Chapter 6

Stopping criteria, adaptivity, accelerations

Convex Optimization for Computer Vision
SS 2016

Michael Moeller
Thomas Möllenhoff
Emanuel Laude
Computer Vision Group
Department of Computer Science
TU München

Customized proximal point algorithms

Structured optimization methods for

$$\min_u G(u) + F(Ku)$$

under the assumption of F and G being simple or - in the ADMM case - $(\partial G + \frac{1}{\tau} K^T K)^{-1}$ being easy to compute.

Goal: Find pair $(\hat{u}, \hat{p})$ with

$$-K^T \hat{p} \in \partial G(\hat{u}), \quad K\hat{u} \in \partial F^*(\hat{p})$$

Primal Dual-Hybrid Gradient (PDHG) method:

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Customized proximal point algorithms

Primal ADMM or dual Douglas-Rachford

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau} K^T K & -K^T \\ -K & \tau I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}$$

Question for all these algorithms: What is a good stopping criterion? How do we determine if an algorithm converges?

Stopping customized proximal point algorithms

Generic form:

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \underbrace{\begin{bmatrix} M_1 & -K^T \\ -K & M_2 \end{bmatrix}}_{=:M} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}$$

such that the matrix M is positive (semi-)definite.

Natural considerations:

- How close is $-K^T p^{k+1}$ to being an element of $\partial G(u^{k+1})$?
- How close is Ku^{k+1} to being an element of $\partial F^*(p^{k+1})$?

We define the ***primal** and **dual** residuals*:

$$r_p^{k+1} = M_2(p^{k+1} - p^k) - K(u^{k+1} - u^k)$$
$$r_d^{k+1} = M_1(u^{k+1} - u^k) - K^T(p^{k+1} - p^k)$$

Primal and dual residuals

Based on the ***primal** and **dual** residuals*:

$$r_p^{k+1} = M_2(p^{k+1} - p^k) - K(u^{k+1} - u^k)$$
$$r_d^{k+1} = M_1(u^{k+1} - u^k) - K^T(p^{k+1} - p^k)$$

we could consider our algorithm to be convergent if $\|r_d^{k+1}\|^2 + \|r_p^{k+1}\|^2 \to 0$, because this implies

$$\text{dist}(-K^T p^{k+1}, \partial G(u^{k+1})) \to 0,$$
$$\text{dist}(K u^{k+1}, \partial F^*(p^{k+1})) \to 0.$$

Note that this notion of convergences does not imply convergence of u^k and p^k yet!

Nevertheless, we know PDHG and ADMM do converge, and $\|r_d^{k+1}\|$ and $\|r_p^{k+1}\|$ are good measures for convergence!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Upper bounds on the residuals

How should we use $\|r_d^{k+1}\|$ and $\|r_p^{k+1}\|$ to formalize a stopping criterion?

- Simple option: Iterator until $\|r_d^{k+1}\| \leq \epsilon$ and $\|r_p^{k+1}\| \leq \epsilon$.
- Could be unfair, if $u^k \in \mathbb{R}^n$ and $p^k \in \mathbb{R}^m$ and e.g. $n >> m$. Use $\|r_d^{k+1}\| \leq \sqrt{n}\,\epsilon$ and $\|r_p^{k+1}\| \leq \sqrt{m}\,\epsilon$.
- Could be unfair for different scales! Introduce absolute and relative error criteria:

$$
\begin{aligned}
\|r_d^{k+1}\| &\leq \sqrt{n}\,\epsilon^{abs} + \text{dual scale factor} \cdot \epsilon^{rel} \\
\|r_p^{k+1}\| &\leq \sqrt{m}\,\epsilon^{abs} + \text{primal scale factor} \cdot \epsilon^{rel}
\end{aligned}
$$

But what are reasonable scale factors?

Scaling the primal residuum

The primal residual

$$r_p^{k+1} = M_2(p^{k+1} - p^k) - K(u^{k+1} - u^k)$$

measures how far Ku^{k+1} is away from a particular element in $\partial F^*(p^{k+1})$, and therefore scales with the magnitude of elements in $\partial F^*(p^{k+1})$.

More precisely:

$$\begin{aligned}
&0 \in \partial F^*(p^{k+1}) - Ku^{k+1} + r_p^{k+1} \\
\Rightarrow &0 \in \partial F^*(p^{k+1}) - K^T(2u^{k+1} - u^k) + M_2(p^{k+1} - p^k). \\
\Rightarrow &\underbrace{M_2(p^k - p^{k+1}) + K^T(2u^{k+1} - u^k)}_{=: z^{k+1}} \in \partial F^*(p^{k+1})
\end{aligned}$$

Thus, we can use

$$\|r_p^{k+1}\| \leq \sqrt{m}\, \epsilon^{abs} + \|z^{k+1}\| \cdot \epsilon^{rel}$$

to be scale-independent.

Scaling the dual residuum

The dual residual

$$r_d^{k+1} = M_1(u^{k+1} - u^k) - K^T(p^{k+1} - p^k)$$

measures how far $-K^T p^{k+1}$ is away from a particular element in $\partial G(u^{k+1})$, and therefore scales with the magnitude of elements in $\partial G(u^{k+1})$.

More precisely:

$$\begin{aligned} & 0 \in \partial G(u^{k+1}) + K^T p^{k+1} + r_d^{k+1}. \\ \Rightarrow\ & 0 \in \partial G(u^{k+1}) + K^T p^k + M_1(u^{k+1} - u^k) \\ \Rightarrow\ & \underbrace{M_1(u^k - u^{k+1}) - K^T p^k}_{=:v^{k+1}} \in \partial G(u^{k+1}) \end{aligned}$$

Thus, we can use

$$\|r_d^{k+1}\| \leq \sqrt{n}\, \epsilon^{abs} + \|v^{k+1}\| \cdot \epsilon^{rel}$$

to be scale-independent.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

A scaled absolute and relative stopping criterion

In summary, a good stopping criterion is

$$\|r_p^{k+1}\| \leq \sqrt{m}\, \epsilon^{abs} + \|z^{k+1}\| \cdot \epsilon^{rel},$$
$$\|r_d^{k+1}\| \leq \sqrt{n}\, \epsilon^{abs} + \|v^{k+1}\| \cdot \epsilon^{rel}.$$

Interesting observation in our previous considerations: ADMM, Douglas Rachford, PDHG, and any other ”customized proximal point” algorithm actually generates iterates $(u^{k+1}, p^{k+1}, v^{k+1}, z^{k+1})$ with

$$v^{k+1} \in \partial G(u^{k+1}), \qquad z^{k+1} \in \partial F^*(p^{k+1}).$$

The goal of all algorithms is to achieve convergence

$$\| \underbrace{z^{k+1} - Ku^{k+1}}_{=r_p^{k+1}} \| \to 0 \ \text{ and } \| \underbrace{v^{k+1} + K^T p^{k+1}}_{=r_d^{k+1}} \| \to 0!$$

Note that z is exactly the ”split” variable in the augmented Lagrangian based derivation of ADMM!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Adaptive stepsizes

r_p^{k+1} and r_d^{k+1} determine the convergence of the algorithm.

Can we also use r_d and r_p to accelerate the algorithm?

Adaptive stepsizes:

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau^k} M_1 & -K^T \\ -K & \frac{1}{\sigma^k} M_2 \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}$$

Base the choices of τ^k and σ^k on the residuals r_p^k and r_d^k?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Residual balancing

First option: Residual balancing! Let $(M_1, -K^T; -K, M_2)$ be positive definite. Pick τ^0 and σ^0 with $\tau^0\sigma^0 < 1$ as well as $\mu > 1$, $\alpha > 1$:

- If $\|r_p^k\| > \mu\|r_d^k\|$, do

$$\tau^{k+1} = \frac{1}{\alpha}\tau^k, \quad \sigma^{k+1} = \alpha\sigma^k$$

- If $\|r_d^k\| > \mu\|r_p^k\|$, do

$$\tau^{k+1} = \alpha\tau^k, \quad \sigma^{k+1} = \frac{1}{\alpha}\sigma^k$$

- Keep $\tau^{k+1} = \tau^k$ and $\sigma^{k+1} = \sigma^k$ otherwise.

Why could this make sense?

Unbalanced adaption

Second option: Fougner, Boyd '15: Let $(M_1, -K^T; -K, M_2)$ be positive definite. Pick τ^0 and σ^0 with $\tau^0\sigma^0 < 1$ as well as $\mu > 1$, $\alpha > 1$:

- If $\|r_d^k\| < \epsilon^{thresh}$ and $k > \mu k_1^{prev}$, do

$$\tau^{k+1} = \frac{1}{\alpha}\tau^k, \quad \sigma^{k+1} = \alpha\sigma^k, \quad k_1^{prev} \leftarrow k.$$

- If $\|r_p^k\| < \epsilon^{thresh}$ and $k > \mu k_2^{prev}$, do

$$\tau^{k+1} = \alpha\tau^k, \quad \sigma^{k+1} = \frac{1}{\alpha}\sigma^k, \quad k_2^{prev} \leftarrow k.$$

- Keep $\tau^{k+1} = \tau^k$ and $\sigma^{k+1} = \sigma^k$ otherwise.

Convergence guarantees?

The previous two adaptive step size methods are heuristics that work well in practice.

In general, they have no convergence guarantees!

Common trick: Changing the parameters finitely many times only, reestablishes the convergence guarantees!

More appealing from a theoretical point of view: Decreasing the adaptivity of the stepsizes fast enough.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Convergence guarantees with adaptive step sizes

Goldstein et al. 2015

Consider $M_1 = \frac{1}{\tilde{\tau}} I$, $M_2 = \frac{1}{\tilde{\sigma}} I$ with $\tilde{\sigma}\tilde{\tau} < \|K\|^2$, and define

$$\delta^k = \min\left\{ \frac{\tau^{k+1}}{\tau^k}, \frac{\sigma^{k+1}}{\sigma^k}, 1 \right\}, \quad \phi^k = 1 - \delta^k$$

Let the following three conditions hold:

1. The sequences $\{\tau^k\}$, $\{\sigma^k\}$ remain bounded.
2. The sequence ϕ^k is summable.
3. It holds that $\tau^k \sigma^k < c < 1$.

Then the resulting adaptive PDHG algorithm converges.

Conjecture (for you to prove)

The same result holds for arbitrary M_1, M_2 provided that the matrix $(M_1, -K^T; -K, M_2)$ is positive definite.

Customized proximal point algorithms

Decreasing residual balancing: Let $(M_1, -K^T; -K, M_2)$ be positive definite. Pick τ^0 and σ^0 with $\tau^0\sigma^0 < 1$. Further choose $\mu > 1$, $\alpha^0 < 1$, $\beta < 1$ and adapt as follows

- If $\|r_p^k\| > \mu\|r_d^k\|$, do

$$\tau^{k+1} = (1-\alpha^k)\tau^k, \quad \sigma^{k+1} = \frac{1}{1-\alpha^k}\sigma^k, \quad \alpha^{k+1} = \alpha^k \cdot \beta.$$

- If $\|r_d^k\| > \mu\|r_p^k\|$, do

$$\tau^{k+1} = \frac{1}{1-\alpha^k}\tau^k, \quad \sigma^{k+1} = (1-\alpha^k)\sigma^k, \quad \alpha^{k+1} = \alpha^k \cdot \beta.$$

- Keep $\tau^{k+1} = \tau^k$, $\sigma^{k+1} = \sigma^k$, and $\alpha^{k+1} = \alpha^k$ otherwise.

Convergence proof based on previous theorem.

Sketch of proof

Sketch of the proof:

- The product $\tau^k \sigma^k$ does not change, thus 3. holds.
- It holds that

$$\phi^k = \begin{cases} 0 & \text{if stepsizes were not updated,} \\ \alpha^k & \text{if stepsizes were updated.} \end{cases}$$

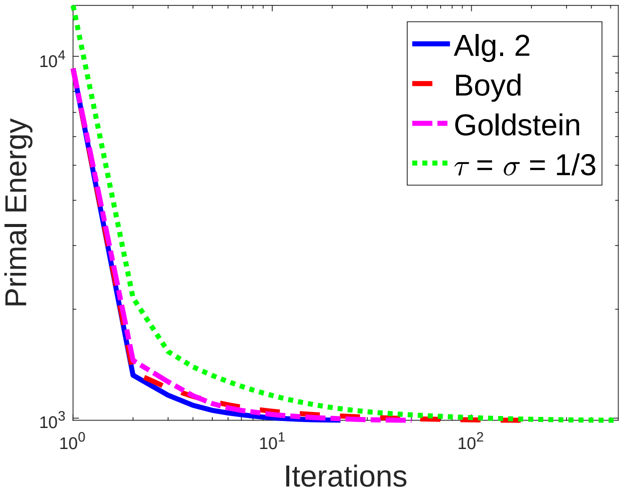
  which means the j-th nonzero entry of $\{\phi^k\}$ is $(\alpha^0)^j$.

- $\sum_k \phi^k = \sum_{j \in I} (\alpha^0)^j < C$, thus condition 2 holds.
- Without restriction of generality we may drop those steps where the stepsize remained unchanged. We find

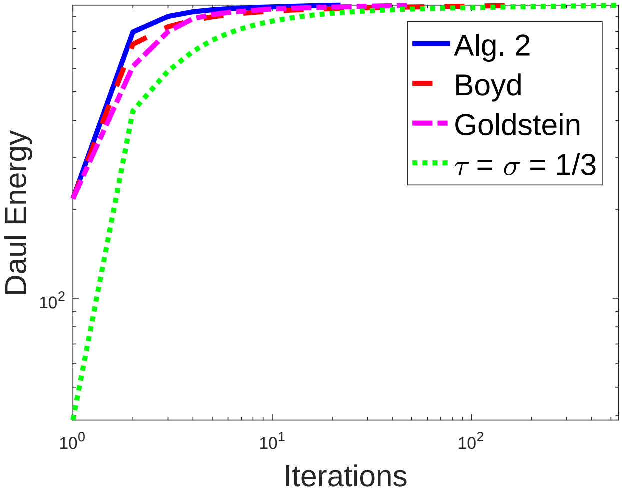
$$\tau^{j+1} \leq \frac{1}{1-\alpha^j} \tau^j \leq \left(\frac{1}{1-\alpha^j} \right)^j \tau^0 = \frac{1}{(1-\alpha^0 \beta^j)^j} \tau^0$$

  The factor $(1-\alpha^0\beta^j)^j$ remains bounded from below and thus condition 1 follows. (For $x \geq -1$: $(1+x)^n \geq 1 + nx$)

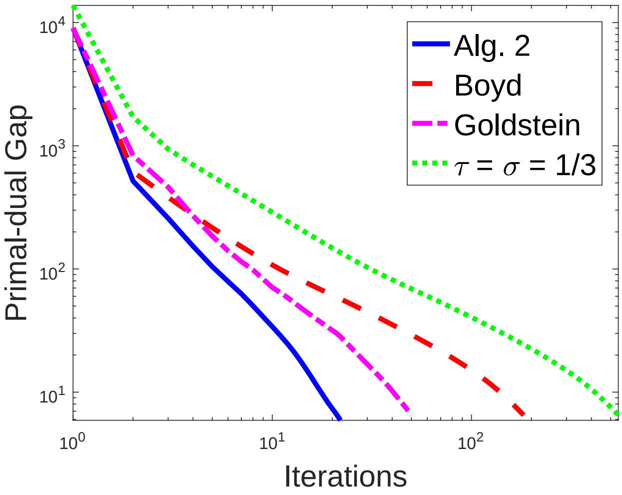
Example plot of convergence for ROF model



Example plot of convergence for ROF model



Example plot of convergence for ROF model



Backtracking

Condition 3 in the previous convergence result for adaptive stepsizes can also be weakened to

3. The saddle point problem

$$\min_u \max_p G(u) + \langle Ku, p \rangle - F^*(p)$$

restricts either u or p to a bounded set. Furthermore there exists a constant c such that for all $k > 0$

$$\left\langle \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}, \begin{bmatrix} \frac{1}{\tau^k} M_1 & -K^T \\ -K & \frac{1}{\sigma^k} M_2 \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix} \right\rangle$$
$$\geq c \left\langle \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}, \begin{bmatrix} \frac{1}{\tau^k} M_1 & 0 \\ 0 & \frac{1}{\sigma^k} M_2 \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix} \right\rangle .$$

Under this condition the convergence result still holds.

Backtracking

The stability condition 3 from the previous slide can used to define a *backtracking* algorithm that works without knowing the constant $\|K\|^2$.

Define

$$b^k = \frac{2\tilde{\tau}\tilde{\sigma}\tau^k\sigma^k\langle p^{k+1} - p^k, K(u^{k+1} - u^k)\rangle}{\gamma\tilde{\sigma}\sigma^k\|u^{k+1} - u^k\|^2 + \gamma\tilde{\tau}\tau^k\|p^{k+1} - p^k\|^2}$$

for some $\gamma \in]0, 1[$.

If $b^k \leq 1$ keep iterating, if $b^k > 1$ update

$$\tau^{k+1} = \beta\tau^k / b^k, \quad \sigma^{k+1} = \beta\sigma^k / b^k$$

for $\beta \in]0, 1[$.

Key insight to prove convergence: $b^k > 1$ can only happen finitely many times.

Next lecture: Preconditioning

Generic customized proximal point algorithm:

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} M_1 & -K^T \\ -K & M_2 \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}$$

We have seen:

- $M_1 = \lambda K^T K$, $M_2 = \frac{1}{\lambda} I$ yields ADMM
- $M_1 = \frac{1}{\tau} I$, $M_2 = \frac{1}{\sigma} I$ yields PDHG

Are there different choices for M_1 and M_2 that make sense and are possibly more efficient?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude