Primal-dual Subgradient Method for Convex Problems with Functional Constraints

Yurii Nesterov, CORE/INMA (UCL)

Workshop on embedded optimization
EMBOPT2014
September 9, 2014 (Lucca)

Outline

Optimization problem: simple constraints

Consider the problem: $\min\limits_{x \in Q} f(x)$, where

- Q is a closed convex set: $x, y \in Q \Rightarrow [x, y] \subseteq Q$,
- f is a subdifferentiable on Q convex function:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad x, y \in Q, \ \nabla f(x) \in \partial f(x).$$

Optimality condition: point $x_* \in Q$ is optimal iff

$$\langle \nabla f(x_*), x - x_* \rangle \geq 0, \quad \forall x \in Q.$$

Interpretation: Function increases along any feasible direction.

Examples

1. Interior solution. Let $x_* \in \operatorname{int} Q$. Then
$\langle \nabla f(x_*), x - x_* \rangle \geq 0$, $\forall x \in Q$ implies $\nabla f(x_*) = 0$.

2. Optimization over positive orthant.
Let $Q \equiv \mathbb{R}^n_+ = \left\{ x \in \mathbb{R}^n : \; x^{(i)} \geq 0, \; i = 1 \ldots, n \right\}$.

Optimality condition: $\langle \nabla f(x_*), x - x_* \rangle \geq 0, \quad \forall x \in \mathbb{R}^n_+$.

Coordinate form: $\nabla_i f(x_*) \left(x^{(i)} - x_*^{(i)} \right) \geq 0, \quad \forall x^{(i)} \geq 0$.

This means that

- $\nabla_i f(x_*) \geq 0, \quad i = 1, \ldots, n,$ (tend $x^{(i)} \to \infty$)
- $x_*^{(i)} \nabla_i f(x_*) = 0, \quad i = 1, \ldots, n,$ (set $x^{(i)} = 0$.)

Optimization problem: functional constraints

Problem: $\min\limits_{x \in Q} \{ f_0(x),\ f_i(x) \le 0,\ i = 1, \dots, m \}$, where

- Q is a closed convex set,
- all f_i are convex and subdifferentiable on Q, $i = 0, \dots, m$:

$$f_i(y) \ge f_i(x) + \langle \nabla f_i(x), y - x \rangle, \quad x, y \in Q,\ \nabla f_i(x) \in \partial f_i(x).$$

Optimality condition (KKT, 1951): point $x_* \in Q$ is optimal iff there exist *Lagrange multipliers* $\lambda_*^{(i)} \ge 0$, $i = 1, \dots, m$, such that

$$(1): \quad \langle \nabla f_0(x_*) + \sum_{i=1}^{m} \lambda_*^{(i)} \nabla f_i(x_*), x - x_* \rangle \ge 0, \quad \forall x \in Q,$$

$$(2): \quad f_i(x_*) \le 0, \quad i = 1, \dots, m, \quad \text{(feasibility)}$$

$$(3): \quad \lambda_*^{(i)} f_i(x_*) = 0, \quad i = 1, \dots, m. \quad \text{(complementary slackness)}$$

Lagrange multipliers: interpretation

Let $\mathcal{I} \subseteq \{1, \ldots, m\}$ be an arbitrary set of indexes.

Denote $f_{\mathcal{I}}(x) = f_0(x) + \sum\limits_{i \in \mathcal{I}} \lambda_*^{(i)} f_i(x)$. Consider the problem

$$\mathcal{P}_{\mathcal{I}}: \qquad \min_{x \in Q} \{ f_{\mathcal{I}}(x) : \; f_i(x) \leq 0, \; i \notin \mathcal{I} \}.$$

Observation: in any case, x_* is the optimal solution of problem $\mathcal{P}_{\mathcal{I}}$.

Interpretation: $\lambda_*^{(i)}$ are the *shadow prices* for resources. **(Kantorovich, 1939)**

Application examples:

- Traffic congestion: car flows on roads $\Leftrightarrow$ size of queues.
- Electrical networks: currents in the wires $\Leftrightarrow$ voltage potentials, etc.

Main question: How to compute (x_*, λ_*)?

Algebraic interpretation

Consider the Lagrangian $\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda^{(i)} f_i(x)$.

Condition KKT(1): $\langle \nabla f_0(x_*) + \sum_{i=1}^{m} \lambda_*^{(i)} \nabla f_i(x_*), x - x_* \rangle \geq 0$,

$\forall x \in Q$, implies

$$\boxed{x_* \in \operatorname{Arg}\min_{x \in Q} \mathcal{L}(x, \lambda_*).}$$

Define the dual function $\phi(\lambda) = \min_{x \in Q} \mathcal{L}(x, \lambda)$, $\lambda \geq 0$. It is concave!

By Danskin's Theorem, $\nabla \phi(\lambda) = (f_1(x(\lambda)), \ldots, f_m(x(\lambda)))$, with

$$x(\lambda) \in \operatorname{Arg}\min_{x \in Q} \mathcal{L}(x, \lambda).$$

Conditions KKT(2,3): $f_i(x_*) \leq 0$, $\lambda_*^{(i)} f_i(x_*) = 0$, $i = 1, \ldots, m$, imply $(x_* = x(\lambda_*))$

$$\boxed{\lambda_* \in \operatorname{Arg}\max_{\lambda \geq 0} \phi(\lambda).}$$

Algorithmic aspects

Main idea: solve the dual problem $\boxed{\max_{\lambda \geq 0} \phi(\lambda)}$

by the *subgradient method*:

1. Compute $x(\lambda_k)$ and define $\nabla \phi(\lambda_k) = (f_1(x(\lambda_k)), \ldots, f_m(x(\lambda_k)))$.

2. Update $\lambda_{k+1} = \text{Project}_{\mathbb{R}^n_+}(\lambda_k + h_k \nabla \phi(\lambda_k))$.

Stepsizes $h_k > 0$ are defined in the usual way.

Main difficulties:

- Each iteration is time consuming.
- Unclear termination criterion.
- Low rate of convergence $\left(O\left(\frac{1}{\epsilon^2}\right)\right.$ upper-level iterations).

Augmented Lagrangian (1970's) [Hestenes, Powell, Rockafellar, Polyak, Bertsekas, ...]

Define the Augmented Lagrangian

$$\widehat{\mathcal{L}}_K(x,\lambda) = f_0(x) + \frac{1}{2K}\sum_{i=1}^{m}\left(\lambda^{(i)} + Kf_i(x)\right)_+^2 - \frac{1}{2K}\|\lambda\|_2^2, \quad \lambda \in \mathbb{R}^m,$$

where $K > 0$ is a penalty parameter.

Consider the dual function $\hat{\phi}(\lambda) = \min\limits_{x\in Q} \widehat{\mathcal{L}}(x,\lambda)$.

- **Main properties.** Function $\hat{\phi}$ is concave. Its gradient is Lipschitz continuous with constant $\frac{1}{K}$.
- Its unconstrained maximum is attained at the optimal dual solution.
- The corresponding point $\hat{x}(\lambda_*)$ is the optimal primal solution.

Hint: Check that the equation $\left(\lambda^{(i)} + Kf_i(x)\right)_+ = \lambda^{(i)}$ is equivalent to KKT(2,3).

Method of Augmented Lagrangians

Note that $\nabla\hat{\phi}(\lambda) = \frac{1}{K}\left(\lambda^{(i)} + Kf_i(x)\right)_+ - \frac{1}{K}\lambda$.

Therefore, the usual gradient method $\lambda_{k+1} = \lambda_k + K\nabla\hat{\phi}(\lambda_k)$ is exactly as follows:

Method: $\lambda_{k+1} = \left(\lambda_k + Kf(\hat{x}(\lambda_k))\right)_+$.

Advantage: Fast convergence of the dual process.

Disadvantages:

- Difficult iteration.
- Unclear termination.
- No global complexity analysis.

Do we have an alternative?

Problem formulation

Problem: $f^* = \inf\limits_{x \in Q} \{f_0(x) : \; f_i(x) \leq 0, \; i = 1, \ldots, m\}$, where

- $f_i(x)$, $i = 0, \ldots, m$, are closed convex functions on Q endowed with a first-order black-box oracles,
- $Q \subset \mathbb{E}$ is a bounded *simple* closed convex set. (We can solve some auxiliary optimization problems over Q.)

Defining the Lagrangian

$$\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda^{(i)} f_i(x), \quad x \in Q, \; \lambda \in \mathbb{R}^m_+,$$

we can introduce the Lagrangian dual problem $\boxed{f_* \stackrel{\text{def}}{=} \sup\limits_{\lambda \in \mathbb{R}^m_+} \phi(\lambda),}$

where $\phi(\lambda) \stackrel{\text{def}}{=} \inf\limits_{x \in Q} \mathcal{L}(x, \lambda)$.

Clearly, $f^* \geq f_*$. Later, we will show $f^* = f_*$ *algorithmically*.

Bregman distances

Prox-function: $d(\cdot)$ is strongly convex on Q with parameter one:

$$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \tfrac{1}{2}\|y - x\|^2, \quad x, y \in Q.$$

Denote by x_0 the prox-center of the set Q: $x_0 = \arg\min_{x \in Q} d(x)$.

Assume $d(x_0) = 0$.

Bregman distance:

$$\boxed{\beta(x, y) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle, \; x, y \in Q.}$$

Clearly, $\beta(x, y) \geq \frac{1}{2}\|x - y\|^2$ for all $x, y \in Q$.

Bregman mapping: for $x \in Q$, $g \in E^*$ and $h > 0$ define

$$\mathcal{B}_h(x, g) = \arg\min_{y \in Q} \{h\langle g, y - x \rangle + \beta(x, y)\}.$$

The first-order condition for point $x_+ \stackrel{\text{def}}{=} \mathcal{B}_h(x, g)$ is as follows:

$$\langle hg + \nabla d(x_+) - \nabla d(x), y - x_+ \rangle \geq 0, \quad y \in Q.$$

Examples

1. Euclidean distance. We choose $\|x\| = \left[\sum_{i=1}^{n} (x^{(i)})^2\right]^{1/2}$ and $d(x) = \frac{1}{2}\|x\|^2$. Then $\beta(x, y) = \frac{1}{2}\|x - y\|^2$, and we have

$$\mathcal{B}_h(x, g) = \text{Projection}_Q(x - hg).$$

2. Entropy distance. We choose $\|x\| = \sum_{i=1}^{n} |x^{(i)}|$ and $d(x) = \ln n + \sum_{i=1}^{n} x^{(i)} \ln x^{(i)}$. Then

$$\beta(x, y) = \sum_{i=1}^{n} y^{(i)} [\ln y^{(i)} - \ln x^{(i)}].$$

If $Q = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^{n} x^{(i)} = 1\}$, then

$$\mathcal{B}_h^{(i)}(x, g) = x^{(i)} e^{-hg^{(i)}} / \left[\sum_{j=1}^{n} x^{(j)} e^{-hg^{(j)}}\right], \; i = 1, \ldots, n.$$

Switching subgradient method

Input parameter: the step size $h > 0$.

Initialization : Compute the prox-center x_0.

Iteration $k \geq 0$**:** a) Define $\mathcal{I}_k = \{i \in \{1, \ldots, m\} : \; f_i(x_k) > h\|\nabla f_i(x_k)\|_*\}$.

b) If $\mathcal{I}_k = \emptyset$, then compute $x_{k+1} = \mathcal{B}_h\left(x_k, \frac{\nabla f_0(x_k)}{\|\nabla f_0(x_k)\|_*}\right)$.

c) If $\mathcal{I}_k \neq \emptyset$, then choose arbitrary $i_k \in \mathcal{I}_k$ and define

$h_k = \frac{f_{i_k}(x_k)}{\|\nabla f_{i_k}(x_k)\|_*^2}$. Compute $x_{k+1} = \mathcal{B}_{h_k}(x_k, \nabla f_{i_k}(x_k))$.

After $t \geq 0$ iterations, define $\mathcal{F}_t = \{k \in \{0, \ldots, t\} : \; \mathcal{I}_k = \emptyset\}$.

Denote $N(t) = |\mathcal{F}(t)|$. It is possible that $N(t) = 0$.

Finding the dual multipliers

if $N(t) > 0$, define the dual multipliers as follows:

- $\lambda_t^{(0)} = h \sum\limits_{k \in \mathcal{F}_t} \frac{1}{\|\nabla f_0(x_k)\|_*}$,
- $\lambda_t^{(i)} = \frac{1}{\lambda_t^{(0)}} \sum\limits_{k \in \mathcal{A}_i(t)} h_k, \quad i = 1, \ldots, m,$

  where $\mathcal{A}_i(t) = \{k \in \{0, \ldots, t\} : \; i_k = i\}$, $0 \leq i \leq m$.

Denote $S_t = \sum\limits_{k \in \mathcal{F}_t} \frac{1}{\|\nabla f_0(x_k)\|_*}$. If $\mathcal{F}_t = \emptyset$, then we define $S_t = 0$.

For proving convergence of the switching strategy, we find an upper bound for the gap

$$\delta_t = \frac{1}{S_t} \sum_{k \in \mathcal{F}(t)} \frac{f_0(x_k)}{\|\nabla f_0(x_k)\|_*} - \phi(\lambda_t),$$

assuming that $N(t) > 0$.

Convergence analysis

Note that $\lambda_t^{(0)} = h \cdot S(t)$. Therefore

$$\lambda_t^{(0)} \cdot \delta_t = \sup_{x \in Q} \left\{ h \sum_{k \in \mathcal{F}(t)} \frac{f_0(x_k)}{\|\nabla f_0(x_k)\|_*} - \lambda_t^{(0)} f_0(x) - \sum_{i=1}^{m} \sum_{k \in \mathcal{A}_i(t)} h_k f_i(x) \right\}$$

$$= \sup_{x \in Q} \left\{ h \sum_{k \in \mathcal{F}(t)} \frac{f_0(x_k) - f_0(x)}{\|\nabla f_0(k)\|_*} - \sum_{k \notin \mathcal{F}(t)} h_k f_{i_k}(x) \right\}$$

$$\leq \sup_{x \in Q} \left\{ h \sum_{k \in \mathcal{F}(t)} \frac{\langle \nabla f_0(x_k), x_k - x \rangle}{\|\nabla f_0(x_k)\|_*} + \sum_{k \notin \mathcal{F}(t)} h_k [\langle \nabla f_{i_k}(x_k), x_k - x \rangle - f_{i_k}(x_k)] \right\}$$

Let us estimate from above the right-hand side of this inequality.

Feasible step

For arbitrary $x \in Q$, denote $r_t(x) = \beta(x_t, x)$. Then

$$\begin{aligned} r_{t+1}(x) - r_t(x) &= [d(x) - d(x_{t+1}) - \langle \nabla d(x_{t+1}), x - x_{t+1} \rangle] \\ &\quad -[d(x) - d(x_t) - \langle \nabla d(x_t), x - x_t \rangle] \\ &= \langle \nabla d(x_t) - \nabla d(x_{t+1}), x - x_{t+1} \rangle \\ &\quad -[d(x_{t+1}) - d(x_t) - \langle \nabla d(x_t), x_{t+1} - x_t \rangle] \\ &\leq \langle \nabla d(x_t) - \nabla d(x_{t+1}), x - x_{t+1} \rangle - \tfrac{1}{2}\|x_t - x_{t+1}\|^2. \end{aligned}$$

In view of optimality condition, for all $x \in Q$ and $k \in \mathcal{F}(t)$ we have

$$\frac{h}{\|\nabla f_0(x_k)\|_*} \langle \nabla f_0(x_k), x_{k+1} - x \rangle \leq \langle \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle.$$

Assume that $k \in \mathcal{F}_t$. In this case,

$$\begin{aligned} r_{k+1}(x) - r_k(x) &\leq -\frac{h}{\|\nabla f_0(x_k)\|_*} \langle \nabla f_0(x_k), x_{k+1} - x \rangle - \tfrac{1}{2}\|x_k - x_{k+1}\|^2 \\ &\leq -\frac{h}{\|\nabla f_0(x_k)\|_*} \langle \nabla f_0(x_k), x_k - x \rangle + \tfrac{1}{2}h^2. \end{aligned}$$

Infeasible step

If $k \notin \mathcal{F}(t)$, then the optimality condition defining the point x_{k+1} looks as follows:

$$h_k \langle \nabla f_{i_k}(x_k), x_{k+1} - x \rangle \quad \leq \quad \langle \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle.$$

Therefore,

$$\begin{aligned} r_{k+1}(x) - r_k(x) \quad &\leq \quad -h_k \langle \nabla f_{i_k}(x_k), x_{k+1} - x \rangle - \tfrac{1}{2} \|x_k - x_{k+1}\|^2 \\ &\leq \quad -h_k \langle \nabla f_{i_k}(x_k), x_k - x \rangle + \tfrac{1}{2} h_k^2 \|\nabla f_{i_k}(x_k)\|_*^2. \end{aligned}$$

Hence,

$$\begin{aligned} h_k [\langle \nabla f_{i_k}(x_k), x_k - x \rangle - f_{i_k}(x_k)] \quad &\leq \quad r_k(x) - r_{k+1}(x) - \frac{f_{i_k}^2(x_k)}{2\|\nabla f_{i_k}(x_k)\|_*^2} \\ &\leq \quad r_k(x) - r_{k+1}(x) - \tfrac{1}{2} h^2. \end{aligned}$$

Convergence result

Summing up all inequalities for $k = 0, \ldots, t$, and taking into account that $r_{t+1}(x) \geq 0$, we obtain

$$\lambda_t^{(0)} \delta_t \leq r_0(x) + \tfrac{1}{2} N(t) h^2 - \tfrac{1}{2}(t - N(t)) h^2 = r_0(x) - \tfrac{1}{2} t h^2 + N(t) h^2.$$

Denote $D = \max\limits_{x \in Q} r_0(x)$.

Theorem. If the number $t \geq \frac{2}{h^2} D$, then $\mathcal{F}(t) \neq \emptyset$.

In this case $\delta_t \leq Mh$ and $\max\limits_{1 \leq i \leq m} f_i(x_k) \leq Mh$, $k \in \mathcal{F}(t)$

where $M = \max\limits_{0 \leq k \leq t} \max\limits_{0 \leq i \leq m} \|\nabla f_i(x_k)\|_*$.

Proof: If $\mathcal{F}(t) = \emptyset$, then $N(t) = 0$. Consequently, $\lambda_t^{(0)} = 0$. This is impossible for t big enough.

Finally, $\lambda_t^{(0)} \geq \frac{h}{M} N(t)$. Therefore, if t is big enough, then

$$\delta_t \leq \frac{N(t) h^2}{\lambda_t^{(0)}} \leq Mh. \quad \square$$

Conclusion

1. Optimal primal-dual solution can be approximated by a simple switching subgradient scheme.

2. Dual process looks as a coordinate-descent method.

3. Approximations of dual multipliers have natural interpretation : relative importance of corresponding constraints during the adjustments process.

4. However, it has optimal worst-case efficiency estimate even if the dual optimal solution does not exist.

5. Many interesting questions (influence of smoothness, strong convexity, etc.)

THANK YOU FOR YOUR ATTENTION!