Convex Optimization Algorithms
An introduction

Wenxun Xing

Department of Mathematical Sciences
Tsinghua University
wxing@tsinghua.edu.cn
Tel: 62787945
Office hour: 4:00-5:00pm, Thursday
Office: The New Science Building, A416

Part I Subgradient Methods

Subgradient methods

- A convex optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathcal{X} \subseteq \mathbb{R}^n, \end{array}$$

  where $f(x)$ is a given convex function and $\mathcal{X}$ is a convex closed set.
- Iteration method.

$$x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$$

  where g_k is a subgradient of $f(x)$ at x_k, α_k is a step size, and $P_{\mathcal{X}}(z)$ denotes Euclidean projection of z on the set $\mathcal{X}$.

Examples

Example

- *Optimization problem:* $\min_{x\geq 0} f(x) = (x-1)^2$.
- *Gradient:* $g(x) = f'(x) = 2(x-1)$.
- *Iteration and projection*

$$x_{k+1} = \max\{x_k - \alpha_k f'(x_k), 0\}, P_{\{x\geq 0\}}(z) = \max\{z, 0\}.$$

- *Iterations with different step sizes*

$$\alpha_k = \frac{1}{2},\ x_0 = 0, x_1 = \max\{1, 0\} = 1, x_k = 1,\ k \geq 2.$$

$$\alpha_k = \frac{1}{3},\ x_0 = 0, x_1 = \frac{2}{3}, x_2 = \frac{8}{9}, x_3 = \frac{26}{27}.$$

$$\alpha_k = 2,\ x_0 = 0, x_1 = 4, x_2 = \max\{-8, 0\} = 0, x_3 = 4.$$

Example

- *Optimization problem:*
$$\min_{x\in\mathbb{R}} f(x) = \begin{cases} x, & x \geq 1 \\ \frac{1}{2}(x+1), & -1 \leq x \leq 1 \\ 0, & x \leq -1 \end{cases}.$$
- *Subgradient:* $\partial f(x) = \begin{cases} 1, & x > 1 \\ \left[\frac{1}{2}, 1\right], & x = 1 \\ \frac{1}{2}, & -1 < x < 1, \\ \left[0, \frac{1}{2}\right] & x = -1 \\ 0, & x < -1. \end{cases}$
- *Iteration and projection*
$$x_{k+1} = x_k - \alpha_k \partial f(x_k), P_{\{x\in\mathbb{R}\}}(z) = z.$$
- *Iterations with different subgradients.*
$$\alpha_k = \frac{1}{3},\ x_0 = 2, x_1 = 2 - \frac{1}{3} = \frac{5}{3}, x_2 = \frac{5}{3} - \frac{1}{3} = \frac{4}{3}, x_3 = 1,$$
$$x_4 = \begin{cases} 1 - \frac{1}{3} = \frac{2}{3}, & \partial f(x_3) = 1 \\ 1 - \frac{1}{3} \times \frac{1}{2} = \frac{5}{6}, & \partial f(x_3) = \frac{1}{2}. \end{cases}$$

Subgradient

Definition

For a given $f(x)$ over $\mathcal{X}$, $g \in \mathbb{R}^n$ is called a subgradient of $f(x)$ at x_0 if

$$f(x) \geq f(x_0) + g^T(x - x_0), \ \forall x \in \mathcal{X}.$$

The set of subgradients at x_0 is denoted as $\partial f(x_0)$.

Definition

Define a directional derivative of $f(x)$ at x in a direction d as

$$f'(x; d) = \lim_{\delta \to 0^+} \frac{f(x + \delta d) - f(x)}{\delta},$$

if it exists.

Subgradient

Theorem

Let $f : \mathbb{R}^n\ (\mathcal{X}) \to \mathbb{R}$ *be a convex function (and* $\mathcal{X}$ *be a convex set). For every* $x \in \mathbb{R}^n\ (\in \mathrm{ri}(\mathcal{X}))$*, the followings hold: (i) The* $\partial f(x)$ *is a nonempty, convex, and compact set, and we have*

$$f'(x; d) = \max_{g \in \partial f(x)} g^T d.$$

(ii) If $f(x)$ *is differentiable at* x *with gradient* $\nabla f(x)$*, then we have* $\nabla f(x) = \partial f(x)$*.*

The optimality condition

Theorem

An $x \in \mathcal{X}$ *minimizes a convex function* $f : \mathbb{R}^n \to \mathbb{R}$ *over a convex set* $\mathcal{X}$ *if and only if there exists a subgradient* $g \in \partial f(x)$ *such that*

$$g^T(z - x) \geq 0, \quad \forall z \in \mathcal{X}.$$

Main convergence results

Theorem

(Non-expansiveness of the projection) Let $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$*,* $\mathcal{X}$ *be a nonempty closed convex set. We have*

$$\|P_{\mathcal{X}}(x) - P_{\mathcal{X}}(y)\| \leq \|x - y\|, \forall x, y \in \mathbb{R}^n.$$

Theorem

Let $\{x_k\}$ *be the sequence generated by* $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$ *and* $f(x)$ *be a convex function over* $\mathcal{X}$*. Then for all* $y \in \mathcal{X}$ *and* $k \geq 0$*:*
(i) We have

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k[f(x_k) - f(y)] + \alpha_k^2\|g_k\|^2.$$

(ii) If $f(y) < f(x_k)$*, we have* $\|x_{k+1} - y\| < \|x_k - y\|$ *for all step-sizes* α_k *such that*

$$0 < \alpha_k < \frac{2[f(x_k) - f(y)]}{\|g_k\|^2}.$$

Discussion on convergence

(ii) If $f(x_k) > f(y)$, we have $\|x_{k+1} - y\| < \|x_k - y\|$ for all step-sizes α_k such that

$$0 < \alpha_k < \frac{2[f(x_k) - f(y)]}{\|g_k\|^2}.$$

- If x_k is not an optimal solution and x^* is an optimal solution, then we can get a point x_{k+1} much closer to the optimal x^* by selecting small enough α_k.
- Suppose $\sup_{g \in \cup_{x \in \mathcal{X}} \partial f(x)} \|g\| \leq c$ and $\alpha_k = \alpha > 0,\ \forall k = 0, 1, \ldots$. If

$$x_k \notin \left\{ x \in \mathcal{X} | f(x) \leq f(x^*) + \frac{\alpha c^2}{2} \right\},$$

  then $\|x_{k+1} - x^*\| < \|x_k - x^*\|$.
- For a given precision $\epsilon > 0$ and $\alpha = \frac{2\epsilon}{c^2}$, then we get an ϵ-optimization solution in

$$\{x \in \mathcal{X} | f(x) \leq f(x^*) + \epsilon\}.$$

Convergence analysis—constant step-size

Assumption A: $\sup_{g \in \cup_{x \in \mathcal{X}} \partial f(x)} \|g\| \leq c$ and $\alpha_k = \alpha > 0, \ \forall k = 0, 1, \ldots$.
Denote $\mathcal{X}^*$ be the set of optimal solutions,

$$f^* = \min_{x \in \mathcal{X}} f(x), \ f_\infty = \lim_{k \to \infty} \inf f(x_k), \ d(x) = \min_{x^* \in \mathcal{X}^*} \|x - x^*\|.$$

Theorem

(Convergence within a neighborhood) (i) If $f^* = -\infty$, then $f_\infty = f^*$.
(ii) If $f^* > -\infty$, then

$$f_\infty \leq f^* + \frac{\alpha c^2}{2}.$$

Note: No guarantee for the convergence of $\{x_k\}$.

Theorem

(Convergence rate) Suppose $\sup_{g \in \cup_{x \in \mathcal{X}} \partial f(x)} \|g\| \leq c$, $\alpha_k = \alpha$ *and* $\mathcal{X}^* \neq \emptyset$. *For any positive* $\epsilon > 0$, *we have*

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \frac{\alpha c^2 + \epsilon}{2},$$

where $K = \lfloor \frac{(d(x_0))^2}{\alpha \epsilon} \rfloor$.

Theorem

(Linear convergence rate) Suppose $\sup_{g \in \cup_{x \in \mathcal{X}} \partial f(x)} \|g\| \leq c$, $\alpha_k = \alpha$, $\mathcal{X}^* \neq \emptyset$ *and*

$$f(x) - f^* \geq \gamma (d(x))^2, \quad \forall x \in \mathcal{X}, \text{strong convexity}$$

for some $\gamma > 0$ *and* $\alpha \leq \frac{1}{2\gamma}$. *Then for all* k

$$(d(x_{k+1}))^2 \leq (1 - 2\alpha\gamma)^{k+1} (d(x_0))^2 + \frac{\alpha c^2}{2\gamma}.$$

Note: We have to check the strong convexity. For example, f *is polyhedral, and* $\mathcal{X}$ *is polyhedral and compact.*

Convergence analysis—nonconstant step-size

Theorem

Under the boundedness assumption of subgradients, if α_k satisfies

$$\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty,$$

then $f_\infty = f^*$. Moreover if $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ and $\mathcal{X}^*$ is nonempty, then $\{x_k\}$ converges to some optimal solution.

Theorem

Under the boundedness assumption of subgradients, if $\mathcal{X}^*$ is nonempty and

$$\alpha_k = \frac{f(x_k) - f^*}{\|g_k\|^2}, \quad (\text{Not easy to get } f^*!)$$

then $\{x_k\}$ converges to some optimal solution.

Theorem

Under the boundedness assumption of subgradients and suppose

$$\alpha_k = \frac{f(x_k) - f_k}{\|g_k\|^2},$$

$$f_k = \min_{0 \le j \le k} f(x_j) - \delta_k,$$

$$\delta_{k+1} = \begin{cases} \theta\delta_k, & \text{if } f(x_{k+1}) \le f(x_k), \\ \max\{\beta\delta_k, \delta\}, & \text{if } f(x_{k+1}) > f(x_k), \end{cases}$$

where $0 < \beta < 1, \theta \ge 1$. *If* $f^* = -\infty$, *then*

$$\inf_{j \ge 0} f(x_j) = f^*,$$

and if $f^* > -\infty$, *then*

$$\inf_{j \ge 0} f(x_j) \le f^* + \delta.$$

Comments on subgradient methods

- Use subgradients and projection $P_{\mathcal{X}}(z)$, suitable for problems with large variable size.
- Easy to get the subgradients and projection $P_{\mathcal{X}}(z)$?
- Some problems: convergent, objective value or iteration point convergent, speed?
- Some research topics: random algorithms like the random selection of directions for objective value decreasing.

Part II Proximal Algorithm

Method

For a given closed proper convex function $f(x)$ over $\mathbb{R}^n$, the optimization problem is $\min_{x\in\mathbb{R}^n} f(x)$, and the proximal algorithm (PA) is

$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^n} \left\{ f(x) + \frac{1}{2c_k}\|x - x_k\|^2 \right\},$$

where c_k is a given control parameter.

Example

$$\min_{0\leq x\leq 3} f(x) = x^2.$$

- *Set* $x_0 = 0$ *and* $c_k = c > 0$. $x_1 = \arg\min_{0\leq x\leq 3}\{x^2 + \frac{x^2}{2c}\}$. *By* $(x^2 + \frac{x^2}{2c})' = 0$, *we get* $x_1 = 0$.
- *Set* $x_0 = 2$ *and* $c > 0$. $x_1 = \arg\min_{0\leq x\leq 3}\{x^2 + \frac{(x-2)^2}{2c}\} = \frac{2}{2c+1}$, *and* $x_2 = \arg\min_{0\leq x\leq 3}\{x^2 + \frac{1}{2c}(x - \frac{2}{2c+1})^2\} = \frac{2}{(2c+1)^2}$.

Proximal function

- Proximal function.

$$F(x, y) = f(x) + \frac{1}{2c_k}(x - y)^T(x - y).$$

- If $f(x)$ is convex, then $F(x, y)$ is strongly convex over x. Strongly convexity: there exists a $\mu > 0$, such that

$$G(z) \geq G(x) + g^T(z - x) + \mu(z - x)^T(z - x), \forall z \in \mathbb{R}^n, g \in \partial G(x).$$

- The minimizer of $F(x, x_k)$ is unique.
- $F(x, y)$ is convex over (x, y).
- If x^* is a global minimizer of $f(x)$, then (x^*, x^*) is a global minimizer of $F(x, y)$.
- If x_k is not a (global or local) minimizer of $f(x)$, then there exists a $\bar{x}$ such that $F(\bar{x}, x_k) < F(x_k, x_k) = f(x_k)$.
- If x_k is not a (global or local) minimizer of $f(x)$, then $f(x_{k+1}) < f(x_k)$.

Convergence

Theorem

If $f(x)$ is convex, x_k and x_{k+1} are two successive iterates of the PA, then

$$\frac{x_k - x_{k+1}}{c_k} \in \partial f(x_{k+1}).$$

Theorem

(Three-term inequality) Consider a closed proper convex function $f(x)$ and the PA for any $x_k \in \mathbb{R}^n$ and $c_k > 0$. Then for any $y \in \mathbb{R}^n$, we have

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2c_k(f(x_{k+1}) - f(y)) - \|x_k - x_{k+1}\|^2.$$

Theorem

(Convergence) Let $f(x)$ be convex and $\{x_k\}$ be a sequence generated by the PA. Then, if $\sum_{k=0}^{\infty} c_k = \infty$, we have $f(x_k) \to f^* = \inf_{x \in \mathbb{R}^n} f(x)$ and if the optimal solution set $\mathcal{X}^* = \arg\min_{x \in \mathbb{R}^n} f(x)$ is nonempty, $\{x_k\}$ converges to some point in $\mathcal{X}^*$.

Theorem

(Rate of convergence) Assume that $f(x)$ is convex, $\mathcal{X}^*$ is nonempty and that for some $\beta > 0,\ \delta > 0$, and $\gamma \geq 1$, we have

$$f^* + \beta(d(x))^{\gamma} \leq f(x), \ \ \forall x \in \mathbb{R}^n \ \text{ with } \ d(x) \leq \delta,$$

where $d(x) = \min_{x^* \in \mathcal{X}^*} \|x - x^*\|$. Suppose

$$\sum_{k=0}^{\infty} c_k = \infty.$$

Then (i) For all k sufficiently large, we have, if $\gamma > 1$,

$$d(x_{k+1}) + \beta c_k (d(x_{k+1}))^{\gamma - 1} \leq d(x_k),$$

and, if $\gamma = 1$ and $x_{k+1} \notin \mathcal{X}^*$,

$$d(x_{k+1}) + \beta c_k \leq d(x_k).$$

Convergence

(ii) (Superlinear convergence) If $1 < \gamma < 2$, $x_k \notin \mathcal{X}^*$ for all k and $\inf_{k \geq 0} c_k > 0$, then

$$\limsup_{k \to \infty} \frac{d(x_{k+1})}{(d(x_k))^{\frac{1}{\gamma - 1}}} < \infty.$$

(iii) (Linear convergence) If $\gamma = 2$, $x_k \notin \mathcal{X}^*$ for all k and $0 < \lim_{k \to \infty} c_k = \bar{c} < \infty$, then

$$\limsup_{k \to \infty} \frac{d(x_{k+1})}{d(x_k)} \leq \frac{1}{1 + \beta \bar{c}},$$

while if $\lim_{k \to \infty} c_k = \infty$, then

$$\lim_{k \to \infty} \frac{d(x_{k+1})}{d(x_k)} = 0.$$

(iv) (Sublinear convergence) If $\gamma > 2$, then

$$\limsup_{k \to \infty} \frac{d(x_{k+1})}{(d(x_k))^{\frac{2}{\gamma}}} = 0.$$

Comments on convergence

$$f^* + \beta(d(x))^\gamma \leq f(x), \quad \forall x \in \mathbb{R}^n \ \text{ with } \ d(x) \leq \delta.$$

- When $f(x)$ is differentiable and x^* be an optimal solution, we have $f'(x^*) = 0$, and by the Taylor formula, for any x near x^*,

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T H(x - x^*) + o(d^2(x)).$$

- If $f(x)$ is strictly convex at x^*, We know that $\gamma \geq 2$.
- c_k is large, the decreasing of $f(x_{k+1})$ may be slow by $f(x_{k+1}) + \frac{1}{2c_k}\|x_{k+1} - x_k\|^2 \leq f(x_k)$.
- The larger c_k is, the weaker the power of the proximal term $\frac{1}{2c_k}\|x - x_k\|^2$ is.
- The larger c_k is, the smaller $d(x_{k+1})$ is by $d(x_{k+1}) + \beta c_k(d(x_{k+1}))^{\gamma-1} \leq d(x_k)$, for $\gamma > 1$, which means x_{k+1} is more closed to the optimal than that of x_k.
- If $\gamma_1 > \gamma_2 \geq 1$, then $(d(x_{k+1}))^{\gamma_1} < (d(x_{k+1}))^{\gamma_2}$ when $d(x_{k+1}) < 1$. Then the convergent rate for the function with γ_2 is higher than that of γ_1.

Further convergence results

Theorem

(Finite convergence) Assume $f(x)$ is convex, $\mathcal{X}^*$ is nonempty and that there exists a scalar $\beta > 0$ such that

$$f^* + \beta d(x) \le f(x), \quad \forall x \in \mathbb{R}^n.$$

If $\sum_{k=0}^{\infty} c_k = \infty$, then the PA converges to $\mathcal{X}^*$ finitely. Furthermore, if $c_0 \ge d(x_0)/\beta$, the algorithm converges in a single step.

Theorem

(Sharp minimum condition for polyhedral functions) Let $f(x)$ be a polyhedral (extended) functions and $\mathcal{X}^* \neq \emptyset$. Then there exists a scalar $\beta > 0$ such that

$$f^* + \beta d(x) \le f(x), \quad \forall x \notin \mathcal{X}^*.$$

Gradient interpretation of PA

- PA: $\phi_c(x_k) = \min_{x\in\mathbb{R}^n}\{f(x) + \frac{1}{2c}\|x - x_k\|^2\}$.
- $x_{k+1} = \arg\min_{x\in\mathbb{R}^n}\{f(x) + \frac{1}{2c}\|x - x_k\|^2\}$.
- Gradient: $x_{k+1} = x_k - cv$ where $v = \nabla\phi_c(x_k)$.

Theorem

The function $\phi_c(x_k)$ is convex and differentiable, and we have

$$\inf_{x\in\mathbb{R}^n} f(x) \leq \phi_c(z) \leq f(z), \quad \forall z \in \mathbb{R}^n,$$

$$\nabla\phi_c(z) = \frac{z - x_c(z)}{c}, \quad \forall z \in \mathbb{R}^n,$$

where $x_c(z)$ is the unique minimizer of the PA iteration. Moreover

$$\nabla\phi_c(z) \in \partial f(x_c(z)), \quad \forall z \in \mathbb{R}^n.$$

Remarks

- The proximal point function

$$F_{c_k}(x, x_k) = f(x) + \frac{1}{2c_k}\|x - x_k\|^2$$

is strongly convex. The next iterate point is unique.

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \{f(x) + \frac{1}{2c_k}\|x - x_k\|^2\}.$$

- Stop criteria: (1) $x_k = x_{k+1}$, (2) $f(x_k) - f(x_{k+1}) \leq \epsilon$ for a given precision $\epsilon > 0$, (3) $\|x_k - x_{k+1}\| \leq \epsilon$ for a given precision $\epsilon > 0$, (4) a finite K, etc..
- So the final iteration point of PPA depends on the initial point x_0.
- An example: $f(x) = 0, x \in [0, 1]$. Any point in $[0, 1]$ is an optimal solution.

$$x_{k+1} = \arg\min\{f(x) + \frac{1}{2c_k}(x - x_k)^2\} = \arg\min\{\frac{1}{2c_k}(x - x_k)^2\} = x_k.$$

So the final and the initial points are the same.

Fixed point interpretation of PA

- In view of the fixed point concept. Define

$$P_{c,f}(z) = \arg\min_{x\in\mathbb{R}^n} \left\{ f(x) + \frac{1}{2c}\|x - z\|^2 \right\}, \quad z \in \mathbb{R}^n.$$

  The fixed point is $\bar{z}$ such that $z = P_{c,f}(z)$.
- We can use the fixed point theory to design algorithms to get optimal solutions.
- Variants

$$N_{c,f}(z) = 2P_{c,f}(z) - z, \ \ \forall z \in \mathbb{R}^n.$$

Theorem

For any $c > 0$ and closed proper convex function $f(x)$, the mapping $N_{c,f}(z) = 2P_{c,f}(z) - z$ is nonexpansive, i.e.,

$$\|N_{c,f}(z_1) - N_{c,f}(z_2)\| \leq \|z_1 - z_2\|, \ \ \forall z_1, z_2 \in \mathbb{R}^n.$$

Moreover, any interpolated mapping as following is nonexpansive

$$(1-\alpha)z + \alpha N_{c,f}(z), \ \ \forall\ 0 \leq \alpha \leq 1.$$

Fixed point theory of PA

Theorem

(Krasnosel'skii-Mann theorem for nonexpansive iterations) Consider a mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ *that is nonexpansive with respect to the Euclidean norm, i.e.,*

$$\|T(x) - T(y)\| \leq \|x - y\|, \ \ \forall x, y \in \mathbb{R}^n,$$

and has at least one fixed point. Then the iteration

$$x_{k+1} = (1 - \alpha_k)x_k + \alpha_k T(x_k), \ \ 0 \leq \alpha_k \leq 1$$

with $\sum_{k=0}^{\infty} \alpha_k(1 - \alpha_k) = \infty$, *converges to a fixed point of* T, *starting from any* $x_0 \in \mathbb{R}^n$.

Theorem

(Stepsize relaxation in the PA) The iteration $x_{k+1} = x_k + \gamma_k(P_{c,f}(x_k) - x_k)$, *where* $\gamma_k \in [\epsilon, 2 - \epsilon]$ *for any* $0 < \epsilon < 2$, *converges to a minimum of* $f(x)$, *assuming at least one minimum exists.*

Dual proximal algorithm (DPA)

- Equivalent reformulation. For

$$\min_{x\in\mathbb{R}^n} \left\{ f(x) + \frac{1}{2c_k}\|x - x_k\|^2 \right\},$$

let $f_1(x) = f(x)$, $f_2(x) = \frac{1}{2c_k}\|x - x_k\|^2$. The equivalent reformulation is

$$\begin{array}{rl} \min & f_1(x_1) + f_2(x_2) \\ \text{s.t.} & x_2 = x_1 \\ & x_1, x_2 \in \mathbb{R}^n. \end{array}$$

- Lagrangian dual

$$\begin{aligned} &\max\nolimits_{\lambda\in\mathbb{R}^n} \min\nolimits_{x_1,x_2\in\mathbb{R}^n} f_1(x_1) + f_2(x_2) + \lambda^T(x_2 - x_1) \\ &= \max\nolimits_{\lambda\in\mathbb{R}^n} \min\nolimits_{x_1,x_2\in\mathbb{R}^n} f_1(x_1) - \lambda^T x_1 + f_2(x_2) - (-\lambda)^T x_2 \\ &= \max\nolimits_{\lambda\in\mathbb{R}^n} -f_1^*(\lambda) - f_2^*(-\lambda) \\ &= -\min\nolimits_{\lambda\in\mathbb{R}^n} f_1^*(\lambda) + f_2^*(-\lambda). \end{aligned}$$

Dual proximal algorithms (DPA)

- Dual functions $f_1^*(\lambda)$ and $f_2^*(\lambda)$.

$$f_1^*(\lambda) = f^*(\lambda) = \sup_{x\in\mathbb{R}^n} \{x^T\lambda - f(x)\}. \text{ Depends on } f(x).$$

$$f_2^*(\lambda) = \sup_{x\in\mathbb{R}^n} \{x^T\lambda - \frac{1}{2c_k}\|x - x_k\|^2\} = x_k^T\lambda + \frac{c_k}{2}\|\lambda\|^2$$

- Lagrangian dual

$$\begin{array}{rl} -\min & f_1^*(\lambda) - x_k^T\lambda + \frac{c_k}{2}\|\lambda\|^2 \\ \text{s.t.} & \lambda \in \mathbb{R}^n. \end{array}$$

- If the primal is proper convex, then the conditions of Fenchel Theorem are satisfied $f_1(x_1), f_2(x_2)$ are proper convex (there exists a lower bound of minimal value), there exists a feasible relative point $x_1 = x_2$, and a cone $\{(x_1, x_2)|x_1 = x_2\}$. Suppose x_{k+1} and λ_{k+1} be optimal solutions of the above.

$$x_{k+1} \in \partial f_1^*(\lambda_{k+1}), x_{k+1} \in \partial f_2^*(-\lambda_{k+1}), (Comments)$$

$$\lambda_{k+1} \in \partial f_1(x_{k+1}), \quad -\lambda_{k+1} \in \partial f_2(x_{k+1}).$$

Dual proximal algorithms (DPA)

- $-\lambda_{k+1} \in \partial f_2(x_{k+1})$ is equivalent to

$$x_{k+1} = \arg \sup_{x \in \mathbb{R}^n} \{-\lambda_{k+1}^T x - f_2(x)\},$$

from which, we get

$$\lambda_{k+1} = \frac{x_k - x_{k+1}}{c_k}.$$

- Dual proximal algorithm. Find

$$\lambda_{k+1} = \arg \min_{\lambda \in \mathbb{R}^n} \left\{ f_1^*(\lambda) - x_k^T \lambda + \frac{c_k}{2} \|\lambda\|^2 \right\},$$

and then $x_{k+1} = x_k - c_k \lambda_{k+1}$.
- Computational complexity depends on $f_1^*(\lambda)$.

Part III Augmented Lagrangian Methods and Alternative Direction Methods of Multipliers

Augmented Lagrangian methods

- Primal problem

$$\begin{array}{rl} \min & f(x) \\ \text{s.t.} & Ax = b \\ & x \in \mathcal{X}, \end{array}$$

  where $f(x)$ is an extended convex function, $\mathcal{X}$ is a convex set, $A \in M(m, n)$ and $b \in \mathbb{R}^m$.

- Reformulation: To solve the following $p(0)$.

$$\begin{array}{rl} p(u) = \min & f(x) \\ \text{s.t.} & Ax - b = u \\ & x \in \mathcal{X}, \end{array}$$

- Lagrangian relaxation for the primal problem at any $\lambda \in \mathbb{R}^m$

$$q(\lambda) = \min_{x \in \mathcal{X}} \left\{ f(x) + \lambda^T (Ax - b) \right\}.$$

Comments on augmented Lagrangian methods

- Augmented and proximal.

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \left\{ f(x) + \lambda_k^T(Ax-b) + \frac{c_k}{2}\|Ax-b\|^2 \right\},$$

$$u_{k+1} = Ax_{k+1} - b,$$

$$\lambda_{k+1} = \arg\min_{\lambda\in\mathbb{R}^m} \{-q(\lambda) + \frac{1}{2c_k}\|\lambda-\lambda_k\|^2\}.\text{(Not solve this!)}$$

$$\lambda_{k+1} = \lambda_k + c_k(Ax_{k+1} - b).$$

- The bigger c_k is, a slower convergent rate of λ may be, and a more feasible solution of x will be.
- Augmented Lagrangian algorithm (ALA):

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \left\{ f(x) + \lambda_k^T(Ax-b) + \frac{c_k}{2}\|Ax-b\|^2 \right\},$$

$$u_{k+1} = Ax_{k+1} - b,$$

$$\lambda_{k+1} = \lambda_k + c_k(Ax_{k+1} - b).$$

Convergence properties of augmented Lagrangian algorithm

Theorem

Suppose $\sum_{k=0}^{\infty} c_k = \infty$, $\inf\{c_k\} > 0$, $p(u)$ *be closed proper and* $p(0)$ *be finite. For the sequence* $\{x_k, \lambda_k\}$, *the sequence* $\{q(\lambda_k)\}$ *converges to the augmented Lagrangian primal and dual optimal value. Moreover if the dual problem* $\max_{\lambda \in \mathbb{R}^m} \min_{x \in \mathcal{X}} \{f(x) + \lambda^T(Ax - b)\}$ *has at least one optimal solution, the following hold: (i)* $\{\lambda_k\}$ *converges to an optimal dual solution. Furthermore, convergence in a finite number of iterations is obtained if* $q(\lambda)$ *is polyhedral. (ii) Every limit point of* $\{x_k\}$ *is an optimal solution of the primal problem* $\min_{x \in \mathcal{X}, Ax=b} f(x)$.

Comments

- $\{x_k\}$ may not be convergent. An example: the objective function is $f(x) = e^{x^1}$, and the constraints are $x^1 + x^2 = 0, x^2 \geq 0$. There is no primal optimal solution.
- The assumptions 'Suppose $p(u)$ be closed proper and $p(0)$ be finite' can be replaced by the primal problem is feasible and lower bounded.

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & Ax = b \\ & x \in \mathcal{X}, \end{array}$$

- The concavity of $q(\lambda)$ can be proved by the function analysis.
- If $f(x)$ is a linear function and $\mathcal{X}$ is a polyhedral or $\mathcal{X} = \mathbb{R}^n$, i.e.,$\min_{Ax=b, x\geq 0} c^T x$ then $q(\lambda)$ is polyhedral. The APA gets its optimal solution in a finite number of steps.

$$q(\lambda) = \min_{x \in \mathbb{R}^n_+} \{c^T x + \lambda^T (Ax - b)\}.$$

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n_+} \left\{ c^T x + \lambda_k^T (Ax - b) + \frac{c_k}{2} \|Ax - b\|^2 \right\}.$$

Variants of ALA—inequality constraints

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathcal{X}, a_i^T x \le b_i, i = 1, 2, \ldots, r. \end{array}$$

- Reformulation

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathcal{X}, z^i \ge 0,\ a_i^T x + z^i = b_i, i = 1, 2, \ldots, r. \end{array}$$

- The AL method

$$\bar{L}_c(x, z, \mu) = f(x) + \sum_{j=1}^{r} \left\{ \mu^j (a_j^T x - b_j + z_j) + \frac{c}{2} (a_j^T x - b_j + z^j)^2 \right\}.$$

$$\min_{x \in \mathcal{X}, z \ge 0} \bar{L}_c(x, z, \mu).$$

- Two stages

$$L_c(x, \mu) = \min_{z \ge 0} \bar{L}_c(x, z, \mu),$$

$$\min_{x \in \mathcal{X}, z \ge 0} \bar{L}_c(x, z, \mu) = \min_{x \in \mathcal{X}} L_c(x, \mu).$$

Variants of ALA—the first-order method

- Augmented Lagrangian method.

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \lambda_k^T(Ax - b) + \frac{c_k}{2}\|Ax - b\|^2 \right\},$$

$$u_{k+1} = Ax_{k+1} - b,$$

$$\lambda_{k+1} = \arg\min_{\lambda \in \mathbb{R}^m} \{-q(\lambda) + \frac{1}{2c_k}\|\lambda - \lambda_k\|^2\}.$$

$$\lambda_{k+1} = \lambda_k + c_k(Ax_{k+1} - b).$$

- Directly solve the following to get $\nabla q_{c_k}(\lambda)$

$$q_{c_k}(\lambda_k) = \min_{\lambda \in \mathbb{R}^m} \{-q(\lambda) + \frac{1}{2c_k}\|\lambda - \lambda_k\|^2\}.$$

Then we have

$$\nabla q_{c_k}(\lambda_k) = \frac{\lambda_{k+1} - \lambda_k}{c_k},$$

and $\lambda_{k+1} = \lambda_k + c_k \nabla q_{c_k}(\lambda_k)$.

Alternative direction methods of multipliers

An example

$$\begin{array}{ll} \min & \sum_{i=1}^{m} f_i(x) \\ \text{s.t.} & x \in \cap_{i=1}^{m} \mathcal{X}_i, \end{array}$$

where $f_i(x)$ are convex functions and $\mathcal{X}_i$ are closed convex sets.

- Reformulation

$$\begin{array}{ll} \min & \sum_{i=1}^{m} f_i(z^i) \\ \text{s.t.} & x = z^i, i = 1, 2, \ldots, m \\ & z^i \in \mathcal{X}_i, i = 1, 2, \ldots, m. \end{array}$$

- The augmented Lagrangian model

$$\begin{array}{l} (x_{k+1}, z_{k+1}^1, \ldots, z_{k+1}^m) = \\ \arg\min \;\; \sum_{i=1}^{m} \left\{ f_i(z^i) + (\lambda_k^i)^T (x - z^i) + \frac{c_k}{2} \|x - z^i\|^2 \right\} \\ \text{s.t.} \;\; x \in \mathbb{R}^n, z^i \in \mathcal{X}_i, i = 1, 2, \ldots, m, \end{array}$$

(parallel or not?)

$$\lambda_{k+1}^i = \lambda_k^i + c_k(x_{k+1} - z_{k+1}^i), \; i = 1, 2, \ldots, m.$$

ADMM: Definition

- Problem
$$\begin{array}{ll} \min & f_1(x) + f_2(Ax) \\ \text{s.t.} & x \in \mathbb{R}^n, \end{array}$$
where $f_1(x), f_2(y)$ are closed proper convex functions.
- Reformulation
$$\begin{array}{ll} \min & f_1(x) + f_2(z) \\ \text{s.t.} & Ax = z \\ & x \in \mathbb{R}^n, z \in \mathbb{R}^m. \end{array}$$
- Augmented Lagrangian function
$$L_c(x, z, \lambda) = f_1(x) + f_2(z) + \lambda^T(Ax - z) + \frac{c}{2}\|Ax - z\|^2.$$
- The ADMM iterations
$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^n} L_c(x, z_k, \lambda_k),$$
$$z_{k+1} \in \arg\min_{z\in\mathbb{R}^m} L_c(x_{k+1}, z, \lambda_k),$$
$$\lambda_{k+1} = \lambda_k + c(Ax_{k+1} - z_{k+1}).$$

ADMM: Convergence and applications

Theorem

Assume that there exists a primal and dual optimal solution pair, and either $\mathrm{dom}(f_1)$ *is compact or* A^TA *is invertible. Then (i) The sequence* $\{x_k, z_k, \lambda_k\}$ *generated by ADMM is bounded, and every limit point of* $\{x_k\}$ *is an optimal solution of the primal problem. Furthermore* $\{\lambda_k\}$ *converges to an optimal dual solution. (ii) The residual sequence* $\{Ax_k - z_k\}$ *converges to 0, and if* A^TA *is invertible, then* $\{x_k\}$ *converges to an optimal primal solution.*

- Base pursuit.

$$
\begin{array}{rl}
\min & \|x\|_1 \\
\text{s.t.} & Cx = b \\
& x \in \mathbb{R}^n.
\end{array}
$$

Reformulation

$$
\begin{array}{rl}
\min & f_1(x) + f_2(z) \\
\text{s.t.} & x = z \\
& x \in \mathbb{R}^n,
\end{array}
$$

where $f_1(x) = \begin{cases} 0, & x \in \{x | Cx = b\} \\ \infty, & \text{otherwise,} \end{cases} \quad f_2(z) = \|z\|_1.$

ADMM: Base pursuit

- The augmented Lagrangian function

$$L_c(x, z, \lambda) = \begin{cases} \|z\|_1 + \lambda^T(x - z) + \frac{c}{2}\|x - z\|^2, & \text{if } Cx = b \\ +\infty, & \text{otherwise,} \end{cases}$$

- The ADMM iterations

$$x_{k+1} \in \arg\min_{Cx=b} \left\{ \lambda_k^T x + \frac{c}{2}\|x - z_k\|^2 \right\}, \textit{Oracle}$$

$$\begin{aligned} z_{k+1} \in &\quad \arg\min_{z\in\mathbb{R}^n} \left\{ \|z\|_1 - \lambda_k^T z + \tfrac{c}{2}\|x_{k+1} - z\|^2 \right\} \\ = &\quad \arg\min_{z\in\mathbb{R}^n} \left\{ \|z\|_1 + \tfrac{c}{2}\|z - (x_{k+1} + \tfrac{\lambda_k}{c})\|^2 \right\} \end{aligned} \quad (\textit{Easy})$$

$$\lambda_{k+1} = \lambda_k + c(Ax_{k+1} - z_{k+1}).(\textit{Easy})$$

- The ith component of z_{k+1} is calculated by

$$(z_{k+1})_i = \begin{cases} (x_{k+1} + \frac{\lambda_k}{c})_i - \frac{1}{c}, & \text{if } (x_{k+1} + \frac{\lambda_k}{c})_i > \frac{1}{c} \\ 0, & \text{if } |(x_{k+1} + \frac{\lambda_k}{c})_i| \le \frac{1}{c} \\ (x_{k+1} + \frac{\lambda_k}{c})_i + \frac{1}{c}, & \text{if } |(x_{k+1} + \frac{\lambda_k}{c})_i| < -\frac{1}{c} \end{cases}$$

ADMM: l_1-regularization

$$\begin{array}{ll} \min & f(x) + \gamma\|x\|_1 \\ \text{s.t.} & x \in \mathbb{R}^n. \end{array}, \quad \begin{array}{ll} \min & f_1(x) + f_2(z) \\ \text{s.t.} & x = z \\ & x \in \mathbb{R}^n, \end{array}$$

where $f_1(x) = f(x)$ and $f_2(z) = \gamma\|z\|_1$.

- The augmented Lagrangian function

$$L_c(x, z, \lambda) = f(x) + \gamma\|z\|_1 + \lambda^T(x - z) + \frac{c}{2}\|x - z\|^2.$$

- The ADMM iterations

$$x_{k+1} \in \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \lambda_k^T x + \frac{c}{2}\|x - z_k\|^2 \right\},$$

$$z_{k+1} \in \arg\min_{z \in \mathbb{R}^n} \left\{ \gamma_k\|z\|_1 - \lambda_k^T z + \frac{c}{2}\|x_{k+1} - z\|^2 \right\},$$

$$\lambda_{k+1} = \lambda_k + c(x_{k+1} - z_{k+1}).$$

- The ith component of z_{k+1} is calculated by

$$(z_{k+1})_i = \begin{cases} (x_{k+1} + \frac{\lambda_k}{c})_i - \frac{\gamma}{c}, & \text{if } (x_{k+1} + \frac{\lambda_k}{c})_i > \frac{\gamma}{c} \\ 0, & \text{if } |(x_{k+1} + \frac{\lambda_k}{c})_i| \leq \frac{\gamma}{c} \\ (x_{k+1} + \frac{\lambda_k}{c})_i + \frac{\gamma}{c}, & \text{if } (x_{k+1} + \frac{\lambda_k}{c})_i < -\frac{\gamma}{c} \end{cases}$$

ADMM: Least absolute deviation problems

$$\begin{array}{cl} \min & \|Cx - b\|_1 \\ \text{s.t.} & x \in \mathbb{R}^n, \end{array} \qquad \begin{array}{cl} \min & f_1(x) + f_2(z) \\ \text{s.t.} & Cx - b = z \\ & x \in \mathbb{R}^n, \end{array}$$

where $C_{m\times n}$ is of rank n, $f_1(x) = 0$ and $f_2(z) = \|z\|_1$.

- The augmented Lagrangian function

$$L_c(x, z, \lambda) = \|z\|_1 + \lambda^T(Cx - z - b) + \frac{c_k}{2}\|Cx - z - b\|^2.$$

- The ADMM iterations

$$x_{k+1} = (C^T C)^{-1} C^T (z_k + b - \frac{\lambda_k}{c_k}),$$

$$z_{k+1} \in \arg\min_{z\in\mathbb{R}^m} \left\{ \|z\|_1 - \lambda_k^T z + \frac{c_k}{2}\|Cx_{k+1} - z - b\|^2 \right\},$$

$$\lambda_{k+1} = \lambda_k + c_k(Cx_{k+1} - z_{k+1} - b).$$

- The ith component of z_{k+1} is calculated by

$$(z_{k+1})_i = \begin{cases} (Cx_{k+1} - b + \frac{\lambda_k}{c_k})_i - \frac{1}{c_k}, & \text{if } (Cx_{k+1} - b + \frac{\lambda_k}{c})_i > \frac{1}{c}, \\ 0, & \text{if } |(Cx_{k+1} - b + \frac{\lambda_k}{c})_i| \le \frac{1}{c} \\ (Cx_{k+1} - b + \frac{\lambda_k}{c})_i + \frac{1}{c}, & \text{if } (Cx_{k+1} - b + \frac{\lambda_k}{c})_i < -\frac{1}{c} \end{cases}$$

ADMM: Separate problems

$$\begin{array}{ll} \min & \sum_{i=1}^m f_i(x^i) \\ \text{s.t.} & \sum_{i=1}^m A_i x^i = b \\ & x^i \in \mathcal{X}_i, \end{array}$$

where $f_i(x^i) : \mathbb{R}^{n_i} \to \mathbb{R}$ are convex function, $\mathcal{X}_i$ are closed convex sets.

- The augmented Lagrangian function

$$L_c(x^1, x^2, \ldots, x^m, \lambda) = \sum_{i=1}^m f_i(x^i) + \lambda^T(\sum_{i=1}^m A_i x^i - b) + \frac{c}{2}\|\sum_{i=1}^m A_i x^i - b\|^2.$$

- The ADMM iterations

$$x_{k+1}^i \in \arg\min_{x^i \in \mathcal{X}_i} L_c(x_{k+1}^1, \ldots, x_{k+1}^{i-1}, x^i, x_{k+1}^{i+1}, \ldots, x_{k+1}^m, \lambda_k),$$

$$i = 1, 2, \ldots, m,$$

$$\lambda_{k+1} = \lambda_k + c(\sum_{i=1}^m A_i x_{k+1}^i - b).$$

Separate problems: Reformulation

- Reformulation

$$
\begin{array}{ll}
\min & \sum_{i=1}^m f_i(x^i) \\
\text{s.t.} & A_i x^i = z^i \\
& \sum_{i=1}^m z^i = b \\
& x^i \in \mathcal{X}_i, z^i \in \mathbb{R}^r,
\end{array}
$$

  where $b \in \mathbb{R}^r$ is given.
- The augmented Lagrangian function

$$
L_c(x, z, p) = \sum_{i=1}^m f_i(x^i) + (p^i)^T (A_i x^i - z^i) + \frac{c}{2} \|A_i x^i - z^i\|^2, \;\; x \in \mathcal{X}, \; z \in \mathcal{Z},
$$

  where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m$ and $\mathcal{Z} = \left\{ z \mid \sum_{i=1}^m z^i = b \right\}$.
- ADMM iteration

$$
x_{k+1}^i \in \arg \min_{x^i \in \mathcal{X}_i} \left\{ f_i(x^i) + (A_i x^i - z_k^i)^T p_k^i + \frac{c}{2} \|A_i x^i - z_k^i\|^2 \right\},
$$

$$
i = 1, 2, \ldots, m.
$$

Separate problems: Reformulation (cont'd-1)

- ADMM iterations: z and p iterations

$$z_{k+1} \in \arg \min_{\sum_{i=1}^m z^i = b} \left\{ \sum_{i=1}^m (A_i x_{k+1}^i - z^i)^T p_k^i + \frac{c}{2} \|A_i x_{k+1}^i - z^i\|^2 \right\},$$

$$p_{k+1}^i = p_k^i + c(A_i x_{k+1}^i - z_{k+1}^i), i = 1, 2, \ldots, m.$$

- To get z_{k+1} by the z iteration, to relax $\sum_{i=1}^m z^i = b$ and to solve

$$\min_{z^i \in \mathbb{R}^r} \left\{ (A_i x_{k+1}^i - z^i)^T p_k^i + \frac{c}{2} \|A_i x_{k+1}^i - z^i\|^2 + \lambda_{k+1}^T z^i \right\},$$

we get

$$z_{k+1}^i = A_i x_{k+1}^i + \frac{p_k^i - \lambda_{k+1}}{c}.$$

Then

$$\lambda_{k+1} = p_k^i + c(A_i x_{k+1}^i - z_{k+1}^i) = p_{k+1}^i, \; i = 1, 2, \ldots, m.$$

Separate problems: Reformulation (cont'd-2)

- λ_{k+1} (using $\sum_{i=1}^m z_{k+1}^i = b, p_k^i = \lambda_k$)

$$\lambda_{k+1} = \lambda_k + \frac{c}{m}(\sum_{i=1}^m A_i x_{k+1}^i - b).$$

- $z_{k+1}^i,\ i = 1, 2, \ldots, m.$

$$z_{k+1}^i = A_i x_{k+1}^i + \frac{\lambda_k - \lambda_{k+1}}{c} = A_i x_{k+1}^i - \frac{1}{m}(\sum_{i=1}^m A_i x_{k+1}^i - b).$$

- $p_{k+1}^i,\ i = 1, 2, \ldots, m.$

$$p_{k+1}^i = p_k^i + c(A_i x_{k+1}^i - z_{k+1}^i).$$

- An application: constrained ADMM

$$\begin{array}{ll} \min & f_1(x) + f_2(Ax) \\ \text{s.t.} & Ex = d, x \in \mathcal{X}. \end{array} \rightarrow \begin{array}{ll} \min & f_1(x^1) + f_2(x^2) \\ \text{s.t.} & \begin{pmatrix} A \\ E \end{pmatrix} x^1 + \begin{pmatrix} -I \\ 0 \end{pmatrix} x^2 = \begin{pmatrix} 0 \\ d \end{pmatrix} \\ & x^1 \in \mathcal{X}, x^2 \in \mathbb{R}^n. \end{array}$$

Comments on ADMM

- The convergence conditions.
- The complexity to solve the key optimization problem depends on the properties of $f_i(x^i)$ and $\mathcal{X}_i$. For example,

$$x_{k+1}^i \in \arg\min_{x^i \in \mathcal{X}_i} \left\{ f_i(x^i) + (A_i x^i - z_k^i)^T p_k^i + \frac{c}{2} \|A_i x^i - z_k^i\|^2 \right\}.$$

When $\mathcal{X}_i$ is a polyhedral set, $f_i(x^i)$ is a convex quadratic function, then it is solved in polynomial time.