Idier, Moussaoui 2011]
  [Chouzenoux, Jezierska, Pesquet, Talbot 2013]
- Approximate minimization of $s \mapsto f(x^{(k)} + D^{(k)} s)$ by MM procedure.
- Sequentially approximate f by quadratic (tangent majorizers) functions around current trial step size $s^{(k,j)}$ and minimize these quadratic approximations.
- **Yields** monotonically non-increasing objective values, and gradient vanishes.

Accelerations of Forward–Backward Splitting
— Part 3: Non-smooth Optimization —

Peter Ochs
Saarland University
ochs@math.uni-sb.de

— June 11th – 13th, 2018 —

www.mop.uni-saarland.de

Table of Contents

Table of Contents

This part is mainly based on the books of

- [R. T. Rockafellar: *Convex Analysis*. Princeton University Press, 1970.]
- [R. T. Rockafellar, R. J.-B. Wets: *Variational Analysis*. Springer, 1998.]
- [H. H. Bauschke and P. L. Combettes: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.]

Extended real numbers

Definition:

- *Extended real numbers* $\overline{\mathbb{R}} := [-\infty, +\infty]$

$$\begin{array}{rcll}
a + (+\infty) = +\infty + a & = & +\infty & \text{for} \quad -\infty < a \leq +\infty \\
a + (-\infty) = -\infty + a & = & -\infty & \text{for} \quad -\infty \leq a < +\infty \\
a(+\infty) = (+\infty)a & = & +\infty & \text{for} \quad 0 < a \leq +\infty \\
a(-\infty) = (-\infty)a & = & -\infty & \text{for} \quad 0 < a \leq +\infty \\
a(+\infty) = (+\infty)a & = & -\infty & \text{for} \quad -\infty \leq a < 0 \\
a(-\infty) = (-\infty)a & = & +\infty & \text{for} \quad -\infty \leq a < 0 \\
0(\pm\infty) = (\pm\infty)0 & = & 0 & \\
-(-\infty) & = & +\infty & \\
\inf \emptyset & = & +\infty & \\
\sup \emptyset & = & -\infty &
\end{array}$$

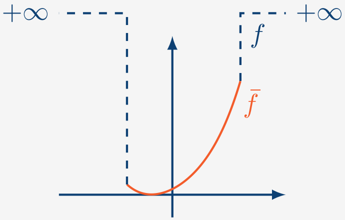
- Operations $+\infty + (-\infty)$ and $-\infty + (+\infty)$ are **not** defined.
- Familiar laws of arithmetic, if all binary operations are well-defined:

$$a + b = b + a\,, \quad (a + b) + c = a + (b + c)\,,$$
$$ab = ba\,, \quad (ab)c = a(bc)\,, \qquad a(b + c) = ab + ac$$

Extended real numbers

- Extend functions $\bar{f}\colon C \to \mathbb{R}$ with $C \subset \mathbb{R}^N$ to the whole space $\mathbb{R}^N$ by

$$f(x) = \begin{cases} \bar{f}(x), & \text{if } x \in C\,; \\ +\infty, & \text{otherwise.} \end{cases}$$



- **Definition:**

  A function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is called *proper*, if

$$\begin{cases} f(x) < +\infty \text{ for at least one } x \in \mathbb{R}^N \text{ and} \\ f(x) > -\infty \text{ for all } x \in \mathbb{R}^N\,, \end{cases}$$

  and *improper* otherwise.

Domain, Epigraph, and Level Sets

Definition:

- The *(effective) domain* is the set

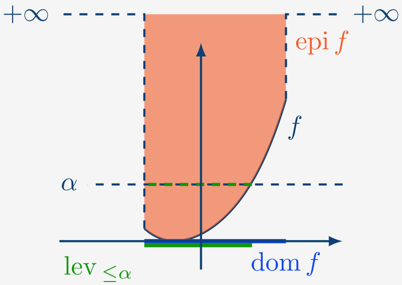
$$\operatorname{dom} f := \{x \in \mathbb{R}^N \,|\, f(x) < +\infty\}\,.$$

- The *epigraph* is the set

$$\operatorname{epi} f := \{(x,\alpha) \in \mathbb{R}^N \times \mathbb{R} \,|\, \alpha \geq f(x)\}\,.$$

- The *lower level set* is the set

$$\operatorname{lev}_{\leq\alpha} f := \{x \in \mathbb{R}^N \,|\, f(x) \leq \alpha\}\,.$$



Definition:

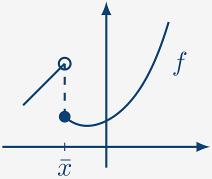
- The *lower limit* of a function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ at $\bar x$ is the value in $\overline{\mathbb{R}}$ defined by

$$\liminf_{x\to\bar x} f(x) := \lim_{\delta\searrow 0}\left[\inf_{x\in B_\delta(\bar x)} f(x)\right] = \sup_{\delta>0}\left[\inf_{x\in B_\delta(\bar x)} f(x)\right].$$

- $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is *lower semi-continuous (lsc) at $\bar x$* if

$$\liminf_{x\to\bar x} f(x) \geq f(\bar x),$$

and *lsc on $\mathbb{R}^N$* if this holds for every $\bar x$.



lsc / not usc

Theorem: (Characterization of lower semi-continuity)

The following properties of a function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ are equivalent:

(a) f is lower semi-continuous on $\mathbb{R}^N$.

(b) The epigraph $\operatorname{epi} f$ is closed in $\mathbb{R}^N \times \mathbb{R}$.

(c) The level sets of type $\operatorname{lev}_{\leq\alpha} f$ are all closed in $\mathbb{R}^N$.

Definition:

A function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is (lower) *level-bounded*, if for every $\alpha \in \mathbb{R}$ the set $\operatorname{lev}_{\leq\alpha} f$ is bounded (possibly empty).

Theorem: (Attainment of minimizers)

Suppose $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is lsc, level-bounded, and proper. Then the value $\inf_{x\in\mathbb{R}^N}\ f(x)$ is finite and the set $\arg\min_{x\in\mathbb{R}^N}\ f(x)$ is nonempty and compact.

Definition

The *infimal convolution* (or inf-convolution) is defined by

$$(f \,\Box\, g)(x) := \inf_{w \in \mathbb{R}^N} f(x-w) + g(w) = \inf_{w \in \mathbb{R}^N} f(w) + g(x-w)\,.$$

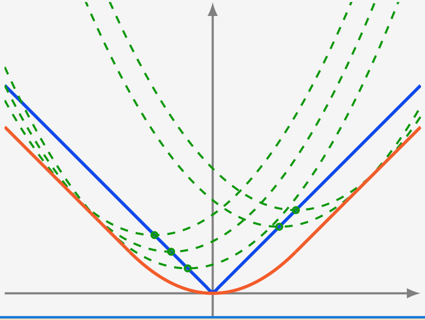
- $f \,\Box\, g$ is the point-wise infimum of functions $h_w(x) = f(w) + g(x-w)$.
- $\mathrm{epi}\,(f \,\Box\, g) = \mathrm{epi}\, f + \mathrm{epi}\, g$, if the infimum in $f \,\Box\, g$ is attained when finite.

Example:

Let $f(x) = |x|$ and $g(x) = \frac{1}{2\lambda}|x|^2$.

$$(f \,\Box\, g)(x) = \inf_{w \in \mathbb{R}^N} |w| + \frac{1}{2\lambda}|x-w|^2$$

$$= \begin{cases} \frac{1}{2\lambda}x^2\,, & \text{if } |x| \le \lambda \\ |x| - \frac{\lambda}{2}\,, & \text{otherwise.} \end{cases}$$



Moreau envelope and proximal mapping

Definition:

For a proper, lsc function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ and parameter value $\lambda > 0$ the *Moreau envelope* function $e_\lambda f$ and the *proximal mapping* $\mathrm{prox}_{\lambda f}$ are defined by

$$e_\lambda f(x) := \inf_{w\in\mathbb{R}^N} f(w) + \frac{1}{2\lambda}|w-x|^2$$
$$\mathrm{prox}_{\lambda f}(x) := \arg\min_{w\in\mathbb{R}^N} f(w) + \frac{1}{2\lambda}|w-x|^2$$

Remark:

In general, $e_\lambda f$ is extended-valued, and $\mathrm{prox}_{\lambda f}$ is set-valued.

Example:

Let $\emptyset \neq C \subset \mathbb{R}^N$ be a closed convex set and δ_C the associated indicator function. Then, for any $\bar{x} \in \mathbb{R}^N$ and $\lambda > 0$, it holds that

$$\mathrm{prox}_{\lambda\delta_C}(\bar{x}) = \operatorname*{argmin}_{x\in C} \frac{1}{2\lambda}|x-\bar{x}|^2 = \mathrm{proj}_C(\bar{x})\,.$$

Calculation Rules for the Proximal Mapping:

Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ and $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be proper, lsc functions and $b \in \mathbb{R}$.

- If $f(x,y) = f_1(x) + f_2(y)$, then $\mathrm{prox}_{\lambda f}(x,y) = (\mathrm{prox}_{\lambda f_1}(x), \mathrm{prox}_{\lambda f_2}(y))$.
- If $f(x) = \alpha g(x) + b$ with $\alpha > 0$, then $\mathrm{prox}_f(x) = \mathrm{prox}_{\alpha g}(x)$.
- If $f(x) = g(\alpha x + b)$ with $\alpha \neq 0$, then $\mathrm{prox}_f(x) = \frac{1}{\alpha}(\mathrm{prox}_{\alpha^2 g}(\alpha x + b) - b)$.
- If $f(x) = g(Qx)$ with Q orthogonal (such that $Q^\top Q = Q^\top Q = \mathrm{id}$), then
$$\mathrm{prox}_f(x) = Q^\top \mathrm{prox}_g(Qx)\,.$$
- If $f(x) = g(x) + \langle a, x\rangle + b$ with $a \in \mathbb{R}^N$, then $\mathrm{prox}_f(x) = \mathrm{prox}_g(x - a)$.
- If $f(x) = g(x) + \frac{\gamma}{2}|x - a|^2$ with $\gamma > 0$ and $a \in \mathbb{R}^N$, then
$$\mathrm{prox}_f(x) = \mathrm{prox}_{\tilde\gamma g}(\tilde\gamma x + \tilde\gamma\gamma a)$$
with $\tilde\gamma := 1/(1+\gamma)$.

Examples for the Proximal Mapping:

- $f(x) = \frac{\lambda}{2}|x|^2$:

$$\mathrm{prox}_{\tau f}(\bar{x}) = \operatorname*{argmin}_{x \in \mathbb{R}^N} \frac{\tau\lambda}{2}|x|^2 + \frac{1}{2}|x - \bar{x}|^2$$

Optimality condtion:

$$\tau\lambda x + (x - \bar{x}) = 0 \quad \Leftrightarrow \quad x = \frac{\bar{x}}{1 + \tau\lambda}\,.$$

- **Nuclear norm**: $f(X) = \|X\|_* := \sum_{i=1}^N \sigma_i$ with SVD

$$X = U\mathrm{diag}(\sigma_1, \ldots, \sigma_N)V^\top \qquad \sigma_i \geq 0\,.$$

We can show that $(g(\sigma_i) = \sigma_i + \delta_{[\sigma_i \geq 0]}(\sigma_i))$

$$\mathrm{prox}_{\tau f}(\bar{X}) = U\mathrm{diag}\big([\mathrm{prox}_{\tau g}(\bar{\sigma}_i)]_{i=1}^N\big)V^\top \qquad \text{with } \bar{X} = U\mathrm{diag}(\bar{\sigma}_1, \ldots, \bar{\sigma}_N)V^\top$$

and

$$\mathrm{prox}_{\tau g}(\bar{\sigma}_i) = \operatorname*{argmin}_{\sigma_i \geq 0} \tau\sigma_i + \frac{1}{2}(\sigma_i - \bar{\sigma}_i)^2 = \max(0, \bar{\sigma}_i - \tau)\,.$$

Generalized Projection Theorem

Theorem: (Generalized Projection Theorem)
Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be lsc, proper, and **convex**, and $x \in \mathbb{R}^N$, $\lambda > 0$. Then, $\mathrm{prox}_{\lambda f}(x) \in \mathbb{R}^N$ is the unique point that satisfies

$$e_\lambda f(x) = f(\mathrm{prox}_{\lambda f}(x)) + \frac{1}{2\lambda}|\mathrm{prox}_{\lambda f}(x) - x|^2 \,.$$

Moreover,

$$p = \mathrm{prox}_{\lambda f}(x) \quad \Leftrightarrow \quad \forall y \in \mathbb{R}^N\colon\ \langle x - p, y - p\rangle + \lambda f(p) \leq \lambda f(y)\,.$$

The envelope function $e_\lambda f$ is continuously differentiable and

$$\nabla e_\lambda f(x) = \frac{1}{\lambda}(x - \mathrm{prox}_{\lambda f}(x))$$

is λ^{-1}-Lipschitz continuous.

The same formula holds locally, for **prox-regular** functions. ($\rightsquigarrow$ *later*)

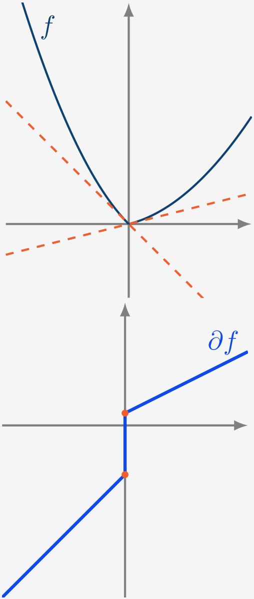
Definition:

- Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be **convex.**
- v is a **subgradient** of f at $\bar{x}$, i.e. $v \in \partial f(\bar{x})$, **if** the following holds:
  subgradient inequality:

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x} \rangle \,, \quad \forall x \in \mathbb{R}^N$$

- **Subdifferential** $\partial f\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ (set-valued mapping) of f given by

$$\operatorname{Graph} \partial f := \{(x, v) \in \mathbb{R}^N \times \mathbb{R}^N \,|\, v \in \partial f(x)\}$$



Set-valued mapping

Definition:

A *set-valued mapping* $F\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^M$ is a mapping that maps each $x \in \mathbb{R}^N$ to a subset of $\mathbb{R}^M$. The graph of the mapping F is given by

$$\operatorname{Graph} F := \{(x, u) \in \mathbb{R}^N \times \mathbb{R}^M \,|\, u \in F(x)\} \subset \mathbb{R}^N \times \mathbb{R}^M \,.$$

For a set-valued mapping the *(effective) domain* is defined by

$$\operatorname{dom} F := \{x \in \mathbb{R}^N \,|\, F(x) \neq \emptyset\} \subset \mathbb{R}^N \,.$$

Subgradients for nonconvex functions

Definition:

- Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be a function and $\bar x$ a point with $f(\bar x)$ finite.
- v is a **regular subgradient** of f at $\bar x$, i.e. $v \in \widehat{\partial} f(\bar x)$, if

$$\liminf_{\substack{x\to\bar x\\ x\neq\bar x}} \frac{f(x) - f(\bar x) - \langle x-\bar x, v\rangle}{|x-\bar x|} \geq 0$$

$$\Big(\Leftrightarrow f(x) \geq f(\bar x) + \langle v, x-\bar x\rangle + \mathrm{o}(|x-\bar x|) \Big)\,.$$

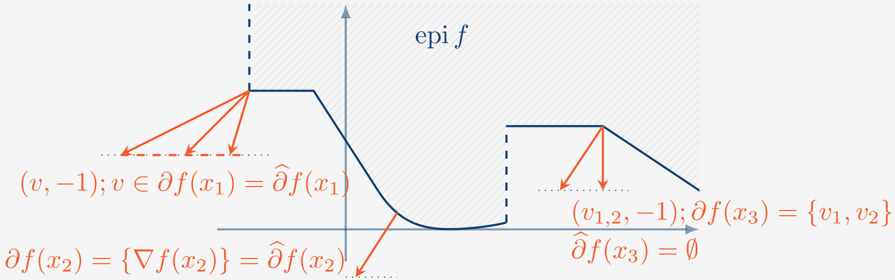
- v is a **(limiting) subgradient** of f at $\bar x$, i.e. $v \in \partial f(\bar x)$, if

$$\exists\, x^\nu \to \bar x\colon\; f(x^\nu) \to f(\bar x),\, v^\nu \to v,\, v^\nu \in \widehat{\partial} f(x^\nu)$$

- v is a **horizon subgradient** of f at $\bar x$, i.e. $v \in \partial^\infty f(\bar x)$, if

$$\exists\, x^\nu \to \bar x,\, \lambda^\nu \searrow 0\colon\; f(x^\nu) \to f(\bar x),\, \lambda^\nu v^\nu \to v,\, v^\nu \in \widehat{\partial} f(x^\nu)$$

Example: (Subgradients for nonconvex functions)



Properties:

- f differentiable at $\bar{x}$, then $\widehat{\partial} f(\bar{x}) = \{\nabla f(\bar{x})\}$, and $\nabla f(\bar{x}) \in \partial f(\bar{x})$.
- f smooth in a neighborhood of $\bar{x}$, then $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$.
- f proper, convex, then $\widehat{\partial} f(\bar{x}) = \partial f(\bar{x})$.

Examples for the Subdifferential

Example:

- The subdifferential of $f\colon \mathbb{R}^N \to \mathbb{R}$, $x \mapsto \frac{1}{2}|x|^2$ is given by

$$\partial f(x) = \{x\}\,.$$

- The subdifferential of $|\cdot|$ in $\mathbb{R}^N$ is

$$\partial|\cdot|(x) = \begin{cases} \left\{\dfrac{x}{|x|}\right\}, & \text{if } x \neq 0\,; \\ B_1(0), & \text{if } x = 0\,. \end{cases}$$

- The subdiffferential of $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto \sqrt{|x|}$ is given by

$$\widehat{\partial}\sqrt{|\cdot|}(x) = \partial\sqrt{|\cdot|}(x) = \begin{cases} \{\frac{1}{2\sqrt{x}}\}\,, & \text{if } x > 0\,; \\ \{\frac{-1}{2\sqrt{-x}}\}\,, & \text{if } x < 0\,; \\ (-\infty, \infty)\,, & \text{if } x = 0\,. \end{cases}$$

Proposition: (Subdifferential Calculus)

- If $f(x) = f_1(x_1) + f_2(x_2)$ with $x = (x_1, x_2)$, then
$$\widehat{\partial} f(x) = \widehat{\partial} f_1(x_1) \times \widehat{\partial} f_2(x_2) \quad \text{and} \quad \partial f(x) = \partial f_1(x_1) \times \partial f_2(x_2)\,.$$
- If $f = f_1 + f_2$ with proper lsc functions f_1 and f_2 and $\bar{x} \in \operatorname{dom} f$, then
$$\widehat{\partial} f(\bar{x}) \supset \widehat{\partial} f_1(\bar{x}) + \widehat{\partial} f_2(\bar{x})\,.$$
If the only combination of $v_i \in \partial^\infty f_i(\bar{x})$ with $v_1 + v_2 = 0$ is $v_1 = v_2 = 0$, then
$$\partial f(\bar{x}) \subset \partial f_1(\bar{x}) + \partial f_2(\bar{x})\,.$$
If each f_i is regular at $\bar{x}$, i.e. $\widehat{\partial} f(\bar{x}) = \partial f(\bar{x})$, then
$$\partial f(\bar{x}) = \partial f_1(\bar{x}) + \partial f_2(\bar{x})\,.$$
- If $f = f_1 + f_2$ with f_1 finite at $\bar{x}$ and f_2 smooth on a neighborhood of $\bar{x}$, then
$$\widehat{\partial} f(\bar{x}) = \widehat{\partial} f_1(\bar{x}) + \nabla f_2(\bar{x}) \quad \text{and} \quad \partial f(\bar{x}) = \partial f_1(\bar{x}) + \nabla f_2(\bar{x})\,.$$

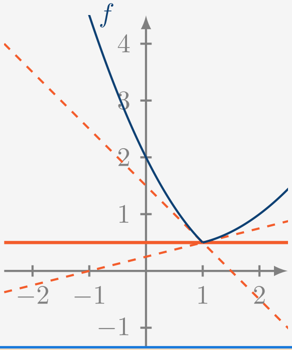
Optimality condition: Fermat's rule

Theorem: (Fermat's Rule)
Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper functions with a local minimum at $\bar{x}$, then

$$0 \in \partial f(\bar{x})\,.$$

If f is convex, then

$$\bar{x} \in \underset{x\in\mathbb{R}^N}{\operatorname{argmin}}\; f(x) \quad \Leftrightarrow \quad 0 \in \partial f(\bar{x})\,.$$



Smooth Minimization with Geometric Constraint

Smooth Minimization with Geometric Constraint:

- $f\colon \mathbb{R}^N \to \mathbb{R}$ continuously differentiable and $\emptyset \neq C \subset \mathbb{R}^N$ be a closed set.
- Then, we have the following necessary optimality condition

$$
\begin{aligned}
0 \in \partial(f + \delta_C)(x) = \nabla f(x) + \partial\delta_C(x) =: \nabla f(x) + N_C(x) \\
\Leftrightarrow \quad -\nabla f(x) \in N_C(x)\,.
\end{aligned}
$$

Example:

For $C = [0, +\infty)^N$, we have

$$
(N_C(x))_i = \begin{cases} (-\infty, 0]\,, & \text{if } x_i = 0 \\ 0 & \text{otherwise.} \end{cases}
$$

or $(N_C(x))_i = \{v_i \,:\, x_i \geq 0 \text{ and } v_i \leq 0 \text{ and } x_i v_i = 0\}$.
Therefore, $-\nabla f(x) \in N_C(x)$ is equivalent to the *complementary condition*:

$$
(\nabla f(x))_i \geq 0\,, \quad x_i \geq 0\,, \quad \text{and} \quad (\nabla f(x))_i x_i = 0\,.
$$

Example: Fermat's Rule

- Compute $\operatorname{prox}_{\tau f}(\bar{x})$ for $f(x) = |x|$.
- Can be computed coordinate-wise. Thus, w.l.o.g. $x \in \mathbb{R}^1$.
- Optimality condition of $\min_x \; \tau|x| + \frac{1}{2}(x - \bar{x})^2$:

$$
\begin{aligned}
& 0 \in \tau\partial|\cdot|(x) + x - \bar{x} \\
\Leftrightarrow \;\; & x = \bar{x} - \partial|\cdot|(x) = \begin{cases} \bar{x} - \tau & \text{if } x > 0 \; (\Leftrightarrow \bar{x} > \tau)\,; \\ \bar{x} + \tau & \text{if } x < 0 \; (\Leftrightarrow \bar{x} < -\tau)\,; \\ \bar{x} - \tau[-1, 1] & \text{if } x = 0 \; (\Leftrightarrow \bar{x} \in [-\tau, \tau])\,. \end{cases}
\end{aligned}
$$

- The solution is the **Soft Shrinkage-Thresholding Operator**:

$$\operatorname{prox}_{\tau f}(\bar{x}) = \max(0, |\bar{x}| - \tau)\operatorname{sign}(\bar{x})\,.$$

An Algorithm for Non-smooth Functions: (Convex Optimization)

- Return to the **gradient dynamical system**:

$$\dot{X}(t) + \nabla f(X(t)) = 0 \,.$$

- **Explicit** discretization yields **Gradient Descent**: (aka. forward step)

$$\frac{x^{(k+1)} - x^{(k)}}{\tau_k} + \nabla f(x^{(k)}) = 0 \quad \Leftrightarrow \quad x^{(k+1)} = (\mathrm{id} - \tau_k \nabla f)(x^{(k)}) \,.$$

- **Implicit** discretization yields **Proximal Algorithm**: (aka. backward step)

$$\frac{x^{(k+1)} - x^{(k)}}{\tau_k} + \nabla f(x^{(k+1)}) = 0 \quad \Leftrightarrow \quad (\mathrm{id} + \tau_k \nabla f)(x^{(k+1)}) = x^{(k)} \,.$$

Proximal Algorithm / Proximal Point Algorithm

- **Proximal Algorithm** can be written as

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} f(x) + \frac{1}{2\tau_k} |x - x^{(k)}|^2 \,.$$

- **Optimality condition**:

$$0 = \nabla f(x) + \frac{1}{\tau_k}\left(x - x^{(k)}\right) \quad \Leftrightarrow \quad (\mathrm{id} + \tau_k \nabla f)x = x^{(k)} \,.$$

- The proximal algorithm does not require f to be differentiable.

- **Optimality condition**: (f proper, lsc)

$$0 \in \partial f(x) + \frac{1}{\tau_k}\left(x - x^{(k)}\right) = 0 \quad \Leftrightarrow \quad x^{(k)} \in (\mathrm{id} + \tau_k \partial f)x$$

$$\overset{f \text{ convex}}{\Leftrightarrow} \quad x = (\mathrm{id} + \tau_k \partial f)^{-1}(x^{(k)}) \,.$$

Proximal Point Algorithm (PPA)

Algorithm: (Proximal Minimization Algorithm)

- **Optimization problem**: $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ proper, lsc
- **Iterations** $(k \geq 0)$: Update ($x^{(0)} \in \mathbb{R}^N$)

$$x^{(k+1)} \in \operatorname{prox}_{\tau_k f}(x^{(k)}) = \arg\min_{w\in\mathbb{R}^N}\; f(w) + \frac{1}{2\tau_k}|w - x^{(k)}|^2$$

- **Parameter setting**: $\tau_k > 0$ step size parameter.

- Very general (**conceptual**) algorithm.
- Note that a single iteration is usually as hard as solving the original problem.
- In a more general form, it applies to **maximal monotone operators**. See [Rockafellar 1976].
- Many algorithms are actually special cases of the proximal point algorithm.

Structured Optimization Problems: (**Splitting**)

- Common Structure in Applications:

$$\min_{x\in\mathbb{R}^N} \; f(x) \; + \; g(x)$$

$\mathbb{R}^N \to \mathbb{R}$
smooth
∇f Lipschitz

$\mathbb{R}^N \to \overline{\mathbb{R}}$
non-smooth
simple prox

- **Lasso, Group Lasso, ...**:

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}|Ax-b|^2 + \lambda\|x\|_1 \quad \text{or} \quad \min_{x\in\mathbb{R}^N} \frac{1}{2}|Ax-b|^2 \quad s.t.\ \|x\|_1 \leq \lambda\,.$$

- **Non-negative Least Squares:**

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}|Ax-b|^2 \quad s.t.\ x_i \geq 0\ \forall i = 1,\dots,N\,.$$

Applications of Forward–Backward Splitting

- **Logistic Regression:**

$$\min_{w\in\mathbb{R}^N} \log(1+\exp(-y_i \langle x_i, w\rangle)) + \lambda\|w\|_1 \,.$$

- **Low Rank Approximation:** (e.g. Matrix completion)

$$\min_{X\in\mathbb{R}^{M\times N}} \frac{1}{2}\|A-X\|_F^2 + \lambda\|X\|_* \,.$$

- **Regularized Non-linear Regression:**

$$\min_{w\in\mathbb{R}^N} \frac{1}{2}\sum_{i=1}^{M} |\mathcal{N}_w(x_i) - y_i|^2 + \lambda g(w) \,.$$

- **Feasibility Problem:** Find $x \in C\cap D$ for closed set $C \neq \emptyset$ and a closed convex set $D \neq \emptyset$.

$$\min_{x\in\mathbb{R}^N} e_1\delta_D(x) \quad s.t.\; x\in C \qquad = \min_{x\in C} \operatorname{dist}(x,D)^2$$

Algorithm: (Forward–Backward Splitting (FBS)) (Convex Problem)

- **Optimization problem**: $\min_x \; f(x) + g(x)$
  - $f\colon \mathbb{R}^N \to \mathbb{R}$ continuously differentiable, convex, with ∇f L-Lipschitz.
  - $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ proper, lsc, convex with simple proximal mapping.
- **Iterations** $(k \geq 0)$: Update $(x^{(0)} \in \mathbb{R}^N)$, $\varepsilon \leq \tau_k \leq \frac{2-\varepsilon}{L}$ for some $\varepsilon > 0$:

$$x^{(k+1)} = \mathrm{prox}_{\tau_k g}(x^{(k)} - \tau_k \nabla f(x^{(k)}))$$

Proposition: [Combettes, Pesquet 2011], [Combettes, Wajs 2005]
If $f + g$ is coercive, then any sequence generated by **FBS converges to a solution of** $\min_x f + g$.

Method traces back to:
[P. L. Lions and B. Mercier: *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.]

Forward–Backward Splitting

Naming:

$$x^{(k+1)} = \underbrace{\mathrm{prox}_{\tau_k g}}_{\text{backward step}} \underbrace{(x^{(k)} - \tau_k \nabla f(x^{(k)}))}_{\text{forward step}}$$

- Other frequently used name: **Proximal Gradient Descent**.

Equivalent update rules:

$$\begin{aligned}
x^{(k+1)} &= \mathrm{prox}_{\tau_k g}(x^{(k)} - \tau_k \nabla f(x^{(k)})) \\
&= (\mathrm{id} + \tau_k \partial g)^{-1}\big(x^{(k)} - \tau_k \nabla f(x^{(k)})\big) \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^N} g(x) + f(x^{(k)}) + \Big\langle \nabla f(x^{(k)}), x - x^{(k)} \Big\rangle + \frac{1}{2\tau_k}|x - x^{(k)}|^2 \\
&= x^{(k)} - \tau_k \Big[\frac{1}{\tau_k}\Big(x^{(k)} - \mathrm{prox}_{\tau_k g}\big(x^{(k)} - \tau_k \nabla f(x^{(k)})\big)\Big)\Big] \\
&= (\mathrm{id} - \tau_k \nabla e_{\tau_k} g)(\mathrm{id} - \tau_k \nabla f)(x^{(k)})
\end{aligned}$$

Accelerations of Forward–Backward Splitting
— Part 4: Single Point Convergence —

Peter Ochs
Saarland University
ochs@math.uni-sb.de

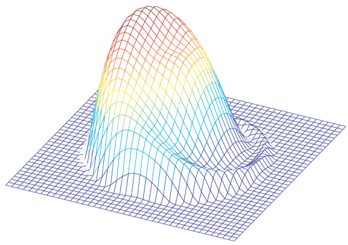
— June 11th – 13th, 2018 —

www.mop.uni-saarland.de

Table of Contents

4. Single Point Convergence

- Łojasiewicz Inequality
- Kurdyka-Łojasiewicz Inequality
- Abstract Convergence Theorem
- Convergence of Non-convex Forward–Backward Splitting
- A Generalized Abstract Convergence Theorem
- Convergence of iPiano
- Local Convergence of iPiano



counterexample for convergence to a single point for Gradient Descent

Łojasiewicz and smooth Kurdyka-Łojasiewicz inequality

Theorem:[[Łojasiewicz, 1963]]

Let $f\colon U \subset \mathbb{R}^N \to \mathbb{R}$ be a **real analytic**, U open, and $\hat{x} \in U$ a critical point of f. Then, there exists $\theta \in [\frac{1}{2}, 1)$, $C > 0$, and a neighbourhood W of $\hat{x}$ such that

$$\forall x \in W : \qquad |f(x) - f(\hat{x})|^\theta \le C|\nabla f(x)|\,.$$

- Equivalent formulation: $\varphi(s) := cs^{1-\theta}$ (**desingularization function**)

$$\varphi'(f(x) - f(\hat{x}))|\nabla f(x)| \ge 1\,,$$

- or (assume $f(\hat{x}) = 0$)

$$|\nabla(\varphi \circ f)(x)| \ge 1$$

Łojasiewicz Inequality and Gradient System

- Let $X\colon [0,+\infty) \to W$ be a gradient trajectory (i.e. $\dot X(t) = -\nabla f(X(t))$).

  Lyapunov function: $h(t) := \varphi(f(X(t)) - f(\hat X))$ ($\hat X$ limit point of X) .

- $\dot h(t) = \varphi'(f(X(t)) - f(\hat X)) \left\langle \nabla f(X(t)), \dot X(t) \right\rangle$.

- **Lyapunov property** (non-increasingness along the trajectory):

$$\begin{aligned}
\dot h(t) + |\dot X(t)| &= \dot h(t) + |\nabla f(X(t))| \\
&= \dot h(t) + |\nabla f(X(t))|^{-1} |\nabla f(X(t))|^2 \\
&\leq \dot h(t) + \varphi'(f(X(t)) - f(\hat X)) \left\langle \nabla f(X(t)), -\dot X(t) \right\rangle = 0\,.
\end{aligned}$$

- This yields $\dot X \in L^1(0,+\infty)$:

$$\begin{aligned}
\mathrm{length}(X) = \int_0^{+\infty} |\dot X(t)|\, dt &\leq h(0) - \lim_{t\to+\infty} h(t) \\
&= \varphi(f(X(0)) - f(\hat X)) < +\infty\,.
\end{aligned}$$

Definition:

The lsc function $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ has the KL property at $\hat{x} \in \operatorname{dom}\partial f$, **if**

- there exists $\eta \in (0, +\infty]$,
- a neighborhood U of $\hat{x}$,
- and a continuous concave function $\varphi\colon [0, \eta) \to \mathbb{R}_+$ with

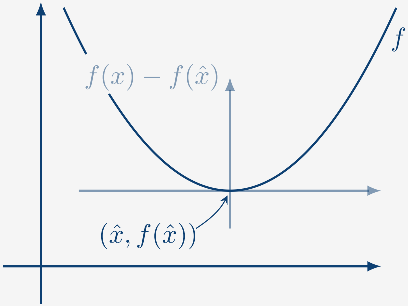
$$\begin{cases} \varphi(0) = 0 \\ \varphi \in C^1((0,\eta)) \\ \varphi'(s) > 0 \ \text{ for all } \ s \in (0, \eta) \end{cases}$$

such that the **(non-smooth) Kurdyka-Łojasiewicz inequality**

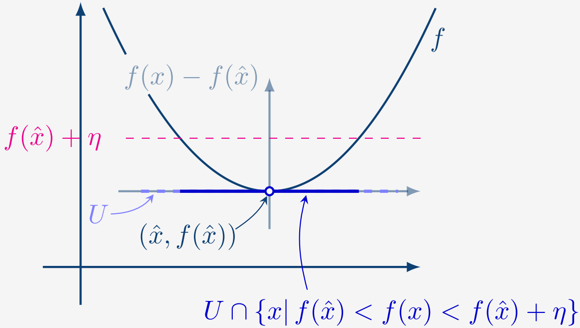
$$\varphi'(f(x) - f(\hat{x}))\operatorname{dist}(0, \partial f(x)) \geq 1$$

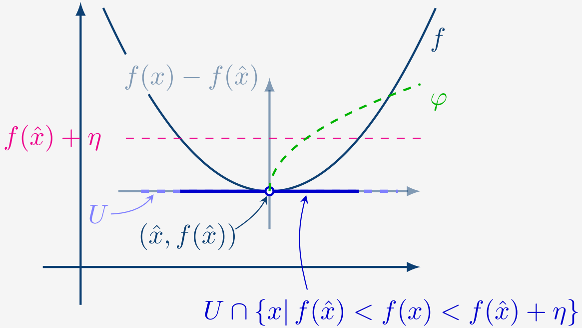
holds, for all $x \in U \cap \{x \in \mathbb{R}^N : f(\hat{x}) < f(x) < f(\hat{x}) + \eta\}$.

KL inequality

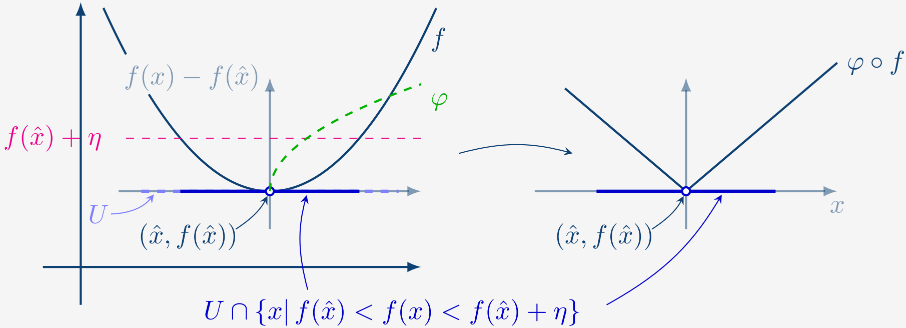


KL inequality





f
$f(x) - f(\hat{x})$
φ
$f(\hat{x}) + \eta$
U
$(\hat{x}, f(\hat{x}))$
$U \cap \{x \mid f(\hat{x}) < f(x) < f(\hat{x}) + \eta\}$



f
$f(x) - f(\hat{x})$
φ
$f(\hat{x}) + \eta$
U
$(\hat{x}, f(\hat{x}))$
$\varphi \circ f$
x
$(\hat{x}, f(\hat{x}))$
$U \cap \{x|\, f(\hat{x}) < f(x) < f(\hat{x}) + \eta\}$

What functions have the KL property?

- Real analytic functions [Łojasiewicz '63]
- Differentiable functions definable in an $\mathcal{o}$-minimal structure [Kurdyka '98]
- Non-smooth lsc functions definable in an $\mathcal{o}$-minimal structure
  - Clarke subgradients [Bolte, Daniilidis, Lewis, Shiota 2007]
  - Limiting subgradients [Attouch, Bolte, Redont, Soubeyran 2010]

$\rightsquigarrow$ **nearly any function in practice**

(*excludes many pathological cases.*)

What functions have the KL property?

Theorem: [Bolte, Daniilidis, Lewis, Shiota 2007]

Any lsc function $f: \mathbb{R}^N \to \overline{\mathbb{R}}$ that is definable in an o-minimal structure $\mathcal{O}$ has the Kurdyka-Łojasiewicz property at each point of $\operatorname{dom} \partial f$. Moreover, the function φ is definable in $\mathcal{O}$.

Examples:

- semi-algebraic functions (*Next slides.*)
  (polynomials, piecewise polynomials, absolute value function, Euclidean distance function, p-norm for $p \in \mathbb{Q}$ (also $p = 0$), ...)

- globally subanlytic functions
  (e.g. $\exp|_{[-1,1]}$)

- $\log$–$\exp$ extension of globally subanalytic structure is an o-minimal structure

- An o-minimal structure is closed under finite sums and products, composition, and several other important operations

Semi-algebraic Functions

Semi-algebraic Structure:

- A set S is **semi-algebraic**, iff there exists polynomials $P_{i,j}$, $Q_{i,j}$ such that
$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{x \in \mathbb{R}^N \,:\, P_{i,j}(x) = 0\,,\; Q_{i,j} < 0\}$$
- $f \colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is semi-algebraic, iff $\mathrm{Graph}\,(f) \subset \mathbb{R}^{N+1}$ is semi-algebraic.
- Finite **union**, **intersection**, **complementary** are again semi-algebraic.
- **Theorem (Tarski-Seidenberg):**
  Canonical **projection** of $S \in \mathbb{R}^{N+1}$ onto $\mathbb{R}^N$ preserves semi-algebraicity.
- **Composition** of semi-algebraic functions: $f = h \circ g$, $\mathbb{R}^N \to \mathbb{R}^M \to \mathbb{R}^L$:
$$\begin{aligned}\mathrm{Graph}\,(f) &= \{(x,z) \in \mathbb{R}^{N\times L} \,:\, z = h(g(x))\} \\ &= \{(x,z) \in \mathbb{R}^{N\times L} \,:\, \exists y \in \mathbb{R}^M \colon\; z = h(y),\, y = g(x)\} \\ &= \Pi_{\mathbb{R}^N \times \mathbb{R}^L}\Big(\{(x,y,z) \,:\, y = g(x)\} \cap \{(x,y,z) \,:\, z = h(y)\}\Big)\end{aligned}$$
- **Desingularization function** of the form $\varphi(s) = cs^{1-\theta}$, $\theta \in [0,1) \cap \mathbb{Q}$.

Definable Functions

Definable Functions: (Axiomatization of the qualitative properties of semi-algebraic sets) [van den Dries, 1998]

Definition:

$\mathcal{O} = \{\mathcal{O}_n\}_{n\in\mathbb{N}}$ is an **o-minimal structure**, if $\mathcal{O}_n$ is a collection of subsets of $\mathbb{R}^n$, and

1. Each $\mathcal{O}_n$ is a boolean algebra: $\emptyset \in \mathcal{O}_n$, $A, B \in \mathcal{O}_n \Rightarrow A \cup B, A \cap B, \mathbb{R}^n \setminus A \in \mathcal{O}_n$.
2. For all $A \in \mathcal{O}_n$, $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{n+1}$.
3. For all $A \in \mathcal{O}_{n+1}$, $\Pi(A) := \{(x_1, \ldots, x_n) \in \mathbb{R}^n \,:\, (x_1, \ldots, x_n, x_{n+1}) \in A\} \in \mathcal{O}_n$.
4. For all $i \neq j$ in $\{1, \ldots, n\}$, $\{(x_1, \ldots, x_n) \in \mathbb{R}^n \,:\, x_i = x_j\} \in \mathcal{O}_n$.
5. The set $\{(x_1, x_2) \in \mathbb{R}^2 \,:\, x_1 < x_2\}$ belongs to $\mathcal{O}_2$.
6. The elements of $\mathcal{O}_1$ are exactly finite unions of intervals.

- A **is definable**, if A belongs to $\mathcal{O}$.
- $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ **is definable**, if $\operatorname{Graph}(f)$ is a definable subset of $\mathbb{R}^{N+1}$.

Single Point Convergence

Single Point Convergence:

- Generalize the result for the gradient trajectory to many other algorithm.
- [Attouch et al. 2013] formulate an abstract descent algorithm.
- Use the (non-smooth) KL inequality.
- Prove a finite length property and single-point convergence.

Abstract descent algorithms: [Attouch et al. 2013]

$$\min_{x \in \mathbb{R}^N} \; f(x)$$

$f : \mathbb{R}^N \to \overline{\mathbb{R}}$ proper, lsc; $a, b > 0$ fixed.
Let $(x^{(k)})_{k \in \mathbb{N}}$ be a sequence that satisfies the following conditions:

(h1) (**Sufficient decrease condition**). For each $k \in \mathbb{N}$,

$$f(x^{(k+1)}) + a|x^{(k+1)} - x^{(k)}|^2 \leq f(x^{(k)})\,;$$

(h2) (**Relative error condition**). For each $k \in \mathbb{N}$,

$$\|\partial f(x^{(k+1)})\|_- \leq b|x^{(k+1)} - x^{(k)}|\,;$$

(h3) (**Continuity condition**). There exists $K \subset \mathbb{N}$ and $\tilde{x}$ such that

$$x^{(k)} \to \tilde{x} \quad \text{and} \quad f(x^{(k)}) \to f(\tilde{x}) \quad \text{as } k \overset{k \in K}{\to} \infty\,.$$

An abstract convergence theorem

Theorem: [Attouch et al. 2013]

- **Let** $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper, lsc.
- **If** $(x^{(k)})_{k\in\mathbb{N}}$ satisfies **(h1)**, **(h2)**, and **(h3)**, i.e.,
  - Sufficient decrease condition,
  - Relative error condition,
  - Continuity condition, and
- f has the Kurdyka-Łojasiewicz property at the cluster point $\tilde{x}$,

then

- $(x^{(k)})_{k\in\mathbb{N}}$ converges to $\bar{x} = \tilde{x}$
- $\bar{x}$ is a critical point of f, i.e., $0 \in \partial f(\bar{x})$, and
- $(x^{(k)})_{k\in\mathbb{N}}$ has a finite length, i.e.,

$$\sum_{k=0}^{\infty} |x^{(k+1)} - x^{(k)}| < +\infty\,.$$

Convergence of Forward–Backward Splitting:

- ∇f is L-Lipschitz, $g\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is proper, lsc., $\inf f + g > -\infty$
- Use this theorem to prove convergence of FBS:

$$x^{(k+1)} \in \underset{x\in\mathbb{R}^N}{\operatorname{argmin}}\; g(x) + f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle + \frac{1}{2\tau}|x - x^{(k)}|^2 .$$

- or an inexact version: Fix $\tau < 1/L$. Find $x^{(k+1)}, v^{(k+1)}$ such that

$$\begin{aligned} & g(x^{(k+1)}) + \left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle + \frac{1}{2\tau}|x^{(k+1)} - x^{(k)}|^2 \leq g(x^{(k)}) \\ & v^{(k+1)} \in \partial g(x^{(k+1)}) \\ & |v^{(k+1)} + \nabla f(x^{(k)})| \leq b|x^{(k+1)} - x^{(k)}| \end{aligned}$$

- Let $(x^{(k)})_{k\in\mathbb{N}}$ be a bounded sequence generated by (inexact) FBS.

Sufficient Decrease Conditions:

- Add update step and Descent Lemma:

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

$$g(x^{(k+1)}) \leq g(x^{(k)}) - \left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle - \frac{1}{2\tau}|x^{(k+1)} - x^{(k)}|^2$$

$$\Rightarrow \ (f+g)(x^{(k+1)}) \leq (f+g)(x^{(k)}) - \Big(\frac{1}{2\tau} - \frac{L}{2}\Big)|x^{(k+1)} - x^{(k)}|^2 \,.$$

Relative Error Condition:

- **Inexact Algorithm**:

$$\|\partial(f+g)(x^{(k+1)})\|_- = \|\partial g(x^{(k+1)}) + \nabla f(x^{(k+1)})\|_-$$
$$\leq |v^{(k+1)} + \nabla f(x^{(k)})| + |\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})| \leq (b+L)|x^{(k+1)} - x^{(k)}|$$

- **Exact Algorithm**: Use optimality of $x^{(k+1)}$:

$$\frac{x^{(k)} - x^{(k+1)}}{\tau} - \nabla f(x^{(k)}) \in \partial g(x^{(k+1)})\,.$$

Convergence of Forward–Backward Splitting

Continuity Condition:

- **Inexact Algorithm**: Assume that g is continuous on $\operatorname{dom} g$.
- **Exact Algorithm**:
  - Let $x^{(k)} \overset{k\in K}{\to} \tilde{x}$ with $K \subset \mathbb{N}$.
  - Since $((f+g)(x^{(k)}))_{k\in\mathbb{N}}$ is monotonically non-increasing, we have

$$\Big(\frac{1}{2\tau} - \frac{L}{2}\Big)|x^{(k+1)} - x^{(k)}|^2 \leq (f+g)(x^{(k)}) - (f+g)(x^{(k+1)}) \to 0\,.$$

  - Then $\limsup_{k \overset{k\in K}{\to} \infty} g(x^{(k+1)}) \leq g(\tilde{x})$ by taking $\limsup$ on both sides of

$$\begin{aligned} g(x^{(k+1)}) + \Big\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \Big\rangle + \frac{1}{2\tau}|x^{(k+1)} - x^{(k)}|^2 \\ \leq g(\tilde{x}) + \Big\langle \nabla f(x^{(k)}), \tilde{x} - x^{(k)} \Big\rangle + \frac{1}{2\tau}|\tilde{x} - x^{(k)}|^2\,. \end{aligned}$$

  - Combined with lower semi-continuity $\lim_{k \overset{k\in K}{\to} \infty} g(x^{(k)}) = g(\tilde{x})$.

Convergence of Forward–Backward Splitting

Theorem:

Let $(x^{(k)})_{k\in\mathbb{N}}$ be a bounded sequence that is generated by FBS or inexact FBS. Then $(x^{(k)})_{k\in\mathbb{N}}$ converges to a critical point x^* of $f+g$. Moreover, $(x^{(k)})_{k\in\mathbb{N}}$ has the finite length property:

$$\sum_{k=0}^{\infty} |x^{(k+1)} - x^{(k)}| < +\infty\,.$$

Generalized Abstract Descent Algorithm

Generalized Abstract Descent Algorithm: [O. 2016]

- **Let** $\mathcal{F}\colon \mathbb{R}^N \times \mathbb{R}^P \to \overline{\mathbb{R}}$ be proper lsc with $\inf \mathcal{F} > -\infty$.

(H1) (**Sufficient decrease condition**) For each $k \in \mathbb{N}$:

$$\mathcal{F}(x^{(k+1)}, u^{(k+1)}) + a_k d_k^2 \le \mathcal{F}(x^{(k)}, u^{(k)}) \,.$$

(H2) (**Relative error condition**) For each $k \in \mathbb{N}$: (set $d_j = 0$ for $j \le 0$)

$$b_{k+1} \|\partial \mathcal{F}(x^{(k+1)}, u^{(k+1)})\|_- \le b \sum_{i \in I} \theta_i d_{k+1-i} + \varepsilon_{k+1} \,.$$

(H3) (**Continuity condition**) There exists $K \subset \mathbb{N}$ and $(\tilde{x}, \tilde{u})$:

$$(x^{(k)}, u^{(k)}) \overset{\mathcal{F}}{\to} (\tilde{x}, \tilde{u}) \quad \text{as } k \overset{k \in K}{\to} \infty \,.$$

(H4) (**Distance condition**) $d_k \to 0 \Rightarrow |x^{(k+1)} - x^{(k)}| \to 0$ and

$$\exists k' \colon \forall k \ge k' \colon d_k = 0 \Rightarrow \exists k'' \colon \forall k \ge k'' \colon x^{(k+1)} = x^{(k)} \,.$$

(H5) (**Parameter condition**)

$$(b_k)_{k \in \mathbb{N}} \notin \ell_1 \,, \quad \sup_{k \in \mathbb{N}} (a_k b_k)^{-1} < \infty \,, \quad \inf_{k \in \mathbb{N}} a_k =: \underline{a} > 0 \,, \quad (\varepsilon_k)_{k \in \mathbb{N}} \in \ell_1 \,.$$

Generalized Abstract Descent Algorithm

Theorem:

Suppose $\mathcal{F}$ is a proper, lsc, Kurdyka-Łojasiewicz function with $\inf \mathcal{F} > -\infty$. Let $(x^{(k)})_{k\in\mathbb{N}}$, $(u^{(k)})_{k\in\mathbb{N}}$ be bounded and satsify (H1)–(H5). Assume that converging subsequences of $(x^{(k)}, u^{(k)})_{k\in\mathbb{N}}$ converge $\mathcal{F}$-attentive. **Then**:

(i) The sequence $(d_k)_{k\in\mathbb{N}}$ satisfies

$$\sum_{k=0}^{\infty} d_k < +\infty\,.$$

(ii) If d_k satisfies $|x^{(k+1)} - x^{(k)}| \leq \bar{c} d_{k+k'}$ for some k', then

$$\sum_{k=0}^{\infty} |x^{(k+1)} - x^{(k)}| < \infty$$

and $(x^{(k)})_{k\in\mathbb{N}}$ converges to $\tilde{x}$.

(iii) If $(u^{(k)})_{k\in\mathbb{N}}$ converges, then $(x^{(k)}, u^{(k)})_{k\in\mathbb{N}}$ converges to a critical point of $\mathcal{F}$.

Inertial proximal algorithm for nonconvex optimization

Algorithm: (iPiano, [O., Chen, Brox, Pock 2014])

- **Optimization problem**: $\min_{x\in\mathbb{R}^N}\ h(x), \quad h(x) := f(x) + g(x)$
  - ∇f is Lipschitz
  - g is proper, lsc, convex and simple
- **Iterations** $(k \geq 0)$: Update $(x^{-1} := x^0 \in \operatorname{dom} g)$

$$x^{(k+1)} = \operatorname{prox}_{\alpha_k g}\big(x^{(k)} - \alpha_k \nabla f(x^{(k)}) + \beta_k (x^{(k)} - x^{(k-1)})\big)$$

- **Parameter setting** for α_k and β_k, *see convergence analysis*

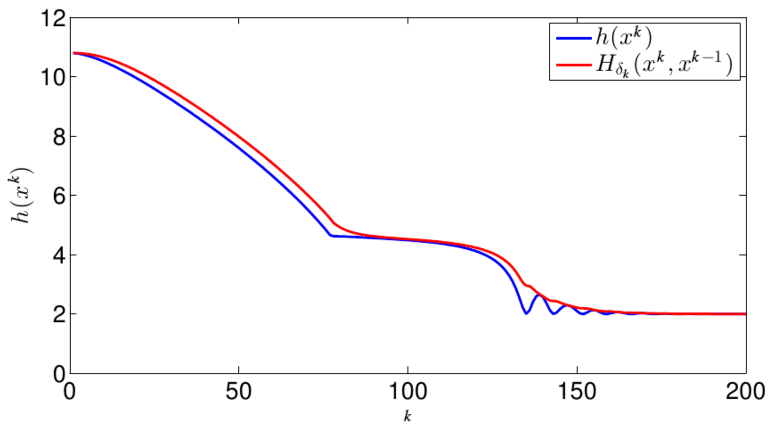
Remark:

- Extension: g **non**-convex in [Bot, Csetnek, Lázló 2016], [O. 2015].
- Other suitable names: “proximal Heavy-ball method”

A Lyapunov function: Define $H_{\delta_k}(x,y) := h(x) + \delta_k|x-y|^2$ ($\delta_k > 0$).

- $(H_{\delta_k}(x^{(k)}, x^{(k-1)}))_{k=0}^{\infty}$ is non-increasing: ($\gamma_k > 0$)

$$H_{\delta_{k+1}}(x^{(k+1)}, x^{(k)}) \leq H_{\delta_k}(x^{(k)}, x^{(k-1)}) - \gamma_k|x^{(k)} - x^{(k-1)}|^2 .$$



Proof of the Lyapunov Property.

- Update step: $x^{(k+1)} \in \arg\min_{x\in\mathbb{R}^N} \; G^{(k)}(x)$ with

$$G^{(k)}(x) := g(x) + \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle + \frac{1}{2\alpha_k}|x - (x^{(k)} + \beta(x^{(k)} - x^{(k-1)})))|^2 \,.$$

- Optimality of $x^{(k+1)}$:

$$G^{(k)}(x^{(k+1)}) + \frac{1}{2\alpha_k}|x^{(k+1)} - x^{(k)}|^2 \leq G^{(k)}(x^{(k)}) = g(x^{(k)})$$

- Descent Lemma:

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle + \frac{L_k}{2}|x^{(k+1)} - x^{(k)}|^2$$

- Combination of optimality and descent lemma:

$$\begin{aligned} h(x^{(k+1)}) \leq{}& h(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle + \frac{L_k}{2}|x^{(k+1)} - x^{(k)}|^2 \\ & - \left\langle \nabla f(x^{(k)}) - \frac{\beta_k}{\alpha_k}(x^{(k)} - x^{(k-1)}), x^{(k+1)} - x^{(k)} \right\rangle - \frac{1}{2\alpha_k}|x^{(k+1)} - x^{(k)}|^2 \,. \end{aligned}$$

Convergence Results – Lyapunov Function for iPiano

- Use $2\langle a, b\rangle \leq |a|^2 + |b|^2$ for vectors $a, b \in \mathbb{R}^N$:

$$\underbrace{h(x^{(k+1)}) + \delta_k |x^{(k+1)} - x^{(k)}|^2}_{H_{\delta_k}(x^{(k+1)}, x^{(k)})} \leq \underbrace{h(x^{(k)}) + \delta_k |x^{(k)} - x^{(k-1)}|^2}_{H_{\delta_k}(x^{(k)}, x^{(k-1)})} - \gamma_k |x^{(k)} - x^{(k-1)}|^2$$

i.e.

$$H_{\delta_{k+1}}(x^{(k+1)}, x^{(k)}) \leq H_{\delta_k}(x^{(k)}, x^{(k-1)}) - \gamma_k |x^{(k)} - x^{(k-1)}|^2$$

where $\gamma_k > 0$ and $(\delta_k)_{k\in\mathbb{N}}$ monotonically non-increasing with

$$\gamma_k := \frac{1}{2}\left(\frac{1 - 2\beta_k}{\alpha_k} - L_k\right) \quad \text{and} \quad \delta_k := \gamma_k + \frac{\beta_k}{2\alpha_k}$$

Yields step size restrictions: $(L_k = L)$

g convex:	$0 < \alpha < \frac{2(1-\beta)}{L}$	$\beta \in [0, 1)$
$g - \frac{m}{2}|\cdot|^2$ convex:	$0 < \alpha < \frac{2(1-\beta)}{L-m}$	$\beta \in [0, 1)$
g non-convex:	$0 < \alpha < \frac{(1-2\beta)}{L}$	$\beta \in [0, \frac{1}{2})$

Convergence Results of iPiano

Theorem: Convergence Results of iPiano:

- The sequence $(h(x^{(k)}))_{k\in\mathbb{N}}$ converges.
- There exists a converging subsequence $(x^{k_j})_{j\in\mathbb{N}}$.
- Any limit point $x^* := \lim_{j\to\infty} x^{k_j}$ is a critical point h and $h(x^{k_j}) \to h(x^*)$ as $j \to \infty$.

If $H_\delta(x, y)$ has the Kurdyka-Łojasiewicz property at (x^*, x^*), then

- $(x^{(k)})_{k\in\mathbb{N}}$ has finite length, i.e.,

$$\sum_{k=1}^{\infty} |x^{(k)} - x^{(k-1)}| < \infty\,,$$

- $x^{(k)} \to x^*$ as $k \to \infty$,
- (x^*, x^*) is a critical point of H_δ, and x^* is a critical point of h, i.e.,

$$0 \in \partial h(x^*)\,.$$

Diffusion based Image Compression

Diffusion based Image Compression:

Encoding:

- store image $\mathbf{g}$ only in some small number of pixel: $\mathbf{c}_i = 1$ if pixel i is stored and 0 otherwise

Decoding:

- use $\mathbf{u}_i = \mathbf{g}_i$ for all i with $\mathbf{c}_i = 1$
- use linear diffusion in unknown region ($\mathbf{c}_i = 0$) (solve Laplace equation $L\mathbf{u} = 0$)

⇝ solve for $\mathbf{u}$ in

$$C(\mathbf{u} - \mathbf{g}) - (I - C)L\mathbf{u} = 0$$

where $C = \mathrm{diag}(\mathbf{c})$, and I the identity matrix

↓**encoding**

↓**decoding**

Diffusion based Image Compression

Diffusion based Image Compression:

Our goal:

- Find a sparse vector $\mathbf{c}$ that yields the best reconstruction.

Non-convex optimization problem:

$$\min_{\mathbf{c}\in\mathbb{R}^N,\mathbf{u}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{u}(\mathbf{c})-\mathbf{g}\|^2+\lambda\|\mathbf{c}\|_1$$

$$s.t.\ C(\mathbf{u}-\mathbf{g})-(I-C)L\mathbf{u}=0$$

or equivalently (setting $A := C+(C-I)L$):

$$\min_{\mathbf{c}\in\mathbb{R}^N} \frac{1}{2}\|A^{-1}C\mathbf{g}-\mathbf{g}\|^2+\lambda\|\mathbf{c}\|_1$$

↓**encoding**

↓**decoding**

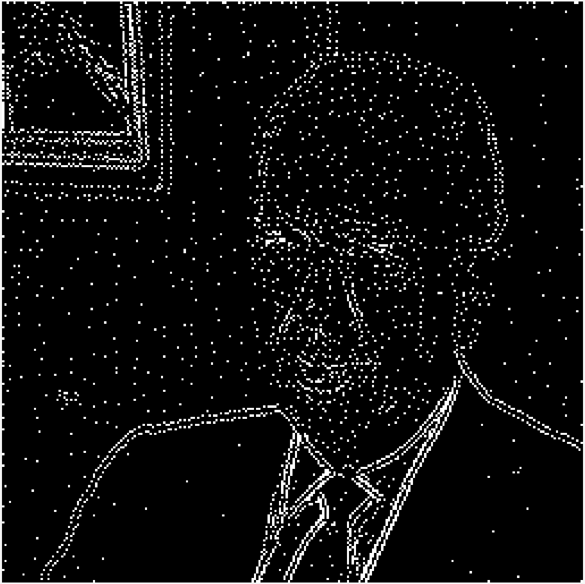
Results for Trui

Results for Trui

Results for Trui

Results for Walter

Results for Walter



Results for Walter

KL Exponent: A measure for the convergence rate:

- **Reminder:** KL inequality for $h\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ at $\bar{x} \in \operatorname{dom} \partial h$:

  There exists [...] and $\varphi\colon [0,\eta) \to \mathbb{R}_+$ with [...] such that

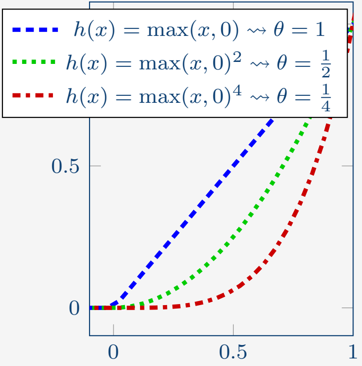
  $$\varphi'(h(x) - h(\bar{x}))\operatorname{dist}(0, \partial h(x)) \geq 1$$

  for x close to $\bar{x}$ and $h(\bar{x}) < h(x) < h(\bar{x}) + \eta$.

- If $\varphi(s) = \frac{c}{\theta} s^\theta$ for $\theta \in (0,1]$, then θ is known as the **KL exponent**. It holds that

  $$\|\partial h(x)\|_- \geq \frac{1}{c}\big(h(x) - h(\bar{x})\big)^{1-\theta} .$$

- **Fact:** e.g. when h is semi-algebraic. *See [Kurdyka, 1998] and [Bolte, Daniilidis, Lewis, Shiota 2007].*



Convergence for iPiano

Theorem: (Local convergence rates for iPiano) [O. 2018] *analogue to [Frankel–Garrigos–Peypouquet, 2014], [Johnstone–Moulin, 2016], [Li–Pong, 2016]*

Let θ be the KL-exponent of H_δ.

- If $\theta = 1$, then $x^{(k)}$ converges to x^* in a **finite number of iterations**.
- If $\frac{1}{2} \leq \theta < 1$, then $H_\delta(x^{(k+1)}, x^{(k)}) \to h(x^*)$ and $x^{(k)} \to x^*$ **linearly**.
- If $0 < \theta < \frac{1}{2}$, then $H_\delta(x^{(k+1)}, x^{(k)}) - h(x^*) \in O(k^{\frac{1}{2\theta-1}})$ and $|x^{(k)} - x^*| \in O(k^{\frac{\theta}{2\theta-1}})$.

Remark: [Liang–Fadili–Peyré, 2016]: local convergence rates using partial smoothness.

Gradient of the Moreau envelope

Theorem: (Local convergence) [O. 2018]

Let x^* be a local (or global) minimizer of h and a certain growth condition holds at x^*.

- Then, if $x^{(k_0)}$ is sufficiently close to x^*, then there exists $r > 0$:

$$x^{(k)} \in B_r(x^*) \quad \text{for all } k \geq k_0 \,.$$

Reminder/Fact:
If f is **prox-regular**, then, **locally**, $e_\lambda f \in \mathcal{C}^{1,+}$ with

$$\nabla e_\lambda f(x) = \frac{1}{\lambda}(x - \mathrm{prox}_{\lambda f}(x)) \,.$$

being λ^{-1}-Lipschitz continuous (for λ small enough).
If f is **convex**, $e_\lambda f$ is finite-valued, and the formula above holds **globally**.

Gradient of the Moreau envelope

Assume from now on:

The gradient of the Moreau envelope can be expressed as above.

Remark:

- Can be true **globally** or on a **neighborhood** of a local (or global) minimum.
- All iterates of iPiano stay within a neighborhood of a local minimum.
- Proximal mappings derived via $\nabla e_\lambda f$ are single-valued.
- Proximal mapping in the backward-step of iPiano may be multi-valued.

We present some **informal results** on the next slides.

Heavy-ball method on the Moreau envelopeof a function:

$$\min_{x\in\mathbb{R}^N} F(x)\,, \qquad F(x) = e_\lambda f(x) = \min_{w\in\mathbb{R}^N} f(w) + \frac{1}{2\lambda}|w-x|^2\,.$$

- Heavy-ball update step (using $\theta := \alpha\lambda^{-1}$)

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha\nabla e_\lambda f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)}) \\ &= x^{(k)} - \alpha\lambda^{-1}(x^{(k)} - \mathrm{prox}_{\lambda f}(x^{(k)})) + \beta(x^{(k)} - x^{(k-1)}) \\ &= (1-\theta)x^{(k)} + \theta\mathrm{prox}_{\lambda f}(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})\,. \end{aligned}$$

$\rightarrow$ **inertial proximal point algorithm** for $\theta = 1$.

- f prox-regular: local convergence.
- f convex: global convergence.

Heavy-ball method on the sum of two Moreau envelopes

Heavy-ball method on the sum of two Moreau envelopes:

$$F(x) = \frac{1}{2}\left(e_\lambda g(x) + e_\lambda f(x)\right) = \min_{w,z\in\mathbb{R}^N} \frac{1}{2}\left(g(z) + f(w) + \frac{1}{2\lambda}|z-x|^2 + \frac{1}{2\lambda}|w-x|^2\right).$$

- Heavy-ball update step:

$$x^{(k+1)} = (1-\theta)x^{(k)} + \frac{\theta}{2}\left(\mathrm{prox}_{\lambda g}(x^{(k)}) + \mathrm{prox}_{\lambda f}(x^{(k)})\right) + \beta(x^{(k)} - x^{(k-1)}).$$

→ **inertial averaged proximal minimization method** for $\theta = 1$.

→ **inertial averaged projection method**, if f and g are indicator functions.

- Obvious extension to the weighted sum of Moreau envelopes.
- f, g prox-regular: local convergence.
- f, g convex: global convergence.

iPiano on an objective involving a Moreau envelope:

$$\min_{x\in\mathbb{R}^N} g(x) + F(x)\,, \qquad F(x) = e_\lambda f(x) = \min_{w\in\mathbb{R}^N} f(w) + \frac{1}{2\lambda}|w-x|^2\,.$$

- iPiano update step:

$$\begin{aligned} x^{(k+1)} &= \mathrm{prox}_{\alpha g}(y^{(k)} - \alpha\nabla e_\lambda f(x^{(k)})) \\ &= \mathrm{prox}_{\theta\lambda g}((1-\theta)x^{(k)} + \theta \mathrm{prox}_{\lambda f}(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})) \end{aligned}$$

→ **inertial alternating proximal minimization method** for $\theta = 1$.

→ **inertial alternating projection method**, if f and g are indicator functions.

- f prox-regular: local convergence.
- f convex: global convergence. (also non-convex g with multi-valued prox)

A Feasibility Problem

A Feasibility Problem:
Find $X \in \mathbb{R}^{N\times M}$ of rank R that satisfies a lin. sys. of eq. $\mathcal{A}(X) = b$:

$$\text{find} \quad X \quad \text{in} \quad \underbrace{\{X \in \mathbb{R}^{N\times M} \,|\, \mathcal{A}(X) = b\}}_{=:\mathscr{A}} \cap \underbrace{\{X \in \mathbb{R}^{N\times M} \,|\, \mathrm{rk}(X) = R\}}_{=:\mathscr{R}} \,.$$

- The projection onto each set is easy:

$$\mathrm{proj}_{\mathscr{A}}(X) = X - \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}(\mathcal{A}(X) - b) \quad \text{and} \quad \mathrm{proj}_{\mathscr{R}}(X) = \sum_{i=1}^{R} \sigma_i u_i v_i^\top \,,$$

- $USV^\top$ is (ordered) singular value decomposition of X ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N$).
- 200 randomly generated problems with $M = 110$, $N = 100$, $R = 4$, $D = 450$.
- max. 1000 iterations.

A Feasibility Problem

Precision 10^p →	-2	-4	-6	-8	-10	-12	-2	-4	-6	-8	-10	-12	-2	-4	-6	-8	-10	-12
Method	iterations						time [sec]						success [%]					
`alternating projection`	235	886	—	—	—	—	1.88	7.03	—	—	—	—	100	97.5	0	0	0	0
`averaged projection`	639	—	—	—	—	—	5.13	—	—	—	—	—	100	0	0	0	0	0
`Douglas-Rachford`	974	—	—	—	—	—	8.10	—	—	—	—	—	2	0	0	0	0	0
`Douglas-Rachford 75`	209	449	696	949	—	—	1.68	3.62	5.63	7.66	—	—	100	100	100	100	0	0
`glob-altproj`, $\alpha = 0.99$	238	894	—	—	—	—	1.92	7.18	—	—	—	—	100	96.5	0	0	0	0
`glob-ipiano-altproj`, $\beta = 0.45$	—	—	—	—	—	—	—	—	—	—	—	—	0	0	0	0	0	0
`glob-ipiano-altproj-bt`, $\beta = 0.45$	45	69	90	115	140	166	0.65	1.03	1.52	2.08	2.63	3.20	100	100	100	100	100	100
`heur-ipiano-altproj`, $\beta = 0.75$	59	212	386	567	749	925	0.79	2.82	5.14	7.52	9.93	12.22	100	100	100	100	100	91
`loc-heavyball-avrgproj-bt`, $\beta = 0.75$	126	297	502	717	929	—	2.29	5.47	9.24	13.21	17.17	—	100	100	100	100	93.5	0
`loc-ipiano-altproj-bt`, $\beta = 0.75$	66	101	138	176	214	252	1.32	2.06	2.80	3.56	4.31	5.06	100	100	100	100	100	100

- Non-convex version of Douglas–Rachford splitting [Li, Pong 2016].

Accelerations of Forward–Backward Splitting
— Part 5: Acceleration and Variants of FBS —

Peter Ochs
Saarland University
ochs@math.uni-sb.de

— June 11th – 13th, 2018 —

www.mop.uni-saarland.de

Table of Contents

FISTA

FISTA: [Beck, Teboull 2009]

- Fast Iterative Shrinkage-Thresholding Algorithm

- Extension of Nesterov's Accelerated Gradient to convex FBS setting:

$$\min_{x\in\mathbb{R}^N} f(x)+g(x)\,, \quad f,g \text{ convex}\,, \quad \nabla f \text{ is } L\text{-Lipschitz.}$$

- **Algorithm**:

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$$
$$y^{(k)} = x^{(k)} + \Big(\frac{t_k-1}{t_{k+1}}\Big)(x^{(k)}-x^{(k-1)})$$
$$x^{(k+1)} = \mathrm{prox}_{g/L}\big(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\big)$$

- **Optimal Algorithm** $O(1/k^2)$: Convergence rate:

$$(f+g)(x^{(k)}) - (f+g)(x^\star) \leq \frac{2L|x^{(0)}-x^\star|^2}{(k+1)^2}\,.$$

FISTA for non-convex problems

FISTA for non-convex problems: [Wen, Chen, Pong 2015]

- **Problem**:

$$\min_{x\in\mathbb{R}^N} f(x) + g(x)$$

with g convex and f (non-convex) satisfies for some $l, L \geq 0$, $L \geq l$

$$f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle - \frac{l}{2}|x-\bar{x}|^2 \quad \forall x, \bar{x}\,,$$
$$f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{L}{2}|x-\bar{x}|^2 \quad \forall x, \bar{x}\,.$$

- For $0 \leq \inf_k \beta_k \leq \sup_k \beta_k < \sqrt{\frac{L}{L+l}}$, the following algorithm

$$y^{(k)} = x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)})$$
$$x^{(k+1)} = \mathrm{prox}_{g/L}\big(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\big)$$

converges to a critical point of $f+g$:

Adaptive FISTA

Update Scheme: FISTA

$$
\begin{aligned}
y_{\beta_k}^{(k)} &= x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}) \\
x^{(k+1)} &= \operatorname*{argmin}_x g(x) + f(y_{\beta_k}^{(k)}) + \left\langle \nabla f(y_{\beta_k}^{(k)}), x - y_{\beta_k}^{(k)} \right\rangle + \frac{1}{2\tau}|x - y_{\beta_k}^{(k)}|^2
\end{aligned}
$$

Adaptive FISTA

Update Scheme: FISTA

$$y_{\beta_k}^{(k)} = x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x g(x) + f(y_{\beta_k}^{(k)}) + \left\langle \nabla f(y_{\beta_k}^{(k)}), x - y_{\beta_k}^{(k)} \right\rangle + \frac{1}{2\tau}|x - y_{\beta_k}^{(k)}|^2$$

Equivalent to

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} g(x) + \frac{1}{2\tau}|x - \left(y_{\beta_k}^{(k)} - \tau\nabla f(y_{\beta_k}^{(k)})\right)|^2 =: \operatorname{prox}_{\tau g}\left(y_{\beta_k}^{(k)} - \tau\nabla f(y_{\beta_k}^{(k)})\right)$$

Update Scheme: **Adaptive FISTA** (**also non-convex**) [O., Pock 2017]

$$y_{\beta_k}^{(k)} = x^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x \min_{\beta_k} g(x) + f(y_{\beta_k}^{(k)}) + \left\langle \nabla f(y_{\beta_k}^{(k)}), x - y_{\beta_k}^{(k)} \right\rangle + \frac{1}{2\tau} |x - y_{\beta_k}^{(k)}|^2$$

Update Scheme: **Adaptive FISTA** (f **quadratic**) [O., Pock 2017]

$$y^{(k)}_{\beta_k} = x^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

$$x^{(k+1)} = \operatorname*{argmin}_x \min_{\beta_k} g(x) + f(y^{(k)}_{\beta_k}) + \left\langle \nabla f(y^{(k)}_{\beta_k}), x - y^{(k)}_{\beta_k} \right\rangle + \frac{1}{2\tau} |x - y^{(k)}_{\beta_k}|^2$$

... Taylor expansion around $x^{(k)}$ and optimize for $\beta_k = \beta_k(x)$ ...

$$x^{(k+1)} = \operatorname*{argmin}_x \; g(x) + \frac{1}{2} |x - (x^{(k)} - \boldsymbol{V}_k^{-1} \nabla f(x^{(k)}))|^2_{\boldsymbol{V}_k}$$

Discussion about Solving the Proximal Mapping

Update Scheme: **Adaptive FISTA** (f **quadratic**)

$$\begin{aligned} x^{(k+1)} &= \operatorname*{argmin}_{x\in\mathbb{R}^N} g(x) + \frac{1}{2}|x - (x^{(k)} - \boldsymbol{V}_k^{-1}\nabla f(x^{(k)}))|^2_{\boldsymbol{V}_k} \\ &=: \mathrm{prox}_g^{\boldsymbol{V}_k}(x^{(k)} - \boldsymbol{V}_k^{-1}\nabla f(x^{(k)})) \end{aligned}$$

with $\boldsymbol{V}_k \in \mathbb{S}_{++}(N)$ as in the **(zero memory) SR1 quasi-Newton method**:

$$\boldsymbol{V} = \boldsymbol{I} - uu^\top \quad \text{(identity minus rank-1)}\,.$$

- SR1 proximal quasi-Newton method proposed by [Becker, Fadili '12] (convex case).
- Special setting is treated in [Karimi, Vavasis '17].
- Unified and extended in [Becker, Fadili, O. '18].

Solving the rank-1 Proximal Mapping: (g convex)

- For general $\boldsymbol{V}$, the main algorithmic step is hard to solve:

$$\hat{x} = \operatorname{prox}_g^{\boldsymbol{V}} := \underset{x\in\mathbb{R}^N}{\operatorname{argmin}}\ g(x) + \frac{1}{2}|x-\bar{x}|_{\boldsymbol{V}}^2$$

- **Theorem:** [Becker, Fadili '12] $\boldsymbol{V} = \boldsymbol{D} \pm uu^\top \in \mathbb{S}_{++}$ for $u \in \mathbb{R}^N$ and $\boldsymbol{D}$ diagonal. Then

$$\operatorname{prox}_g^{\boldsymbol{V}}(\bar{x}) = \boldsymbol{D}^{-1/2} \circ \operatorname{prox}_{g\circ \boldsymbol{D}^{-1/2}}(\boldsymbol{D}^{1/2}\bar{x} \mp v^\star)$$

where $v^\star = \alpha^\star \boldsymbol{D}^{-1/2}u$ and $\alpha^\star$ is the unique root of

$$l(\alpha) = \left\langle u, \bar{x} - \boldsymbol{D}^{-1/2} \circ \operatorname{prox}_{g\circ \boldsymbol{D}^{-1/2}} \circ \boldsymbol{D}^{1/2}(\bar{x} \mp \alpha \boldsymbol{D}^{-1}u)\right\rangle + \alpha\,,$$

which is strictly increasing and Lipschitz continuous with $1 + \sum_i u_i^2 d_i$.

Solving the rank-1 Proximal Mapping for ℓ_1-norm

Example:

- Let $g(x) = |x|_1 = \sum_{i=1}^N |x_i|^2$, $\boldsymbol{D} = \mathrm{diag}(d)$, $u \in \mathbb{R}^N$.
- $V = \boldsymbol{D} - uu^\top$.
- Using the theorem, the proximal mapping

$$\operatorname*{argmin}_{x \in \mathbb{R}^N} |x|_1 + \frac{1}{2}|x - \bar{x}|^2_{\boldsymbol{V}}$$

can be solved by

$$\mathrm{prox}_g^{\boldsymbol{V}}(\bar{x}) = \boldsymbol{D}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{D}^{-1/2}}(\boldsymbol{D}^{1/2}\bar{x} + v^\star)\,.$$

where $v^\star = \alpha^\star \boldsymbol{D}^{-1/2} u$ and $\alpha^\star \in \mathbb{R}$ is the unique root of

$$l(\alpha) = \left\langle u, \bar{x} - \boldsymbol{D}^{-1/2} \circ \mathrm{prox}_{g \circ \boldsymbol{D}^{-1/2}} \circ \boldsymbol{D}^{1/2}(\bar{x} + \alpha \boldsymbol{D}^{-1} u) \right\rangle + \alpha\,.$$

Example: (Solving the rank-1 prox of the ℓ_1-norm)

- The proximal mapping wrt. the diagonal matrix is separable and simple

$$
\begin{aligned}
\mathrm{prox}_{g\circ \boldsymbol{D}^{-1/2}}(z) &= \operatorname*{argmin}_{x\in\mathbb{R}^N} |\boldsymbol{D}^{-1/2}x|_1 + \frac{1}{2}|x-z|^2 \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^N} \sum_{i=1}^N |x_i|/\sqrt{d_i} + \frac{1}{2}(x_i - z_i)^2 \\
&= \Big(\operatorname*{argmin}_{x_i\in\mathbb{R}} |x_i|/\sqrt{d_i} + \frac{1}{2}(x_i - z_i)^2 \Big)_{i=1,\dots,N} \\
&= \Big(\max\big(0, |z_i| - 1/\sqrt{d_i}\big)\mathrm{sign}(z_i) \Big)_{i=1,\dots,N}
\end{aligned}
$$

Solving the rank-1 Proximal Mapping for ℓ_1-norm

The root finding problem in the rank-1 prox of the ℓ_1-norm:

- $\alpha^\star$ is the root of the **1D function** (i.e. $l(\alpha^\star) = 0$)

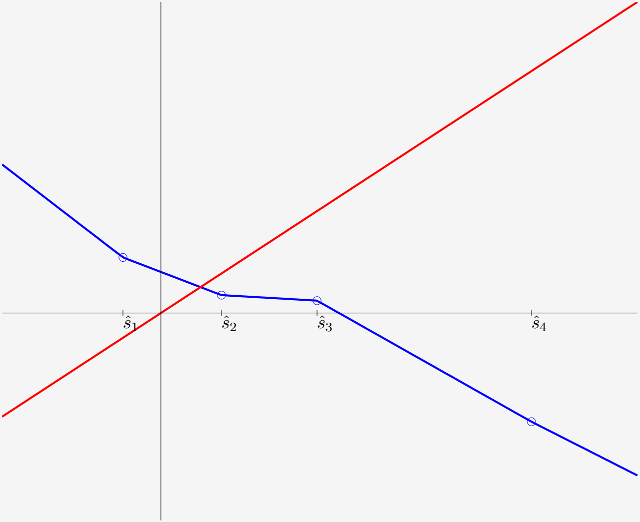
$$l(\alpha) = \left\langle u, \bar{x} - \boldsymbol{D}^{-1/2} \circ \operatorname{prox}_{g \circ \boldsymbol{D}^{-1/2}} \circ \boldsymbol{D}^{1/2}(\bar{x} \mp \alpha \boldsymbol{D}^{-1} u) \right\rangle + \alpha$$
$$= \left\langle u, \bar{x} - \operatorname{PLin}(\bar{x} \mp \alpha \boldsymbol{D}^{-1} u) \right\rangle + \alpha$$

  which is a **piecewise linear function**.

- Construct this function by sorting $K \geq N$ **breakpoints**. Cost: $\mathcal{O}(K \log(K))$.
- The root is determined using **binary search**. Cost: $\mathcal{O}(\log(K))$.
  (*remember:* $l(\alpha)$ is **strictly increasing**)
- Computing $l(\alpha)$ costs $\mathcal{O}(N)$.

$\rightsquigarrow$ **Total cost:** $\mathcal{O}(K \log(K))$.

Solving the rank-1 Proximal Mapping for ℓ_1-norm



from [S. Becker]

Discussion about Solving the Proximal Mapping

Function g	Algorithm
ℓ_1-norm	Separable: exact
Hinge	Separable: exact
ℓ_∞-ball	Separable: exact
Box constraint	Separable: exact
Positivity constraint	Separable: exact
Linear constraint	Nonseparable: exact
ℓ_1-ball	Nonseparable: Semi-smooth Newton + $\mathrm{prox}_{g\circ \boldsymbol{D}^{-1/2}}$ exact
ℓ_∞-norm	Nonseparable: Moreau identity
Simplex	Nonseparable: Semi-smooth Newton + $\mathrm{prox}_{g\circ \boldsymbol{D}^{-1/2}}$ exact

From [Becker, Fadili '12].

Discussion about Solving the Proximal Mapping

Discussion about Solving the Proximal Mapping: (g convex)

- For general $\boldsymbol{V}$, the main algorithmic step is hard to solve:

$$\hat{x} = \operatorname{prox}_g^{\boldsymbol{V}} := \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\; g(x) + \frac{1}{2}|x - \bar{x}|_{\boldsymbol{V}}^2$$

- (L-)**BFGS** uses a rank-r update of the metric with $r > 1$.

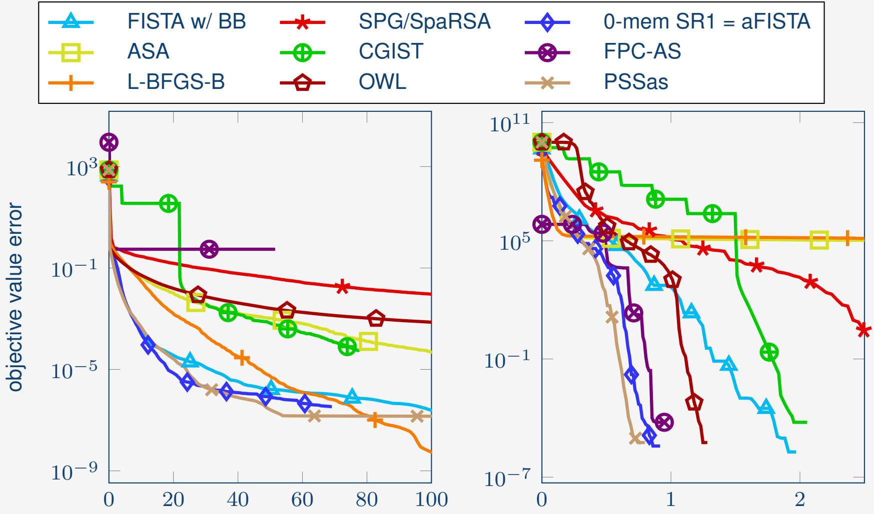
- **Theorem:** [Becker, Fadili, O. '18]
  $\boldsymbol{V} = \boldsymbol{P} \pm \boldsymbol{Q} \in \mathbb{S}_{++}$, $\boldsymbol{P} \in \mathbb{S}_{++}$, $\boldsymbol{Q} = \sum_{i=1}^r u_i u_i^\top$, $\operatorname{rank}(\boldsymbol{Q}) = r$. Then

$$\operatorname{prox}_g^{\boldsymbol{V}}(\bar{x}) = \boldsymbol{P}^{-1/2} \circ \operatorname{prox}_{g \circ \boldsymbol{P}^{-1/2}} \boldsymbol{P}^{1/2}(\bar{x} \mp \boldsymbol{P}^{-1}\boldsymbol{U}\alpha^\star)$$

  where $\boldsymbol{U} = (u_1, \ldots, u_r)$ and $\alpha^\star$ is the unique root of

$$l(\alpha) = \boldsymbol{U}^\top\Big(\bar{x} - \boldsymbol{P}^{-1/2} \circ \operatorname{prox}_{g \circ \boldsymbol{P}^{-1/2}} \circ \boldsymbol{P}^{1/2}(\bar{x} \mp \boldsymbol{P}^{-1}\boldsymbol{U}\alpha)\Big) + \boldsymbol{X}\alpha\,,$$

  where $\boldsymbol{X} := \boldsymbol{U}^\top \boldsymbol{Q}^+ \boldsymbol{U} \in \mathbb{S}_{++}(r)$.



time in seconds
dense problem

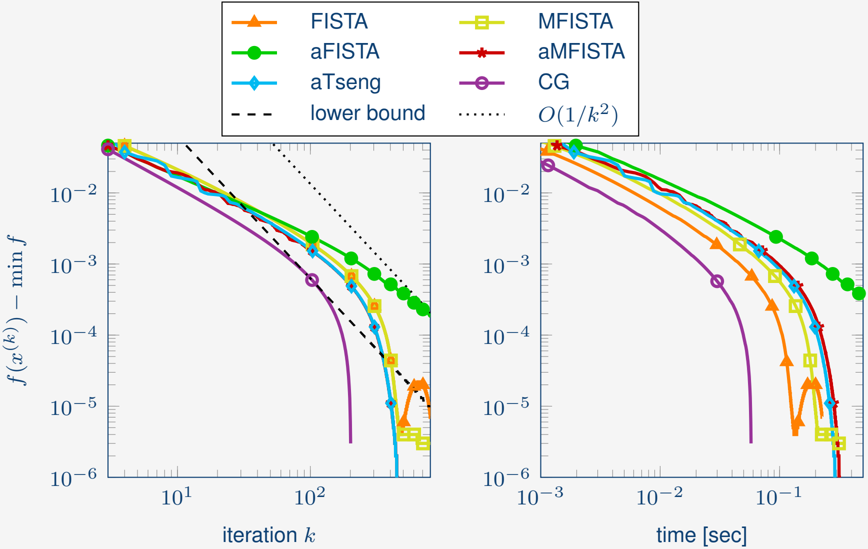
time in seconds
sparse problem

Variants with $O(1/k^2)$-convergence rate

Adaptive FISTA: Variants with $O(1/k^2)$-convergence rate: (convex case)

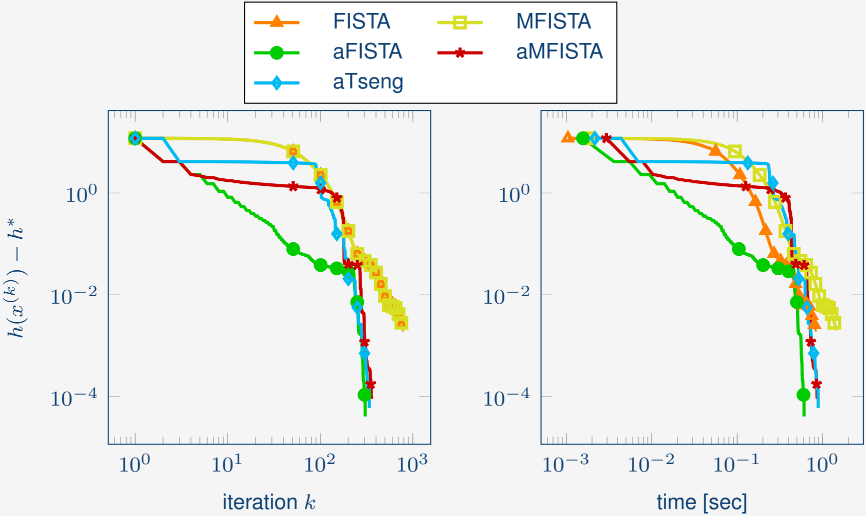
- Adaptive FISTA cannot be proved to have the accelerated rate $O(1/k^2)$.
  - For each point $\bar{x}$, aFISTA decreases the objective more than a FISTA.
  - However, global view on the sequence is lost.
- aFISTA can be embedded into schemes with accelerated rate $O(1/k^2)$.
- **Monotone FISTA version:** (Motivated by [Li, Lin '15], [Beck, Teboulle '09].)
- **Tseng-like version:** (Motivated by [Tseng '08].)

Nesterov's Worst Case Function



LASSO

$$\min_{x\in\mathbb{R}^N} h(x)\,, \quad h(x) = \frac{1}{2}|Ax-b|^2 + \lambda\|x\|_1\,,$$



Proposed Algorithm: (**non-convex setting**)

- **Current iterate** $x^{(k)} \in \mathbb{R}^N$. Step size: $\tau > 0$.
- Define the **extrapolated point** $y_\beta^{(k)}$ that depends on β:

$$y_\beta^{(k)} := x^{(k)} + \beta(x^{(k)} - x^{(k-1)}) .$$

- **Exact version:** Compute $x^{(k+1)}$ as follows:

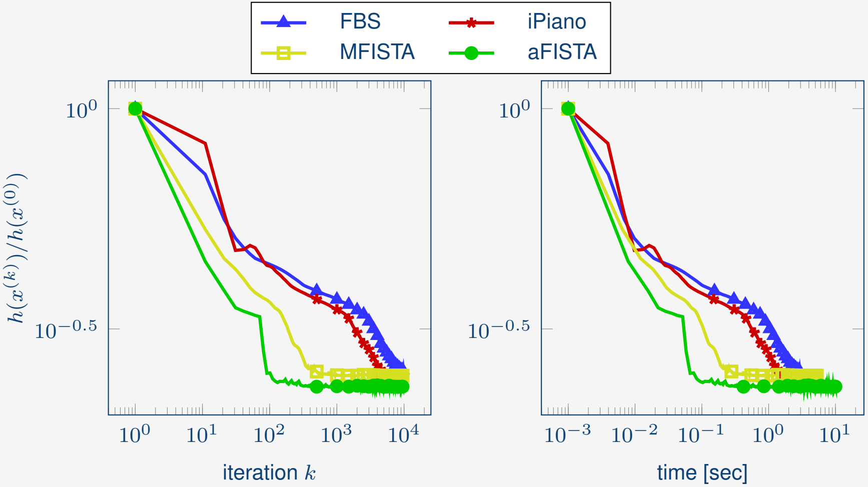
$$x^{(k+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} \min_\beta \ell_f^g(x; y_\beta^{(k)}) + \frac{1}{2\tau}|x - y_\beta^{(k)}|^2 ,$$

$$\ell_f^g(x; y_\beta^{(k)}) := g(x) + f(y_\beta^{(k)}) + \left\langle \nabla f(y_\beta^{(k)}), x - y_\beta^{(k)} \right\rangle$$

- **Inexact version**: Find $x^{(k+1)}$ and β such that

$$\ell_f^g(x^{(k+1)}; y_\beta^{(k)}) + \frac{1}{2\tau}|x^{(k+1)} - y_\beta^{(k)}|^2 \leq f(x^{(k)}) + g(x^{(k)})$$

$$\min_{\substack{W_0,W_1,W2\\ b_0,b_1,b_2}} \sum_{i=1}^{N} \Big(|(W_2\sigma_2(W_1\sigma_1(W_0X+B_0)+B_1)+B_2-\tilde{Y})_{1,i}|^2+\varepsilon^2 \Big)^{1/2} + \lambda \sum_{j=0}^{2} \|W_j\|_1$$



Forward–Backward Envelope: [Stella, Themelis, Patrinos 2017]

- **Forward–Backward Envelope**: (g convex)

$$e_\gamma^{\mathsf{FBS}}(\bar{x}) = \min_{x\in\mathbb{R}^N} \underbrace{g(x) + f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle}_{=:\ell_f^g(x;\bar{x})} + \frac{1}{2\gamma}|x-\bar{x}|^2 \,.$$

- Using

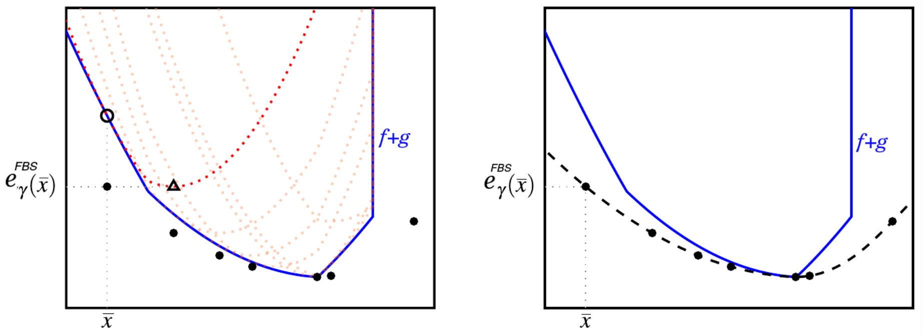
$$\begin{aligned} P_\gamma^{\mathsf{FBS}}(\bar{x}) &:= \operatorname*{argmin}_{x\in\mathbb{R}^N} \ell_f^g(x;\bar{x}) + \frac{1}{2\gamma}|x-\bar{x}|^2 \\ R_\gamma^{\mathsf{FBS}}(\bar{x}) &:= \gamma^{-1}\big(\bar{x} - P_\gamma^{\mathsf{FBS}}(\bar{x})\big) \end{aligned}$$

  the FBS envelope is equivalent to

$$e_\gamma^{\mathsf{FBS}}(\bar{x}) = g(P_\gamma^{\mathsf{FBS}}(\bar{x})) + f(\bar{x}) - \gamma \left\langle \nabla f(\bar{x}), R_\gamma^{\mathsf{FBS}}(\bar{x}) \right\rangle + \frac{\gamma}{2}|R_\gamma^{\mathsf{FBS}}(\bar{x})|^2 \,.$$

- $e_\gamma^{\mathsf{FBS}}(\bar{x})$ is always finite-valued, but not necessarily convex.

Forward–Backward Envelope



modified from [Stella, Themelis, Patrinos 2017]

Properties 1 (Relation of objective values):

- $e_\gamma^{\mathsf{FBS}}(\bar{x}) \leq (f+g)(\bar{x}) - \frac{\gamma}{2}|R_\gamma^{\mathsf{FBS}}(\bar{x})|^2$ for all $\gamma > 0$.
- $(f+g)(P_\gamma^{\mathsf{FBS}}(\bar{x})) \leq e_\gamma^{\mathsf{FBS}}(\bar{x}) - \frac{\gamma}{2}(1-\gamma L)|R_\gamma^{\mathsf{FBS}}(\bar{x})|^2$ for all $\gamma > 0$.
- $(f+g)(P_\gamma^{\mathsf{FBS}}(\bar{x})) \leq e_\gamma^{\mathsf{FBS}}(\bar{x})$ for all $\gamma \in (0, 1/L]$.

Properties 2 (Relation of optimality):

- $(f+g)(z) = e_\gamma^{\mathsf{FBS}}(z)$ for all $\gamma > 0$ and z with $0 \in \partial(f+g)(z)$;
- $\inf(f+g) = \inf e_\gamma^{\mathsf{FBS}}$ and $\operatorname{argmin}(f+g) \subset \operatorname{argmin} e_\gamma^{\mathsf{FBS}}$ for $\gamma \in (0, 1/L]$;
- $\operatorname{argmin}(f+g) = \operatorname{argmin} e_\gamma^{\mathsf{FBS}}$ for all $\gamma \in (0, 1/L)$.

Properties 3 (Differentiability of the forward–backward envelope):

- Assume f is twice continuously differentiable. Then e_γ^{FBS} is **continuously differentiable** and we have

$$\nabla e_\gamma^{\mathsf{FBS}}(\bar{x}) = (\boldsymbol{I} - \gamma \nabla^2 f(\bar{x})) R_\gamma^{\mathsf{FBS}}(\bar{x}).$$

- If $\gamma \in (0, 1/L)$, then the set of stationary points of e_γ^{FBS} equals $\mathrm{zer}\partial(f + g)$.

- e_γ^{FBS} serves as an **exact penalty** formulation for the original objective.

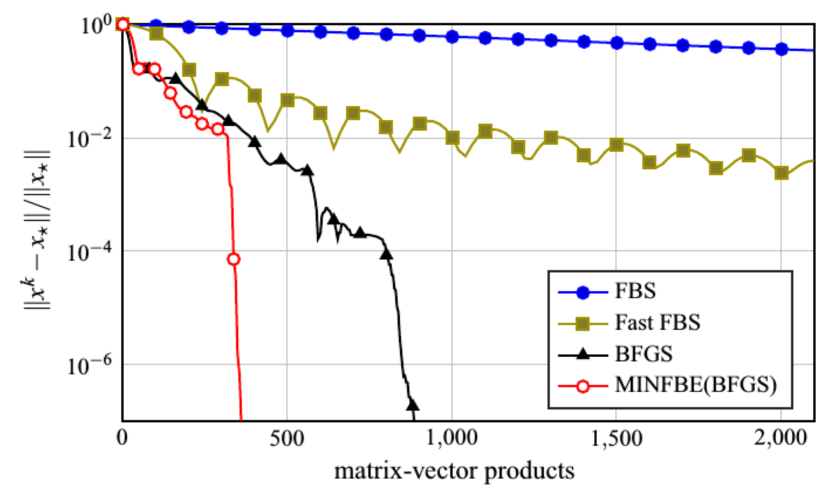
- **Apply variable metric Gradient Descent to e_γ^{FBS}**

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \gamma(\boldsymbol{I} - \gamma \nabla^2 f(x^{(k)}))^{-1} \nabla e_\gamma^{\mathsf{FBS}}(x^{(k)}) \\ &= x^{(k)} - \gamma R_\gamma^{\mathsf{FBS}}(x^{(k)}) \\ &= P_\gamma^{\mathsf{FBS}}(x^{(k)}) \end{aligned}$$

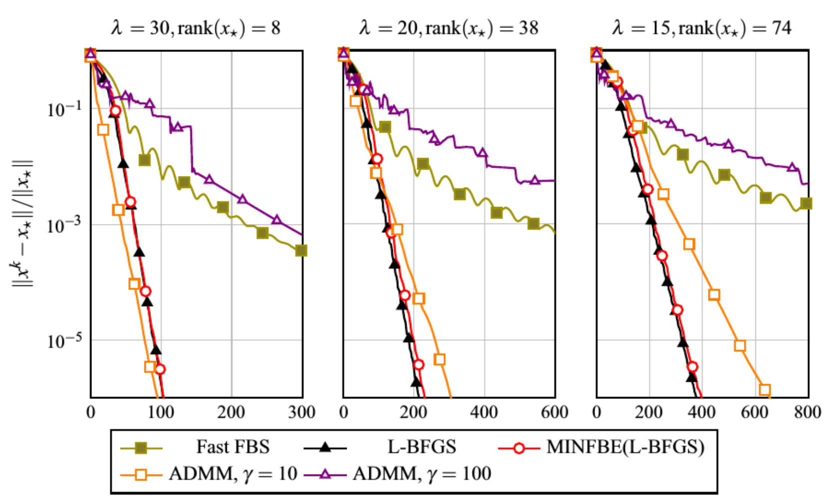
leads to **Forward–Backward Splitting**.

Accelerations using the Forward–Backward Envelope:

- Using the Forward–Backward Envelope, a non-smooth problem is transformed into a **smooth problem**.
- Machinery from smooth optimization can be applied.
- Opens the door for Quasi-Newton Methods and also Newton's method.
- To improve the (weak) convergence properties of quasi-Newton methods, MINFBE interleaves descent steps over the FBE with forward–backward steps, which yields **global convergence**.



LASSO problem from [Stella, Themelis, Patrinos 2017]



Matrix completion problem from [Stella, Themelis, Patrinos 2017]

Generalized Forward–Backward Splitting

Generalized Forward–Backward Splitting: [Raguet, Fadili, Peyré 2013]

- Convex optimization problem:

$$\min_{x\in\mathbb{R}^N} \; f(x) + \sum_{i=1}^{M} g_i(x)\,.$$

- f, g convex; ∇f is L-Lipschitz; g_i are proper lsc convex and simple.

Application Examples:

- Elastic net regularization; e.g. for Linear Regression

$$\min_{x\in\mathbb{R}^N} \; \underbrace{\frac{1}{2}|Ax-b|^2}_{=:f(x)} + \underbrace{\rho|x|_1}_{=:g_1(x)} + \underbrace{\mu|x|_2^2}_{=:g_2(x)}$$

- Block-decomposition: Reformulate

$$\min_{x\in\mathbb{R}^N} \; f(x) + h(x) \quad \text{as} \quad \min_{x,y\in\mathbb{R}^N} \; f(x) + h(y) \quad s.t.\; x = y\,.$$

Generalized Forward–Backward Splitting

Algorithm: (GFBS)

- Fix $\omega \in (0,1]^M$ with $\sum_{i=1}^M \omega_i = 1$, $\gamma \in (0, 2/L)$, $\lambda_k \in (0, \min(\frac{3}{2}, \frac{1}{2} + \frac{1}{\gamma L}))$.
- **Initialize:** $z_i^{(0)} \in \mathbb{R}^N$ and set $x^{(0)} = \sum_{i=1}^M \omega_i z_i^{(0)}$.
- **Update for** $k \geq 0$:
  - For $i = 1, \dots, M$:
$$z_i^{(k+1)} = z_i^{(k)} + \lambda_k \Big(\mathrm{prox}_{\gamma g_i/\omega_i}\big(2x^{(k)} - z_i^{(k)} - \gamma \nabla f(x^{(k)})\big) - x^{(k)}\Big)$$
  - Compute:
$$x^{(k+1)} = \sum_{i=1}^M \omega_i z_i^{(k+1)} \, .$$

Generalized Forward–Backward Splitting

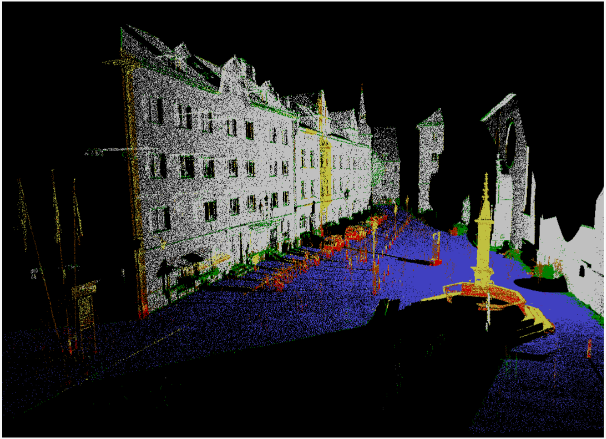
Theorem: (Convergence of Generalized Forward–Backward Splitting)

Under a qualification condition, the sequence $(x^{(k)})_{k\in\mathbb{N}}$ generated by GFBS with erroneuous update steps (with summable error terms) **converges to a solution**.

Properties:

- For $f \equiv 0$: Relaxed **Douglas–Rachford Splitting**.
- For $M = 1$: Relaxed **Forward–Backward Splitting**.

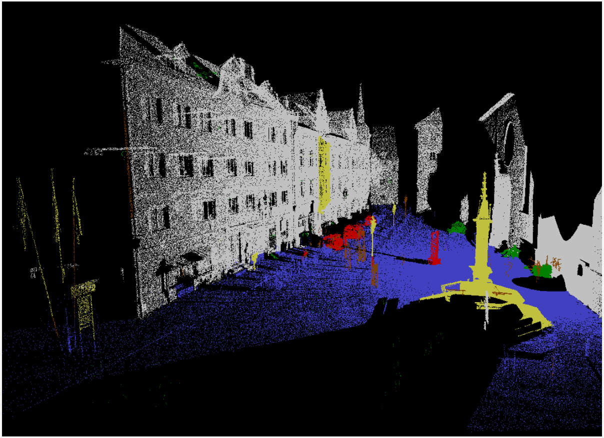
Follow-up work applied to Semantic Labelling of 3D Point Clouds:



Random Forest Classification

[Raguet 2017]

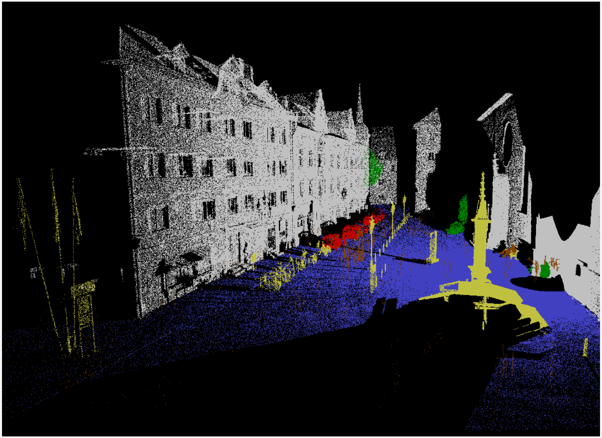
Follow-up work applied to Semantic Labelling of 3D Point Clouds:



Regularized Labelling

[Raguet 2017]

Follow-up work applied to Semantic Labelling of 3D Point Clouds:



Ground Truth Labelling

[Raguet 2017]

Accelerations of Forward–Backward Splitting
— Part 6: Bregman Proximal Minimization —

Peter Ochs
Saarland University
ochs@math.uni-sb.de

— June 11th – 13th, 2018 —

www.mop.uni-saarland.de

Table of Contents

Facts about Gradient Descent

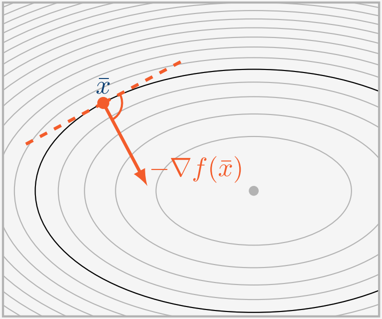
- Smooth optimization problem: (f continuously differentiable)

$$\min_{x\in\mathbb{R}^N} \; f(x)$$

- Update step with step size $\tau > 0$:

$$x^{(k+1)} = x^{(k)} - \tau \nabla f(x^{(k)})\,.$$

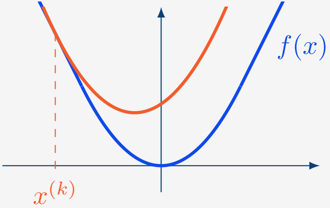
- Step size selection:
  - f **continuously differentiable**
    $\Rightarrow$ line-search is required.
  - ∇f **Lipschitz continuous**
    $\Rightarrow$ feasible range of step sizes can be computed.



Facts about Gradient Descent

- Equivalent to **minimizing a quadratic function**:

$$x^{(k+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle + \frac{1}{2\tau}|x - x^{(k)}|^2 \,.$$



- Optimality condition:

$$\nabla f(x^{(k)}) + \frac{1}{\tau}(x - x^{(k)}) = 0$$
$$\Leftrightarrow x = x^{(k)} - \tau\nabla f(x^{(k)})$$

Facts about Gradient Descent

Another point of view:

- Minimization of a **linear function**

$$f_{x^{(k)}}(x) = f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle$$

with **quadratic penalty on the distance** to $x^{(k)}$:

$$D_h(x, x^{(k)}) = \frac{1}{2\tau} |x - x^{(k)}|^2 \,.$$

- Update step:

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} f_{x^{(k)}}(x) + D_h(x, x^{(k)})$$

Facts about Gradient Descent

Generalization to non-smooth functions f:

- Minimization of a convex **model function**

$$f_{x^{(k)}}(x) \quad \text{with} \quad |f(x) - f_{x^{(k)}}(x)| \le \underbrace{\omega(|x - x^{(k)}|)}_{\text{growth function}}$$

  with **quadratic penalty on the distance** to $x^{(k)}$:

$$D_h(x, x^{(k)}) = \frac{1}{2\tau}|x - x^{(k)}|^2 \,.$$

- Update step:

$$x^{(k+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} f_{x^{(k)}}(x) + D_h(x, x^{(k)})$$

Facts about Gradient Descent

Generalization to non-smooth functions f:

- Minimization of a convex **model function**

$$f_{x^{(k)}}(x) \quad \text{with} \quad |f(x) - f_{x^{(k)}}(x)| \leq \underbrace{\omega(|x - x^{(k)}|)}_{\text{growth function}}$$

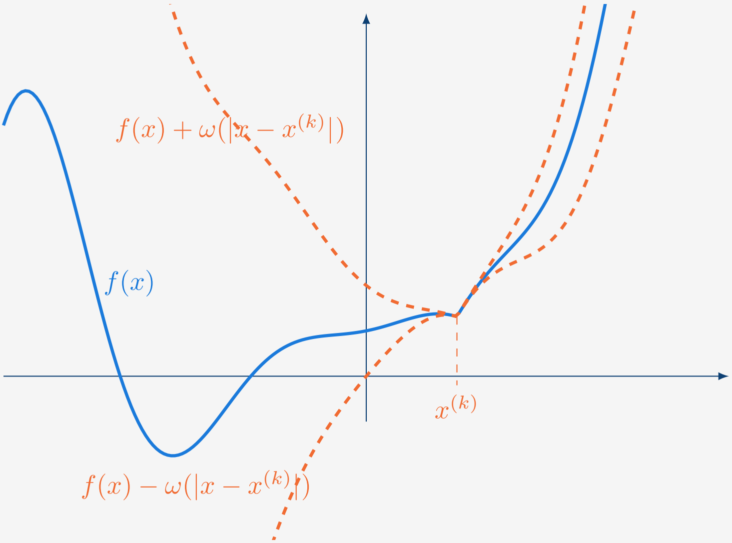
  with **penalty on the distance** to $x^{(k)}$:

$$D_h(x, x^{(k)})\,.$$

- Update step:

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^N} f_{x^{(k)}}(x) + D_h(x, x^{(k)})$$

Model assumption / Growth function



Key Contribution:

The **growth function** and the **distance function**
determine
the **convergence** properties.

Types of growth functions:

(i) *growth function*: $\omega(0) = \omega'(0) = 0$

(ii) *proper growth function*: $\lim_{t \searrow 0} \omega'(t) = \lim_{t \searrow 0} \omega(t)/\omega'(t) = 0$.

(iii) *global growth function* (does not require line-search).

Abstract Algorithm:

$$\tilde{x}^{(k)} = \underset{x\in\mathbb{R}^N}{\operatorname{argmin}}\; f_{x^{(k)}}(x) + D_h(x, x^{(k)})\,.$$

Find $\eta^{(k)} > 0$ using (inexact) **line-search** along

$$x^{(k+1)} = x^{(k)} + \eta^{(k)}(\tilde{x}^{(k)} - x^{(k)})$$

to satisfy an **Armijo-like condition** along.

Remark: (Alternative Line-Search Strategy)

- Replace line-search for $\eta^{(k)} > 0$ by scaling of h in $D_h(x, x^{(k)})$.

Outline

1: Examples for Model Functions

- Gradient Descent, Forward–Backward Splitting, ProxDescent
- Presented with Euclidean distance measure.
- However any distance measure from PART 2 can be used.

2: Examples for Distance Functions

- Bregman distance generated by Legendre functions.

3: Convergence Analysis

- Subsequential convergence to a stationary point.

4: Numerical Examples

- Robust non-linear regression.
- Image deblurring under Poisson noise.

Forward–Backward Splitting

- **Optimization problem**:

$$\min_{x\in\mathbb{R}^N} \underbrace{f_0(x)}_{\substack{\text{non-smooth}\\\text{convex}}} + \underbrace{f_1(x)}_{\substack{\text{diff.}\\\text{non-convex}}}$$

- **Update step:**

$$\tilde{x}^{(k)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} f_0(x) + f_1(x^{(k)}) + \left\langle x - x^{(k)}, \nabla f_1(x^{(k)}) \right\rangle + \frac{1}{2\tau}|x - x^{(k)}|^2$$

- **Model function**:

$$f_{\bar{x}}(x) = f_0(x) + f_1(\bar{x}) + \langle x - \bar{x}, \nabla f_1(\bar{x})\rangle$$

- **Model assumption/error**:

$$|f(x) - f_{\bar{x}}(x)| = |f_1(x) - f_1(\bar{x}) - \langle x - \bar{x}, \nabla f_1(\bar{x})\rangle| \le \omega(|x - \bar{x}|)$$

- FBS case was considered by [Bonettini et al., 2016].

Variable Metric Forward–Backward Splitting

- **Optimization problem**:

$$\min_{x\in\mathbb{R}^N} \underbrace{f_0(x)}_{\substack{\text{non-smooth}\\ \text{convex}}} + \underbrace{f_1(x)}_{\substack{\textbf{twice}\text{ diff.}\\ \text{non-convex}}}$$

- **Model function**:

$$f_{\bar{x}}(x) = f_0(x) + f_1(\bar{x}) + \langle x - \bar{x}, \nabla f_1(\bar{x})\rangle + \frac{1}{2}\langle x - \bar{x}, B(x - \bar{x})\rangle$$

  B is a positive definite approximation to the Hessian $\nabla^2 f_1(\bar{x})$

- **Update step**: (Damped (approx.) Newton Method)

$$\tilde{x}^{(k)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} f_0(x) + f_1(x^{(k)}) + \left\langle x - x^{(k)}, \nabla f_1(x^{(k)})\right\rangle + \frac{1}{2}\left\langle x - x^{(k)}, B(x - x^{(k)})\right\rangle + \frac{1}{2\tau}|x - x^{(k)}|^2$$

ProxDescent

- **Optimization problem**:

$$\min_{x\in\mathbb{R}^N} \underbrace{f_0(x)}_{\substack{\text{non-smooth}\\\text{convex}}} + \underbrace{g}_{\substack{\text{non-smooth}\\\text{convex}\\\text{finite-valued}}}\Big(\underbrace{F(x)}_{\text{diff.}} \Big)$$

- **Model function**: ($DF(\bar{x})$ is the Jacobian matrix of F at $\bar{x}$)

$$f_{\bar{x}}(x) = f_0(x) + g(F(\bar{x}) + DF(\bar{x})(x-\bar{x}))$$

- **Model assumption**:

$$\begin{aligned}|f(x) - f_{\bar{x}}(x)| &= |g(F(x)) - g(F(\bar{x}) + DF(\bar{x})(x-\bar{x}))| \\ &\leq \ell|F(x) - F(\bar{x}) - DF(\bar{x})(x-\bar{x})| \\ &\leq \omega(|x-\bar{x}|)\end{aligned}$$

- **Update step:**

$$\tilde{x}^{(k)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} \; f_0(x) + g(F(x^{(k)}) + DF(x^{(k)})(x - x^{(k)})) + \frac{1}{2\tau}|x - x^{(k)}|^2$$

- [Lewis and Wright, 2016], [Drusvyatskiy and Lewis, 2016]

A Special Case of ProxDescent:

- **Optimization problem:** (Non-linear least-squares problem)

$$\min_{x\in\mathbb{R}^N} \; \frac{1}{2}|F(x)|^2$$

- **Update step:** (Levenberg–Marquardt Algorithm)

$$\tilde{x}^{(k)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} \; \frac{1}{2}|F(x^{(k)}) + DF(x^{(k)})(x - x^{(k)})|^2 + \frac{1}{2\tau}|x - x^{(k)}|^2$$

Composite Optimization: Iterative Reweighting

- **Optimization problem**:

$$\min_{x\in\mathbb{R}^N} \underbrace{f_0(x)}_{\substack{\text{non-smooth}\\ \text{convex}}} + \underbrace{g}_{\substack{\text{smooth}\\ (\nabla g)_i \text{ non-negative}}}\Big(\underbrace{F(x)}_{\substack{\text{Lipschitz}\\ F_i \text{ convex}}} \Big)$$

- **Model function**:

$$f_{\bar{x}}(x) = f_0(x) + g(F(\bar{x})) + \langle \nabla g(F(\bar{x})), F(x) - F(\bar{x}) \rangle$$

- **Model assumption**:

$$\begin{aligned} |f(x) - f_{\bar{x}}(x)| &= |g(F(x)) - g(F(\bar{x})) - \langle \nabla g(F(\bar{x})), F(x) - F(\bar{x}) \rangle| \\ &\leq \omega(|F(x) - F(\bar{x})|) \\ &\leq \omega(|x - \bar{x}|) \end{aligned}$$

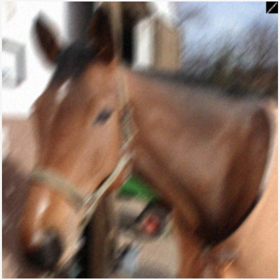
- **Update step:**

$$\tilde{x}^{(k)} = \operatorname*{argmin}_{x\in\mathbb{R}^N} \; f_0(x) + \left\langle \nabla g(F(x^{(k)})), F(x) - F(x^{(k)}) \right\rangle + \frac{1}{2\tau}|x - x^{(k)}|^2 \, .$$

Example: (image deblurring with non-convex regularization)

$$\min_{\mathbf{u}} \; \frac{1}{2}|\mathcal{A}\mathbf{u} - \mathbf{f}|^2 + \rho \sum_{i,j} \log(1 + \mu|(\mathcal{D}\mathbf{u})_{i,j}|)$$

clean



burry/noisy

reconstruction

Distance Measures

Class of Distance Measures:

- *Bregman distance* D_h generated by *Legendre functions* h.

Examples:

- **Euclidean Distance Measure:** $D_h(x,\bar{x}) = \frac{1}{2}|x-\bar{x}|^2$
- **Scaled Euclidean Distance Measure:**

$$D_h(x,\bar{x}) = \frac{1}{2}|x-\bar{x}|_A^2 := \frac{1}{2}\left\langle x-\bar{x}, A(x-\bar{x})\right\rangle$$

- **Burg's Entropy:** (e.g. for non-negativity constraints)

$$D_h(x,\bar{x}) = \sum_{i=1}^{N}\left(\frac{x_i}{\bar{x}_i} - \log\left(\frac{x_i}{\bar{x}_i}\right) - 1\right)$$

- $h(x_i) = -\log(x_i)$ (Barrier) has domain $(0,+\infty)$ and grows towards $+\infty$ for for $x_i \to 0$.

Convergence Results

Seek for stationary point x^*, i.e. $\overline{|\nabla f|}(x^*) = 0$. (Limiting Slope)

Termination of Backtracking Line-Search:

- Backtracking terminates after a finite number of iterations.

Stationarity for Finite Termination:

- Fixed-points of the algorithm are stationary points of f.

Convergence of Objective Values:

- $(f(x^{(k)}))_{k\in\mathbb{N}}$ is non-increasing and converging.

Stationarity of Limit Points

Assumption **to avoid technical details:** D_h has full domain.

Prove Stationarity of Limit Points in Three Settings:

(i) ω is a **growth function**: $\omega(0) = \omega'(0) = 0$ and $|\nabla f|(x^{(k)}) \to 0$.

(ii) ω is a **proper growth function**: $\lim_{t \searrow 0} \omega'(t) = \lim_{t \searrow 0} \omega(t)/\omega'(t) = 0$.

(iii) ω is a **global growth function** (does not require line-search).

Robust Non-linear Regression

Non-smooth non-convex optimization problem:

$$\min_{u:=(a,b)\in\mathbb{R}^P\times\mathbb{R}^P} \sum_{i=1}^{M} \|F_i(u)-y_i\|_1\,, \quad F_i(u) := \sum_{j=1}^{P} b_j \exp(-a_j x_i)$$

- $(x_i, y_i) \in \mathbb{R}\times\mathbb{R}$ noisy non-negative input-output.
- $y_i = F_i(u) + n_i$ with impulse noise n_i.

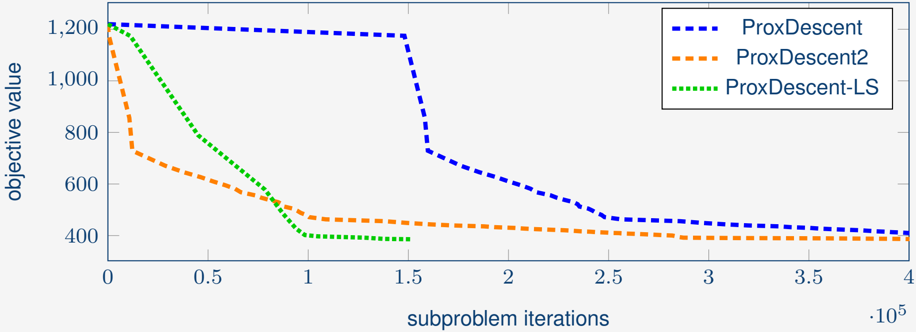
- **Model function** linearizes the inner functions F_i.

- **Convex subproblems** of the form: (solved using dual ascent)

$$\tilde{u} = \operatorname*{argmin}_{u\in\mathbb{R}^P\times\mathbb{R}^P} \sum_{i=1}^{M} \|\mathcal{K}_i u - y_i^\diamond\|_1 + \frac{1}{2\tau}|u-\bar{u}|^2\,, \quad y_i^\diamond := y_i - F(\bar{u}) + \mathcal{K}_i\bar{u}\,.$$

- $\mathcal{K}_i := DF_i(\bar{u})$ is the Jacobian of F_i at $\bar{u}$.

Robust Non-linear Regression



Objective value vs. number of subproblem iterations.

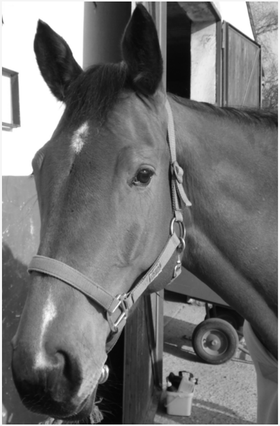
Image Deblurring under Poisson Noise

Constrained smooth optimization problem:

$$\min_{\mathbf{u}\in\mathbb{R}^{n_x\times n_y}} \underbrace{D_{KL}(\mathbf{f},\mathcal{A}\mathbf{u})}_{\text{Kullback–Leibler divergence}} + \frac{\lambda}{2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} \underbrace{\log(1+\mu|(\mathcal{D}\mathbf{u})_{i,j}|^2)}_{\text{smooth non-convex regularizer}} \quad s.t.\ \mathbf{u}_{i,j}\geq 0$$

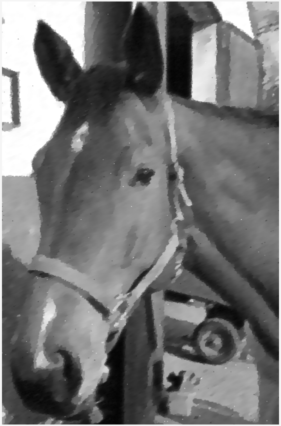
- Even for convex regularization, it is **hard to minimize**.
- Difficulty comes from the **lack of global Lipschitz continuity**.
- For convex regularizer: Use generalized Descent Lemma and Burg's entropy. [Bauschke et al., 2016]
- Burg's entropy is not strongly convex and cannot be used by current FBS.
- Subproblems in our framework have simple **analytic solution**.

Image Deblurring under Poisson Noise



clean

noisy and blurry



reconstruction

Summary:

1. Gradient Descent

- Gradient or Steepest Descent
- Convergence of Gradient Descent
- Convergence to a Single Point
- Speed of Convergence
- Applications
- Structured Optimization Problems
- Unification of Algorithms

2. Acceleration Strategies

- Time Continuous Setting
- Heavy-ball Method
- Nesterov's Acceleration
- Quasi-Newton Methods
- Subspace Acceleration

3. Non-Smooth Optimization

- Basic Definitions
- Infimal Convoution
- Proximal Mapping
- Subdifferential
- Optimality Condition (Fermat's Rule)
- Proximal Point Algorithm
- Forward–Backward Splitting

4. Single Point Convergence

- Łojasiewicz Inequality
- Kurdyka-Łojasiewicz Inequality
- Abstract Convergence Theorem
- Convergence of Non-convex Forward–Backward Splitting
- A Generalized Abstract Convergence Theorem
- Convergence of iPiano
- Local Convergence of iPiano

5. Variants and Acceleration of Forward–Backward Splitting

- FISTA
- Adaptive FISTA
- Proximal Quasi-Newton Methods
- Efficient Solution for Rank-1 Perturbed Proximal Mapping
- Forward–Backward Envelope
- Generalized Forward–Backward Splitting

6. Bregman Proximal Minimization

- Model Function Framework
- Examples of Model Functions
- Examples of Bregman Functions
- Convergence Results
- Applications