TELECOM ParisTech

Institut Mines-Télécom

Programme PGMO

Pour l'optimisation et la recherche opérationnelle

A Smooth Primal-Dual Optimization Framework for Nonsmooth Composite Minimization

Q. Tran-Dinh, O. Fercoq and V. Cevher



Constrained convex optimisation

$$\min_{x \in \mathcal{X}} \; f(x)$$
$$\text{st: } Ax = c$$

- $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is a l.s.c. convex function with a simple proximal operator

$$\text{prox}_f(x) = \arg\max_{z \in \mathcal{X}} f(z) + \frac{1}{2}\|z - x\|_{\mathcal{X}}^2$$

- $A : \mathcal{X} \to \mathcal{Y}$ is a linear map, $\quad c \in \mathcal{Y}$
- Can also be written as $\min_{x \in \mathcal{X}} f(x) + g(Ax)$

where $g(z) = \begin{cases} 0 & \text{if } z = c \\ +\infty & \text{otherwise} \end{cases}$

Duality gap

If $0 \in \mathrm{ri}(\mathrm{dom}\, g - A\,\mathrm{dom}\, f)$ then the optimization problem is equivalent to

$$\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x) + \langle y, Ax \rangle - g^*(y)$$

where g^* is the Fenchel conjugate of g: $g^*(y) = \langle c, y \rangle$.

We define the duality gap as

$$\begin{aligned} G(x, y) &= f(x) + g(Ax) + f^*(-A^\top y) + g^*(y) \\ &= \max_{\bar{y}} \left(f(x) + \langle \bar{y}, Ax \rangle - g^*(\bar{y}) \right) - \min_{\bar{x}} \left(f(\bar{x}) + \langle y, A\bar{x} \rangle - g^*(y) \right) \end{aligned}$$

Famous algorithms: Augmented Lagrangian

At iteration k:

$$x^{k+1} \in \arg\min_{x \in \mathcal{X}} f(x) + \langle y^k, Ax - c \rangle + \frac{\beta}{2}\|Ax - c\|_{\mathcal{Y}}^2$$
$$y^{k+1} = y^k + \beta(Ax^{k+1} - c)$$

Problem: needs to solve a potentially difficult optimization problem at each iteration (same issue with ADMM)

Famous algorithms: Chambolle-Pock

At iteration k:

$$x^{k+1} \in \arg\min_{x\in\mathcal{X}} f(x) + \langle y^k, Ax - c\rangle + \frac{\beta}{2}\|x - x^k\|^2_{\mathcal{X}}$$

$$y^{k+1} = y^k + \frac{\beta - \epsilon}{\|A\|^2_{\mathcal{X},\mathcal{Y}}}\big(A(2x^{k+1} - x^k) - c\big)$$

Convergence: $G(x^k, y^k) \le \frac{1}{k}\left(\frac{\beta}{2}D^2_{\mathcal{X}} + \frac{\|A\|^2_{\mathcal{X},\mathcal{Y}}}{2(\beta-\epsilon)}D^2_{\mathcal{Y}}\right)$
where $D_{\mathcal{X}}$ is the diameter of dom f and $D_{\mathcal{Y}}$ is dom g^*'s.

Problem: Here, $D_{\mathcal{Y}} = +\infty$.

Smoothing the indicator function

- If $Ax^k \neq c$, then $f(x^k) + g(Ax^k) = +\infty$ and so $G(x^k, y^k) = +\infty$.
- We define the smoothed function [Nesterov 2005]

$$g_\beta(z; \dot{y}) = \max_y \langle z, y\rangle - g^*(y) - \frac{\beta}{2}\|y - \dot{y}\|_{\mathcal{Y}}^2$$

- g_β is differentiable wrt z and $\nabla_z g_\beta$ is $\frac{1}{\beta}$-Lipschitz
- $g_\beta(Ax^k, \dot{y}) = \langle \dot{y}, Ax^k - c\rangle + \frac{1}{2\beta}\|Ax^k - c\|_{\mathcal{Y},*}^2$
- Smoothing the duality gap we get $G_{\beta,\gamma}(x^k, y^k, \dot{x}, \dot{y}) < +\infty$

Fundamental theorem

Theorem

Denote $S_\beta(x, \dot{y}) = f(x) + g_\beta(Ax; \dot{y}) - f(x^\star)$. *We have*

$$\|Ax - c\|_{\mathcal{Y},*} \leq \beta\left[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \left(\|y^\star - \dot{y}\|^2_{\mathcal{Y}} + 2\beta^{-1}S_\beta(x; \dot{y})\right)^{1/2}\right]$$

$$f(x) - f(x^\star) \geq -\|y^\star\|_{\mathcal{Y}}\|Ax - c\|_{\mathcal{Y},*}$$

$$f(x) - f(x^\star) \leq S_\beta(x, \dot{y}) + \|y^\star\|_{\mathcal{Y}}\|Ax - c\|_{\mathcal{Y},*} + \frac{\beta}{2}\|y^\star - \dot{y}\|^2_{\mathcal{Y}}$$

If β and $S_\beta(x, \dot{y})$ are small, we have an approximate solution in feasibility and function value

Accelerated Smoothed GAp ReDuction algorithm (ASGARD)

Idea: FISTA on $f(x) + g_\beta(Ax; \dot{y})$ and continuation on β

For $k = 0$ **to** $k_{\max}$:

1: $\tau_{k+1} \in (0,1)$ root of $\tau^3 + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$

2: $y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) = \arg\max_{y\in\mathcal{Y}} \langle A\hat{x}^k, y\rangle - g^*(\hat{y}) - \frac{\beta_{k+1}}{2}\|y - \dot{y}\|^2_{\mathcal{Y}}$

3: $\bar{x}^{k+1} := \mathrm{prox}_{\beta_{k+1}\|A\|^{-2}f}\left(\hat{x}^k - \beta_{k+1}\|A\|^{-2}A^\top y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y})\right)$

4: $\hat{x}^{k+1} = \bar{x}^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(\bar{x}^{k+1} - \bar{x}^k)$

5: $\beta_{k+2} := \frac{\beta_{k+1}}{1+\tau_{k+1}}$

End for

Convergence theorem 1/2

Lemma

If $\tau_0 = 1$ *and* $\forall k \geq 1$, $\tau_k^3 + \tau_k^2 + \tau_{k-1}^2\tau_k - \tau_{k-1}^2 = 0$, $\beta_k := \frac{\beta_{k-1}}{1+\tau_{k-1}}$:

- $$S_{\beta_{k+1}}(\bar{x}^{k+1}, \dot{y}) + \frac{\|A\|^2\tau_k^2}{\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|_{\mathcal{X}}^2$$
$$\leq (1-\tau_k)S_{\beta_k}(\bar{x}^k, \dot{y}) + \frac{\|A\|^2\tau_k^2}{\beta_{k+1}}\|\tilde{x}^k - x^\star\|_{\mathcal{X}}^2$$
where $\tilde{x}^{k+1} = \frac{1}{\tau_k}(\bar{x}^{k+1} - (1-\tau_k)\bar{x}^k)$
- $\beta_k \leq \dfrac{2\beta_1}{k+1}$

Decrease of the smoothed optimality gap and of the parameter

Convergence theorem 2/2

Theorem

The iterates of ASGARD satisfy

$$\|A\bar{x}^k - c\|_{\mathcal{Y},*} \leq \frac{\beta_1}{k+1}\left[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \sqrt{\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 + \frac{\|A\|^2}{\beta_1^2}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2}\right]$$

$$f(\bar{x}^k) - f(x^\star) \geq -\|y^\star\|_{\mathcal{Y}}\|Ax - c\|_{\mathcal{Y},*}$$

$$f(\bar{x}^k) - f(x^\star) \leq \frac{1}{k}\frac{\|A\|^2}{2\beta_1}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2 + \|y^\star\|_{\mathcal{Y}}\|A\bar{x}^k - c\|_{\mathcal{Y},*} + \frac{\beta_1}{k+1}\|y^\star - \dot{y}\|^2$$

$O(1/k)$ convergence in function value and feasibility

Restarting

- Restarting accelerated gradient methods speeds up the convergence when minimizing strongly convex functions [Nesterov 2007, O'Donoghue & Candes 2012]
- Why not restart ASGARD?

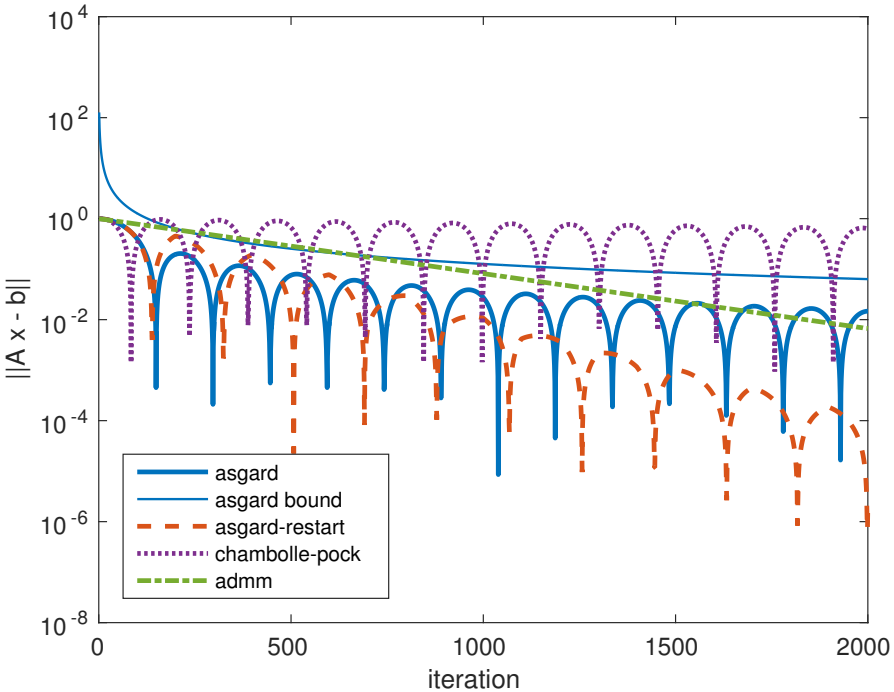
  If $k \equiv 0 \;\; [K]$:

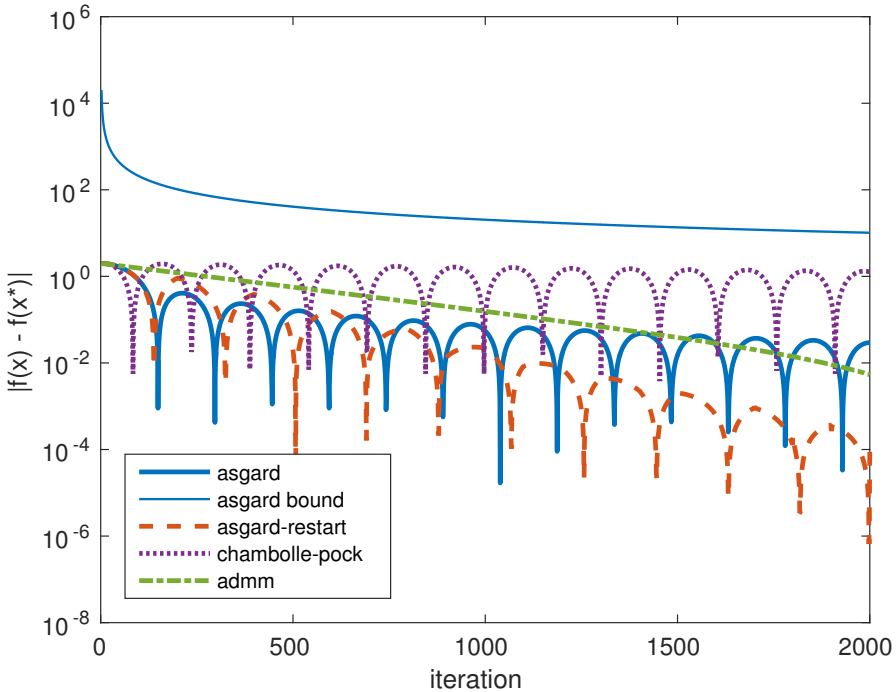
$$\left\{\begin{array}{lcl} \hat{x}^{k+1} & \leftarrow & \bar{x}^{k+1} \\ \tau_{k+1} & \leftarrow & 1 \\ \beta_{k+1} & \leftarrow & \beta_1 \\ \dot{y} & \leftarrow & y^*_{\beta_{k+1}}(A\bar{x}^k; \dot{y}) \end{array}\right.$$

  Update of the center motivated by
  $\frac{1}{2}\|y^*_\beta(A\bar{x}; \dot{y}) - y^\star\|^2_{\mathcal{Y}} \leq \frac{1}{2}\|y^\star - \dot{y}\|^2_{\mathcal{Y}} + \frac{1}{\beta}S_\beta(\bar{x}, \dot{y})$

Numerical experiment on a degenerate LP

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & 2x_n \\ \text{s.t.} \quad & x_n \geq 0 \qquad \sum_{k=1}^{n-1} x_k = 1 \\ & x_n - \sum_{k=1}^{n-1} x_k = 0 \quad (2 \leq j \leq d) \\ (n = 10, \quad & d = 200) \end{aligned}$$





Conclusion

- Convergence speed for constrained problems solved by a first order method using:
  - Smoothing
  - Accelerated gradient methods
  - Continuation
- Extensions we've done:
  - Primal-dual version of the algorithm
  - Augmented Lagrangian smoother
  - Dealing with smooth functions with their gradients
- Perspective (PGMO project): convergence speed for a primal-dual coordinate descent method