Non-smooth Non-convex Bregman Minimization: Unification and new Algorithms

Peter Ochs
Saarland University

— 05.06.2018 —

joint work: Jalal Fadili

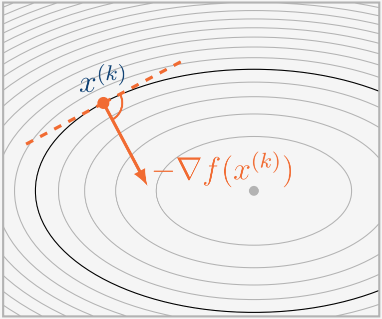
Facts about Gradient Descent

- Smooth optimization problem: (f continuously differentiable)

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

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

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

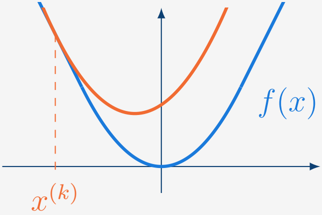


- Step size selection:
  - f **continuously differentiable** $\Rightarrow$ line-search is required.
  - ∇f **Lipschitz continuous** $\Rightarrow$ feasible range of step sizes can be computed.

Facts about Gradient Descent

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

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



- Optimality condition:

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

$$\Leftrightarrow x = x^{(k)} - \tau \nabla f(x^{(k)})$$

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**

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

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

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

- Update step:

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

Facts about Gradient Descent

Generalization to non-smooth functions f:

- Minimization of a convex **model function**

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

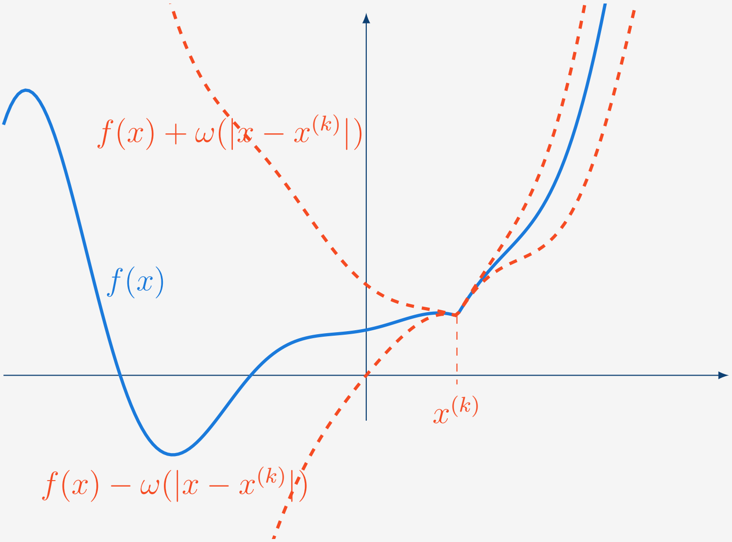
  with **penalty on the distance** to $x^{(k)}$:

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

- Update step:

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

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

Abstract Algorithm:

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

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

PART 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.

PART 2: Examples for Distance Functions

- Bregman distance generated by Legendre functions.

PART 3: Convergence Analysis

- Subsequential convergence to a stationary point.

PART 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:**

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

- **Example:** (Constrained smooth optimization)

$$\min_x f_1(x) \quad s.t.\ x \in C$$

Update step: (Projected gradient descent)

$$\tilde{x}^{(k)} = \operatorname{proj}_C(x^{(k)} - \tau\nabla f_1(x^{(k)}))$$

Forward–Backward Splitting

- **Optimization problem**:

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

- **Model function**:

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

- **Model assumption/error**:

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

- 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**:

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

ProxDescent

- **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)

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

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:**

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

- **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.

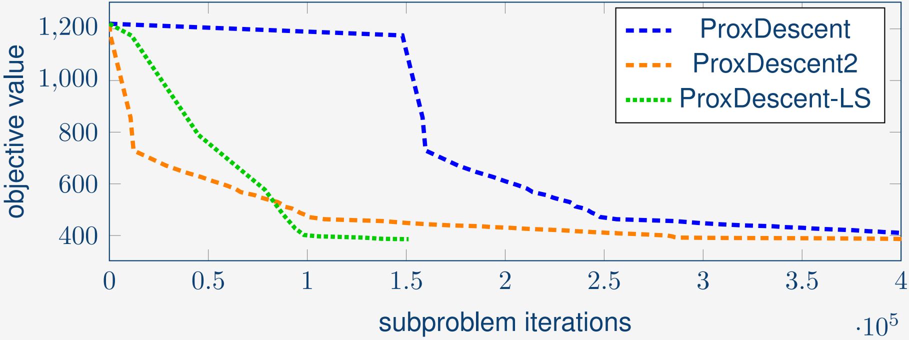
- **Model function** linearizes the inner functions F_i.

- **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.

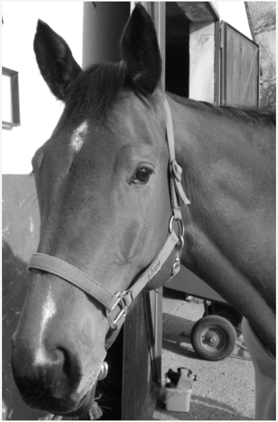
Image Deblurring under Poisson Noise

Constrained smooth optimization problem:

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

- 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

- **Bregman Proximal Minimization Line Search Algorithm**:

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

- **Model assumption:** $f_{\bar{x}}$ is convex and

$$|f(x) - f_{\bar{x}}(x)| \leq \omega(|x - \bar{x}|) \quad \forall x \, .$$

- "Approximation quality" is controlled by a **growth function** ω.
- **Bregman distance** generated by Legendre functions.
- **Unification** of Gradient Descent, FBS, ProxDescent, ..., and variable metric or Bregman versions.