Accelerated primal-dual methods for linearly constrained convex problems

Yangyang Xu

SIAM Conference on Optimization

May 24, 2017

Accelerated proximal gradient

For convex composite problem: $\underset{x}{\text{minimize}}\ F(x) := f(x) + g(x)$

- f: convex and Lipschitz differentiable
- g: closed convex (possibly nondifferentiable) and simple

Proximal gradient:

$$x^{k+1} = \arg\min_x \langle \nabla f(x^k), x\rangle + \frac{L_f}{2}\|x - x^k\|^2 + g(x)$$

- convergence rate: $F(x^k) - F(x^*) = O(1/k)$

Accelerated Proximal gradient [Beck-Teboulle'09, Nesterov'14]:

$$x^{k+1} = \arg\min_x \langle \nabla f(\hat{x}^k), x\rangle + \frac{L_f}{2}\|x - \hat{x}^k\|^2 + g(x)$$

- $\hat{x}^k$: extrapolated point
- convergence rate (with smart extrapolation): $F(x^k) - F(x^*) = O(1/k^2)$

This talk: ways to accelerate primal-dual methods

Part I: accelerated linearized augmented Lagrangian

Affinely constrained composite convex problems

$$\operatorname*{minimize}_{x} F(x) = f(x) + g(x), \quad \text{subject to } Ax = b \qquad \text{(LCP)}$$

- f: convex and Lipschitz differentiable
- g: closed convex and simple

Examples

- nonnegative quadratic programming: $f = \frac{1}{2}x^\top Q x + c^\top x$, $g = \iota_{\mathbb{R}^n_+}$
- TV image denoising: $\min\{\frac{1}{2}\|X - B\|_F^2 + \lambda\|Y\|_1, \text{ s.t. } \mathcal{D}(X) = Y\}$

Augmented Lagrangian method (ALM)

At iteration k,

$$x^{k+1} \leftarrow \arg\min_x f(x) + g(x) - \langle \lambda^k, Ax \rangle + \frac{\beta}{2}\|Ax - b\|^2,$$

$$\lambda^{k+1} \leftarrow \lambda^k - \gamma(Ax^{k+1} - b)$$

- augmented dual gradient ascent with stepsize γ
- β: penalty parameter; dual gradient Lipschitz constant $1/\beta$
- $0 < \gamma < 2\beta$: convergence guaranteed
- also popular for (nonlinear, nonconvex) constrained problems

x-subproblem as difficult as original problem

Linearized augmented Lagrangian method

- Linearize the smooth term f:

$$x^{k+1} \leftarrow \arg\min_x \langle \nabla f(x^k), x\rangle + \frac{\eta}{2}\|x - x^k\|^2 + g(x) - \langle \lambda^k, Ax\rangle + \frac{\beta}{2}\|Ax - b\|^2.$$

- Linearize both f and $\|Ax - b\|^2$:

$$x^{k+1} \leftarrow \arg\min_x \langle \nabla f(x^k), x\rangle + g(x) - \langle \lambda^k, Ax\rangle + \langle \beta A^\top r^k, x\rangle + \frac{\eta}{2}\|x - x^k\|^2,$$

where $r^k = Ax^k - b$ is the residual.

Easier updates and nice convergence speed $O(1/k)$

Accelerated linearized augmented Lagrangian method

At iteration k,

$$\hat{x}^k \leftarrow (1-\alpha_k)\bar{x}^k + \alpha_k x^k,$$

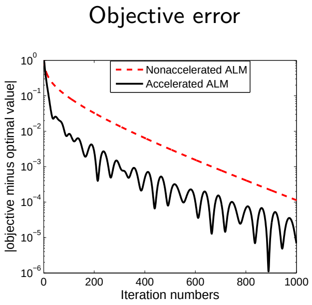
$$x^{k+1} \leftarrow \arg\min_x \langle \nabla f(\hat{x}^k) - A^\top \lambda^k, x\rangle + g(x) + \frac{\beta_k}{2}\|Ax-b\|^2 + \frac{\eta_k}{2}\|x-x^k\|^2,$$

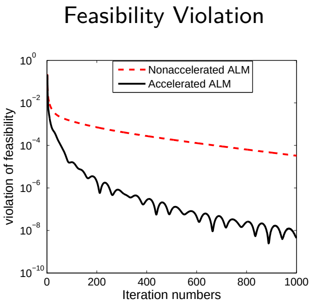
$$\bar{x}^{k+1} \leftarrow (1-\alpha_k)\bar{x}^k + \alpha_k x^{k+1},$$

$$\lambda^{k+1} \leftarrow \lambda^k - \gamma_k(Ax^{k+1}-b).$$

- Inspired by [Lan '12] on accelerated stochastic approximation
- reduces to linearized ALM if $\alpha_k = 1, \beta_k = \beta, \eta_k = \eta, \gamma_k = \gamma, \forall k$
  - convergence rate: $O(1/k)$ if $\eta \geq L_f$ and $0 < \gamma < 2\beta$
- adaptive parameters to have $O(1/k^2)$ (next slides)

Better numerical performance





- Tested on quadratic programming (subproblems solved exactly)
- Parameters set according to theorem (see next slide)
- **Accelerated ALM significantly better**

Guaranteed fast convergence

Assumptions:

- There is a pair of primal-dual solution (x^*, λ^*).
- ∇f is Lipschitz continuous: $\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|$

Convergence rate of order $O(1/k^2)$:

- Set parameters to

$$\forall k: \alpha_k = \frac{2}{k+1}, \gamma_k = k\gamma, \beta_k \geq \frac{\gamma_k}{2}, \eta_k = \frac{\eta}{k},$$

where $\gamma > 0$ and $\eta \geq 2L_f$. Then

$$|F(\bar{x}^{k+1}) - F(x^*)| \leq \frac{1}{k(k+1)} \left(\eta \|x^1 - x^*\|^2 + \frac{4\|\lambda^*\|^2}{\gamma} \right),$$

$$\|A\bar{x}^{t+1} - b\| \leq \frac{1}{k(k+1)\max(1, \|\lambda^*\|)} \left(\eta \|x^1 - x^*\|^2 + \frac{4\|\lambda^*\|^2}{\gamma} \right),$$

Sketch of proof

Let $\Phi(\bar{x}, x, \lambda) = F(\bar{x}) - F(x) - \langle \lambda, A\bar{x} - b \rangle$.

1. Fundamental inequality (for any λ):

$$\begin{aligned} &\Phi(\bar{x}^{k+1}, x^*, \lambda) - (1-\alpha_k)\Phi(\bar{x}^k, x^*, \lambda) \\ \leq& -\tfrac{\alpha_k \eta_k}{2}\left[\|x^{k+1} - x^*\|^2 - \|x^k - x^*\|^2 + \|x^{k+1} - x^k\|^2\right] + \tfrac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2 \\ &+ \tfrac{\alpha_k}{2\gamma_k}\left[\|\lambda^k - \lambda\|^2 - \|\lambda^{k+1} - \lambda\|^2 + \|\lambda^{k+1} - \lambda^k\|^2\right] - \tfrac{\alpha_k \beta_k}{\gamma_k^2}\|\lambda^{k+1} - \lambda^k\|^2, \end{aligned}$$

2. $\alpha_k = \frac{2}{k+1}$, $\gamma_k = k\gamma$, $\beta_k \geq \frac{\gamma_k}{2}$, $\eta_k = \frac{\eta}{k}$ and multiply $k(k+1)$ to the above ineq.:

$$\begin{aligned} &k(k+1)\Phi(\bar{x}^{k+1}, x^*, \lambda) - k(k-1)\Phi(\bar{x}^k, x^*, \lambda) \\ \leq& -\eta\Big[\|x^{k+1} - x^*\|^2 - \|x^k - x^*\|^2\Big] + \frac{1}{\gamma}\Big[\|\lambda^k - \lambda\|^2 - \|\lambda^{k+1} - \lambda\|^2\Big]. \end{aligned}$$

3. Set $\lambda^1 = 0$ and sum the above inequality over k:

$$\Phi(\bar{x}^{k+1}, x^*, \lambda) \leq \frac{1}{k(k+1)}\left(\eta\|x^1 - x^*\|^2 + \frac{1}{\gamma}\|\lambda\|^2\right)$$

4. Take $\lambda = \max\left(1 + \|\lambda^*\|, 2\|\lambda^*\|\right)\frac{A\bar{x}^{k+1} - b}{\|A\bar{x}^{k+1} - b\|}$ and use the optimality condition $\Phi(\bar{x}, x^*, \lambda^*) \geq 0 \Rightarrow F(\bar{x}^{k+1}) - F(x^*) \geq -\|\lambda^*\| \cdot \|A\bar{x}^{k+1} - b\|$

Literature

- [He-Yuan '10]: accelerated ALM to $O(1/k^2)$ for smooth problems
- [Kang et. al '13]: accelerated ALM to $O(1/k^2)$ for nonsmooth problems
- [Huang-Ma-Goldfarb '13]: accelerated linearized ALM (with linearization of augmented term) to $O(1/k^2)$ for strongly convex problems

Part II: accelerated linearized ADMM

Two-block structured problems

Variable is partitioned into two blocks, smooth part involves one block, and nonsmooth part is *separable*

$$\underset{y,z}{\text{minimize}}\, h(y) + f(z) + g(z), \quad \text{subject to } By + Cz = b \qquad \text{(LCP-2)}$$

- f convex and Lipschitz differentiable
- g and h closed convex and simple

Examples:

- Total-variation regularized regression: $\left\{\min_{y,z} \lambda\|y\|_1 + f(z),\ \text{s.t. } \mathcal{D}z = y\right\}$

Alternating direction method of multipliers (ADMM)

At iteration k,

$$y^{k+1} \leftarrow \arg\min_y h(y) - \langle \lambda^k, By \rangle + \frac{\beta}{2}\|By + Cz^k - b\|^2,$$

$$z^{k+1} \leftarrow \arg\min_z f(z) + g(z) - \langle \lambda^k, Cz \rangle + \frac{\beta}{2}\|By^{k+1} + Cz - b\|^2,$$

$$\lambda^{k+1} \leftarrow \lambda^k - \gamma(By^{k+1} + Cz^{k+1} - b)$$

- $0 < \gamma < \frac{1+\sqrt{5}}{2}\beta$: convergence guaranteed [Glowinski-Marrocco'75]
- updating y, z alternatingly: easier than jointly update
  - but z-subproblem can still be difficult

Accelerated linearized ADMM

At iteration k,

$$y^{k+1} \leftarrow \arg\min_y h(y) - \langle \lambda^k, By \rangle + \frac{\beta_k}{2}\|By + Cz^k + -b\|^2,$$

$$z^{k+1} \leftarrow \arg\min_z \langle \nabla f(z^k) - C^\top \lambda^k + \beta_k C^\top r^{k+\frac{1}{2}}, z \rangle + g(z) + \frac{\eta_k}{2}\|z - z^k\|^2,$$

$$\lambda^{k+1} \leftarrow \lambda^k - \gamma_k (By^{k+1} + Cz^{k+1} - b)$$

where $r^{k+\frac{1}{2}} = By^{k+1} + Cz^k - b$.

- reduces to linearized ADMM if $\beta_k = \beta, \eta_k = \eta, \gamma_k = \gamma, \forall k$
  - convergence rate: $O(1/k)$ if $0 < \gamma \leq \beta$ and $\eta \geq L_f + \beta\|C\|^2$
- $O(1/k^2)$ if adaptive parameters and strong convexity on z (next two slides)

Accelerated convergence speed

Assumptions:

- Existence of a pair of primal-dual solution (y^*, z^*, λ^*)
- ∇f Lipschitz continuous: $\|\nabla f(\hat{z}) - \nabla f(\tilde{z})\| \leq L_f \|\hat{z} - \tilde{z}\|$
- f strongly convex with modulus μ_f (not required for y)

Convergence rate of order $O(1/k^2)$

- Set parameters as follows (with $\gamma > 0$ and $\gamma < \eta \leq \mu_f/2$)

  $$\forall k: \;\; \beta_k = \gamma_k = (k+1)\gamma, \;\; \eta_k = (k+1)\eta + L_f,$$

  Then

  $$\max\left(\|z^k - z^*\|^2, \, |F(\bar{y}^k, \bar{z}^k) - F^*|, \|B\bar{y}^k + C\bar{z}^k - b\|\right) \leq O(1/k^2),$$

  where $F(y, z) = h(y) + f(z) + g(z)$ and $F^* = F(y^*, z^*)$.

Sketch of proof

1. Fundamental inequality from optimality conditions of each iterate:

$$\begin{aligned}
&F(y^{k+1}, z^{k+1}) - F(y, z) - \langle \lambda, By^{k+1} + Cz^{k+1} - b\rangle \\
\le& - \left\langle \tfrac{1}{\gamma_k}(\lambda^k - \lambda^{k+1}), \lambda - \lambda^k + \tfrac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1}) - \beta_k C(z^{k+1} - z^k)\right\rangle \\
&+ \tfrac{L_f}{2}\|z^{k+1} - z^k\|^2 - \tfrac{\mu_f}{2}\|z^k - z\|^2 - \eta_k \langle z^{k+1} - z, z^{k+1} - z^k\rangle,
\end{aligned}$$

2. Plug in parameters and bound cross terms:

$$\begin{aligned}
&F(y^{k+1}, z^{k+1}) - F(y^*, z^*) - \langle \lambda, By^{k+1} + Cz^{k+1} - b\rangle \\
&+ \tfrac{1}{2}\left(\eta(k+1)\|z^{k+1} - z^*\|^2 + L_f\|z^{k+1} - z^*\|^2\right) + \tfrac{1}{2\gamma(k+1)}\|\lambda - \lambda^{k+1}\|^2 \\
\le& \tfrac{1}{2}\left(\eta(k+1)\|z^k - z^*\|^2 + (L_f - \mu_f)\|z^k - z^*\|^2\right) + \tfrac{1}{2\gamma(k+1)}\|\lambda - \lambda^k\|^2.
\end{aligned}$$

3. Multiply $k + k_0$ (here $k_0 \sim \frac{2L_f}{\mu_f}$) and sum the inequality over k:

$$F(\bar{y}^{k+1}, \bar{z}^{k+1}) - F(y^*, z^*) - \langle \lambda, B\bar{y}^{k+1} + C\bar{z}^{k+1} - b\rangle \le \frac{\phi(y^*, z^*, \lambda)}{k^2}$$

4. Take a special λ and use KKT conditions

Literature

- [Ouyang et. al'15]: $O(L_f/k^2 + C_0/k)$ with only weak convexity
- [Goldstein et. al'14]: $O(1/k^2)$ with strong convexity on both y and z
- [Chambolle-Pock'11, Chambolle-Pock'16, Dang-Lan'14, Bredies-Sun'16]: accelerated first-order methods on bilinear saddle-point problems

Open question: weakest conditions to have $O(1/k^2)$

Numerical experiments

(More results in paper)

Accelerated (linearized) ADMM

Tested problem: total-variation regularized image denoising

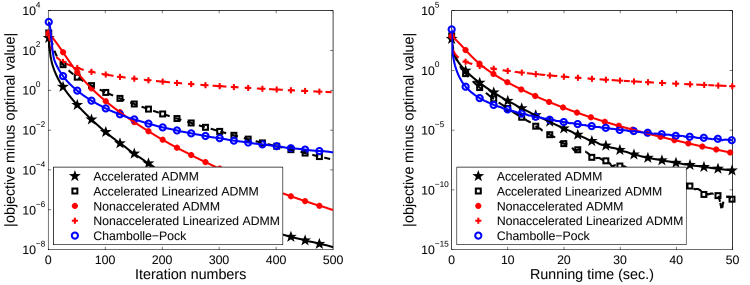
$$\underset{X,Y}{\text{minimize}}\ \frac{1}{2}\|X - B\|_F^2 + \mu\|Y\|_1, \quad \text{subject to } \mathcal{D}X = Y. \qquad \text{(TVDN)}$$

- B observed noisy Cameraman image, and $\mathcal{D}$ finite difference operator

Compared methods:

- original ADMM
- accelerated ADMM
- linearized ADMM
- accelerated linearized ADMM
- accelerated Chambolle-Pock

Performance of compared methods



- Accelerated (linearized) ADMM significantly better than nonaccelerated one
- (accelerated) ADMM faster than (accelerated) linearized ADMM regarding iteration number (but the latter takes less time)

Conclusions

- accelerated linearized ALM to $O(1/k^2)$ from $O(1/k)$ with merely convexity
- accelerated (linearized) ADMM to $O(1/k^2)$ from $O(1/k)$ with strong convexity on one block variable
- performed numerical experiments

References

1. **Y. Xu**. *Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming*, SIAM J. Optimization, 2017.
2. T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. *Fast alternating direction optimization methods*, SIAM J. on Imaging Sciences, 2014.
3. B. He and X. Yuan. *On the acceleration of augmented Lagrangian method for linearly constrained optimization*, Optimization Online, 2010.
4. B. Huang, S. Ma, and D. Goldfarb. *Accelerated linearized Bregman method*, Journal of Scientific Computing, 2013.
5. M. Kang, S. Yun, H. Woo, and M. Kang. *Accelerated bregman method for linearly constrained ℓ_1-ℓ_2 minimization*, Journal of Scientific Computing, 2013.