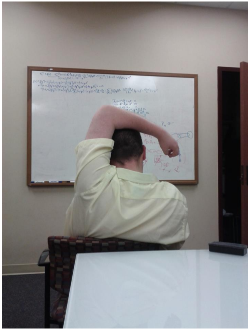
Accelerated Projected Gradient Descent method

Lukáš Pospíšil

Department of Applied Mathematics,
VŠB - Technical University of Ostrava
Czech Republic

Summer School, 14 July - 11 August 2013, Wisconsin - Madison

My Summer School research



- study of Accelerated projected gradient descent method
- attempt to modify it using QP

APGD

APGD method

- for solving

$$\min_{x \in \Omega} f(x)$$

  where
  - $\Omega \subset \mathbb{R}^n$ is closed convex set
  - $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is Lipschitz continuous continuously differentiable convex function
- uses estimate sequences and projected gradients
- primal algorithm (does not use duality and Lagrange multipliers)
- I understood almost every equality and inequality and found tricky parts of derivation (in my report).
- in this presentation just basic ideas

Basic notations

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *Lipschitz continuous* if there exist constant $L > 0$ such that

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

Any such L is referred to as a *Lipschitz constant* for the function f.

Definition

A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called *strongly convex* on $\mathbb{R}^n$ if there exist constant $\mu > 0$ such that

$$\forall x, y \in \mathbb{R}^n : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \ .$$

Any such μ is referred to as a *constant of strong convexity* for the function f.

First theorem about projected gradient

Theorem

Let f is a Lipschitz continuous function, $\gamma \geq L$, $y \in \mathbb{R}^n$, then

$$\forall x \in \Omega : \quad f(x) \geq f(y^P) + \langle g^P(y), x - y \rangle + \frac{1}{2\gamma} \|g^P(y)\|^2 + \frac{\mu}{2} \|x - y\|^2 \ ,$$

where

$$\begin{aligned} g^P(x) \quad &:= \quad \gamma(x - x^P) \ , \\ x^P \quad &:= \quad P_\Omega(x - \tfrac{1}{\gamma} \nabla f(x)) \ . \end{aligned}$$

Corollary

For all $y \in \mathbb{R}^n$

$$f(y^P) \leq f(y) - \frac{1}{2\gamma} \|g^P(y)\|^2$$

In algorithm can be used for adaptive step-size computation.

Estimate sequence

Definition

A pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}, \phi_k \in \mathcal{F}$ and $\{\sigma_k\}_{k=0}^{\infty}, \sigma_k \in \mathbb{R}^+$ is an *estimate sequence* of object function f if

$$\begin{aligned} & \sigma_k && \to && 0 \\ \forall k \geq 0, \forall x \in \Omega : \quad & \phi_k(x) && \leq && (1-\sigma_k)f(x) + \sigma_k \phi_0(x) \ . \end{aligned}$$

Lemma

If for a sequence $\{x_k\}$ we have

$$f(x_k) \leq \min_{x \in \Omega} \phi_k(x),$$

then

$$f(x_k) - f(x^*) \leq \sigma_k(\phi_0(x^*) - f(x^*)) \to 0 \ .$$

Short proof

Lemma

If for a sequence $\{x_k\}$ we have

$$f(x_k) \leq \min_{x \in \Omega} \phi_k(x),$$

then

$$f(x_k) - f(x^*) \leq \sigma_k(\phi_0(x^*) - f(x^*)) \longrightarrow 0 \ .$$

Proof:

$$\begin{aligned} f(x_k) \quad \leq \quad & \min_{x \in \Omega} \phi_k(x) \\ & \text{using the definition of estimate sequence} \\ & \leq \min_{x \in \Omega} [(1-\sigma_k) f(x) + \sigma_k \phi_0(x)] \\ & x^* \text{ is solution, so } \forall x \in \Omega : f(x) \geq f(x^*) \\ & \leq (1-\sigma_k) f(x^*) + \sigma_k \phi_0(x^*) \\ & = f(x^*) + \sigma_k(\phi_0(x^*) - f(x^*)) \end{aligned}$$

Moreover, the $(\phi_0(x^*) - f(x^*))$ is independent of k, i.e. constant, so if $\sigma_k \longrightarrow 0$, then

$$\sigma_k(\phi_0(x^*) - f(x^*)) \quad \longrightarrow \quad 0 \ .$$

How to choose estimate sequence

Theorem

Let

- $\phi_0(x) \in \mathcal{F}$ *is arbitrary*
- $\{y_k\}_{k=0}^{\infty}$ *is arbitrary sequence in* Ω
- $\{\alpha_k\}_{k=0}^{\infty}: \quad \alpha_k \in (0,1), \sum_{k=0}^{\infty} \alpha_k = \infty$
- $\sigma_0 = 1$

Then the pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}, \{\sigma_k\}_{k=0}^{\infty}$ *defined by*

$$\begin{aligned} \sigma_{k+1} \quad &:= \quad (1-\alpha_k)\sigma_k \\ \phi_{k+1}(x) \quad &:= \quad (1-\alpha_k)\phi_k(x) + \alpha_k[f(y_k^P) + \langle g^P(y_k), x - y_k \rangle \\ & \qquad + \tfrac{1}{2L}\|g^P(y_k)\|^2 + \tfrac{\mu}{2}\|x - y_k\|^2] \end{aligned}$$

is estimate sequence.

Proof:
based on previous theorem $f(x) \geq f(y^P) + \langle g^P(y), x - y \rangle + \frac{1}{2\gamma}\|g^P(y)\|^2 + \frac{\mu}{2}\|x - y\|^2$.

Some simple simplifications

$$\begin{aligned}\sigma_{k+1} &:= (1-\alpha_k)\sigma_k \\ \phi_{k+1}(x) &:= (1-\alpha_k)\phi_k(x) + \alpha_k(f(y_k^P) + \langle g^P(y_k), x - y_k\rangle \\ &\quad + \tfrac{1}{2L}\|g^P(y_k)\|^2 + \tfrac{\mu}{2}\|x-y_k\|^2)\end{aligned}$$

is eqivalent to

Equivalent algorithm

Let $\phi_0^{\min} \in \mathbb{R}$, $v_0 \in \mathbb{R}^n$.

$$\phi_k(x) := \phi_k^{\min} + \frac{\gamma_k}{2}\|x - v_k\|^2 \ ,$$

where

$$\begin{aligned}\gamma_{k+1} &:= (1-\alpha_k)\gamma_k + \alpha_k\mu \ , \\ v_{k+1} &:= \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k g^P(y_k)\right] \quad = \arg\min_x \phi_{k+1}(x) \ , \\ \phi_{k+1}^{\min} &:= (1-\alpha_k)\phi_k^{\min} + \alpha_k f(y_k^P) + \left(\frac{\alpha_k}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|g^P(y_k)\|^2 \\ &\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle g^P(y_k), v_k - y_k\rangle\right) \ .\end{aligned}$$

Final simplifications

From previous, we can obtain (after substitutions)

$$\phi_{k+1}^{\min} \geq f(y_k^P) + (1-\alpha_k)\langle g^P(y_k), \frac{\gamma_k \alpha_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\rangle + \left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right) \|g^P(y_k)\|^2 .$$

Final simplifications

From previous, we can obtain (after substitutions)

$$\phi_{k+1}^{\min} \geq f(y_k^P) + (1-\alpha_k)\langle g^P(y_k), \frac{\gamma_k \alpha_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\rangle + \left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|g^P(y_k)\|^2 .$$

Idea

If we choose γ_{k+1} and α_k in a such way that

$$\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}} = 0 \ ,$$

then we get good estimation.

This way is

$$\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}} = 0 \quad \Leftrightarrow \quad \gamma_{k+1} = L\alpha_k^2$$

From previous formulations, we get

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k \mu \qquad =: \gamma_{k+1} \ .$$

Final simplifications

From previous, we can obtain (after substitutions)

$$\phi_{k+1}^{\min} \geq f(y_k^P) + (1-\alpha_k)\langle g^P(y_k), \frac{\gamma_k \alpha_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \rangle + \left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}} \right) \|g^P(y_k)\|^2 .$$

Idea 2

If we choose y_k in a such way that

$$\frac{\gamma_k \alpha_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \; = \; 0 \, ,$$

then we get good estimation.

This way is (after simlifications)

$$y_k : \; = \frac{\gamma_k \alpha_k v_k + \gamma_{k+1} x_k}{\gamma_k + \alpha_k \mu} \; .$$

The first algorithm

- $x_0 \in \Omega$
- $\gamma_0 > 0$
- $v_0 = x_0$
- $k := 0$
- WHILE ...
  - $\alpha_k \in (0,1)$ solves $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k \mu$
  - $\gamma_{k+1} := (1-\alpha_k)\gamma_k + \alpha_k \mu$
  - $y_k := \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + \alpha_k \mu}$
  - $x_{k+1} = y_k^P$
  - $v_{k+1} : = \frac{1}{\gamma_{k+1}} \left[(1-\alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k g^P(y_k) \right]$
  - $k := k+1$

More simplification

After some more simplifications we obtain

- $x_0 \in \Omega$
- $\Theta_0 = 1$
- $y_0 = x_0$
- $k := 0$
- WHILE ...
  - $x_{k+1} := P_\Omega(y_k - \frac{1}{L}\nabla f(y_k))$
  - Θ_{k+1} solves $\Theta_{k+1}^2 = (1 + \Theta_{k+1})\Theta_k + \Theta_k \frac{\mu}{L}$
  - $\beta_k := \frac{\Theta_k(1-\Theta_k)}{\Theta_k^2 + \Theta_{k+1}}$
  - $y_{k+1} := x_{k+1} + \beta_{k+1}(x_{k+1} - x_k)$
  - $k := k + 1$

APGD for QP

Equality in QP

If $f(x) :\, = \frac{1}{2} x^T A x - b^T x$ with SPSD $A \in \mathbb{R}^{n,n}$, then

$$\forall x, y \in \mathbb{R}^n : f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \|x - y\|_A^2$$

Variant of the first estimation for QP

$\gamma \in \left(0, \frac{1}{\lambda_{\max}^A}\right\rangle$, $y \in \mathbb{R}^n$, then

$$\forall x \in \Omega : \quad f(x) \geq f(y^P) + \langle g^P(y), x - y \rangle + \frac{\gamma}{2} \|g^P(y)\|^2 + \frac{1}{2} \|x - y\|_A^2 \ ,$$

where

$$\begin{aligned} g^P(x) \quad &: = \quad \frac{1}{\gamma}(x - x^P) \ , \\ x^P \quad &: = \quad P_\Omega(x - \gamma \nabla f(x)) \ . \end{aligned}$$

Choice of estimate functions

Theorem

Let

- $\phi_0(x) \in \mathcal{F}$ *is arbitrary*
- $\{y_k\}_{k=0}^{\infty}$ *is arbitrary sequence in* Ω
- $\{\alpha_k\}_{k=0}^{\infty}: \quad \alpha_k \in (0,1), \sum_{k=0}^{\infty} \alpha_k = \infty$
- $\sigma_0 = 1$

Then the pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}, \{\sigma_k\}_{k=0}^{\infty}$ *defined by*

$$\begin{aligned} \sigma_{k+1} \quad &:= \quad (1-\alpha_k)\sigma_k \\ \phi_{k+1}(x) \quad &:= \quad (1-\alpha_k)\phi_k(x) + \alpha_k[f(y_k^P) + \langle g^P(y_k), x - y_k\rangle \\ &\qquad + \tfrac{\gamma}{2}\|g^P(y_k)\|^2 + \tfrac{1}{2}\|x - y_k\|_A^2] \end{aligned}$$

is estimate sequence.

Attempt to the same some simple simplifications

$$\begin{array}{rcl} \sigma_{k+1} & := & (1-\alpha_k)\sigma_k \\ \phi_{k+1}(x) & := & (1-\alpha_k)\phi_k(x) + \alpha_k(f(y_k^P) + \langle g^P(y_k), x - y_k\rangle \\ & & +\frac{\gamma}{2}\|g^P(y_k)\|^2 + \frac{1}{2}\|x - y_k\|_A^2) \end{array}$$

is eqivalent to

Equivalent algorithm

Let $\phi_0^{\min} \in \mathbb{R}$, $v_0 \in \mathbb{R}^n$.

$$\phi_k(x) := \phi_k^{\min} + \frac{\gamma_k}{2}\|x - v_k\|_A^2 \ ,$$

where

$$\begin{array}{rcl} \gamma_{k+1} & := & (1-\alpha_k)\gamma_k + \alpha_k \ , \\ v_{k+1} & := & \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k A^{-1} g^P(y_k)\right] \quad = \arg\min_x \phi_{k+1}(x) \ , \\ \phi_{k+1}^{\min} & := & (1-\alpha_k)\phi_k^{\min} + \alpha_k f(y_k^P) + \frac{\alpha_k}{2L}\|g^P(y_k)\|^2 \\ & & +\frac{\alpha_k(1-\alpha_k)\gamma_k}{2\gamma_{k+1}}\left(\frac{1}{2}\|y_k - v_k\|_A^2 + \langle g^P(y_k), v_k - y_k\rangle\right) \\ & & -\frac{\alpha_k^2}{2\gamma_{k+1}}\langle A^{-1} g^P(y_k), g^P(y_k)\rangle \ . \end{array}$$

And the same idea as before?

From previous, we can obtain (after substitutions)

$$\phi_{k+1}^{\min} \geq f(y_k^P) + (1-\alpha_k)\langle g^P(y_k), \frac{\gamma_k \alpha_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\rangle$$
$$+\left\langle \left(\frac{L}{2} I - \frac{\alpha_k^2}{2\gamma_{k+1}} A^{-1}\right) g^P(y_k), g^P(y_k) \right\rangle .$$

?

Conclusion

- APGD is robust algorithm for solving optimising problems
- but for QP are the estimates very rough and can be improved (the speed of algorithm depends on this estimation sequence)

- but here should came the idea, which does not come yet

Conclusion

- APGD is robust algorithm for solving optimising problems
- but for QP are the estimates very rough and can be improved (the speed of algorithm depends on this estimation sequence)
- but here should came the idea, which does not come yet

. . . and the best beer is really in Czech Republic, sorry guys

Thank you for your attention!