Subgradient Methods

9:20 PM

Contents of this page:

Slide1

[Subgradient method description]

Contents of this page:

$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$

$x^{(k)}$: kth iterate

$g^{(k)} \in \partial f(x^{(k)})$

$\alpha_k \in \mathbb{R}_{++}$, step size // rule for step size:

- nonsummable diminishing: step sizes satisfy $\lim_{k\to\infty} \alpha_k = 0$, $\sum_{k=1}^{\infty} \alpha_k = \infty$
- $\alpha_k = \alpha$ (constant)
- $\alpha_k = \frac{\gamma}{\|g^{(k)}\|_2}$
- square summable but not summable $\left(\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \sum_{k=1}^{\infty} \alpha_k = \infty\right)$ e.g. $\alpha_k = \frac{1}{k}$

$f_{best}^{(k)} = \min_{i=1,\dots,k} f(x^{(i)})$ (যেহেতু descent method না, so best known point কে keep track করতে হবে.

Assumptions: * finite minimum exists $\leftrightarrow f^* = \inf_x f(x) > -\infty$, $f(x^*) = f^*$

and obviously the minimizer lies in domain of f

* $\underbrace{\forall_{g \in \partial f} \|g\|_2 \le G}_{\text{the subgradients are bounded at all points}}$ $\leftrightarrow$ equivalent to Lipschitz condition on f

A function $f(x)$ satisfies the Lipschitz condition of order β at $x = 0$ if

$$|f(h) - f(0)| \le B\,|h|^{\beta}$$

for all $|h| < \epsilon$, where B and β are independent of h, $\beta > 0$, and α is an upper bound for all β for which a finite B exists.

* We are at a finite distance from the optimum point at the beginning, i.e., $\|x^{(1)} - x^*\|_2 \le R$

[Convergence Proof: Subgradient Method]

Contents of this page: ret: [Projected Subgradient Method]

*Convergence results:

* $\bar{f} = \lim_{k\to\infty} f_{best}^{(k)}$ হবে নিচে:

- constant step size: $(\alpha_k = \alpha)$ $0 \le \bar{f} - f^* \le \frac{G^2\alpha}{2}$ → upper bound of suboptimality; $\frac{G^2\alpha}{2}$ suboptimal; α small enough হলে optimality achieve হবে.
- constant step length: $(\alpha_k = \gamma / \|g^{(k)}\|_2)$ $0 \le \bar{f} - f^* \le \frac{G\gamma}{2}$ ($\frac{G\gamma}{2}$ suboptimal)
- diminishing step sizes: $\bar{f} = f^*$: converges
(either $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, $\sum_{k=1}^{\infty} \alpha_k = \infty$
or $\lim_{k\to\infty} \alpha_k = 0$, $\sum_{k=1}^{\infty} \alpha_k = \infty$)

Convergence proof:

যা কমবে with iterations: distance from the optimal set.

$\Theta = \{x^* \mid \forall_{x \in \mathrm{dom} f}\ f(x^*) \le f(x)\}$

let $x^* \in \Theta$

$\|x^{(k+1)} - x^*\|_2^2 = \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 = \left((x^{(k)} - x^*) - \alpha_k g^{(k)}\right)^T \left((x^{(k)} - x^*) - \alpha_k g^{(k)}\right)$

$x^{(k+1)} = x^{(k)} + \alpha^k(-g^{(k)})$

$g^{(k)}$ subgradient at $f(x^{(k)})$

$= \left((x^{(k)} - x^*)^T - \alpha_k g^{(k)T}\right)\left((x^{(k)} - x^*) - \alpha_k g^{(k)}\right)$

$= \|x^{(k)} - x^*\|_2^2 + \alpha_k^2 \|g^{(k)}\|_2^2 \underbrace{- 2\alpha_k g^{(k)T}(x^{(k)} - x^*)}_{= +2\alpha_k(-g^{(k)T}(x^{(k)} - x^*))}$

$\le \|x^{(k)} - x^*\|_2^2 + \alpha_k^2 \|g^{(k)}\|_2^2 + 2\alpha_k\left(f(x^*) - f(x^{(k)})\right)$

$$\leq \|x^{(k)}-x^*\|_2^2+\alpha_k^2\|g^{(k)}\|_2^2+2\alpha_k\left(f(x^*)-f(x^{(k)})\right)$$

$$= \|x^{(k)}-x^*\|_2^2+\alpha_k^2\|g^{(k)}\|_2^2-2\alpha_k\left(f(x^{(k)})-f(x^*)\right)$$

$$g^{(k)} \in \partial f(x^{(k)}) \leftrightarrow \forall_{\bar{y}\in \mathrm{dom} f}\ f(\bar{y}) \geq f(x^{(k)})+g^{(k)T}(\bar{y}-x^{(k)})$$

$$\bar{y}=x^*$$

$$\rightarrow f(x^*) \geq f(x^{(k)})+g^{(k)T}(x^*-x^{(k)}) \ \wedge\ f(x^*)\leq f(x^{(k)})$$

$$\rightarrow g^{(k)T}(x^*-x^{(k)}) \leq f(x^*)-f(x^{(k)}) \leq 0$$

$$\rightarrow -g^{(k)T}(x^{(k)}-x^*) \leq f(x^*)-f(x^{(k)}) \leq 0$$

$$\therefore \|x^{(k+1)}-x^*\|_2^2 \leq \|x^{(k)}-x^*\|_2^2+\alpha_k^2\|g^{(k)}\|^2-2\alpha_k\left(f(x^{(k)})-f(x^*)\right) \leq \|x^{(k-1)}-x^*\|_2^2+\alpha_{k-1}^2\|g^{(k-1)}\|_2^2-2\alpha_{k-1}\left(f(x^{(k-1)})-f(x^*)\right)+\alpha_k^2\|g^{(k)}\|_2^2-2\alpha_k\left(f(x^{(k)})-f(x^*)\right)$$

$$= \|x^{(k-1)}-x^*\|_2^2+\sum_{i=k}^{k-1}\alpha_i^2\|g^{(i)}\|_2^2-2\sum_{i=k}^{k-1}\alpha_i\left(f(x^{(i)})-f(x^*)\right)$$

similarly

$$\|x^{(k)}-x^*\|_2^2 \leq \|x^{(k-1)}-x^*\|_2^2+\alpha_{k-1}^2\|g^{(k-1)}\|^2-2\alpha_{k-1}\left(f(x^{(k-1)})-f(x^*)\right)$$

So:

$$\|x^{(k+1)}-x^*\|_2^2 \leq \|x^{(k-1)}-x^*\|_2^2+\sum_{i=k}^{k-1}\alpha_i^2\|g^{(i)}\|_2^2-2\sum_{i=k}^{k-1}\alpha_i\left(f(x^{(i)})-f(x^*)\right)$$

Same recursion বারবার প্রয়োগ করে এবং এদের structure কে exploit করে আমরা easily generalize করতে পারি:

$$\|x^{(k+1)}-x^*\|_2^2 \leq \|x^{(1)}-x^*\|_2^2+\sum_{i=k}^{1}\alpha_i^2\|g^{(i)}\|_2^2-2\sum_{i=k}^{1}\alpha_i\left(f(x^{(i)})-f(x^*)\right)$$

$$\rightarrow 0 \leq \|x^{(k+1)}-x^*\|_2^2 \leq \|x^{(1)}-x^*\|_2^2+\sum_{i=1}^{k}\alpha_i^2\|g^{(i)}\|_2^2-2\sum_{i=1}^{k}\alpha_i\left(f(x^{(i)})-f(x^*)\right)$$

$$\rightarrow 0 \leq \underbrace{\|x^{(1)}-x^*\|_2^2}_{\leq R^2}+\sum_{i=1}^{k}\alpha_i^2\|g^{(i)}\|_2^2-2\sum_{i=1}^{k}\alpha_i\left(f(x^{(i)})-f(x^*)\right) \leq R^2+\sum_{i=1}^{k}\alpha_i^2\|g^{(i)}\|_2^2-2\sum_{i=1}^{k}\alpha_i\left(f(x^{(i)})-f(x^*)\right)$$ equation 1

$$\Big\{ \text{now, } \forall_{i\in\{1,\dots,k\}}\ f_{best}^{(k)}=\min_{i=1,\dots,k} f(x^{(i)}) \leq f(x^{(i)})$$

$$\rightarrow \forall_{i\in\{1,\dots,k\}}\ f_{best}^{(k)}-f(x^*) \leq f(x^{(i)})-f(x^*)$$

$$\rightarrow \forall_{i\in\{1,\dots,k\}}\ \alpha_i\left(f_{best}^{(k)}-f(x^*)\right) \leq \alpha_i\left(f(x^{(i)})-f(x^*)\right) \quad [\because \alpha_i>0]$$

$$\rightarrow \sum_{i=k}^{1}\alpha_i\left(f_{best}^{(k)}-f(x^*)\right) \leq \sum_{i=k}^{1}\alpha_i\left(f(x^{(i)})-f(x^*)\right)$$

constant so বাইরে আনা যায়

$$\rightarrow -2\sum_{i=k}^{1}\alpha_i\left(f_{best}^{(k)}-f(x^*)\right) \geq -2\sum_{i=k}^{1}\alpha_i\left(f(x^{(i)})-f(x^*)\right)\Big\}$$

$$\equiv -2\sum_{i=k}^{1}\alpha_i\left(f(x^{(i)})-f(x^*)\right) \leq -2\left(f_{best}^{(k)}-f(x^*)\right)\sum_{i=k}^{1}\alpha_i$$

$$\rightarrow\ 0 \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2 - 2\left(f_{best}^{(k)} - f(x^*)\right)\sum_{i=1}^{k}\alpha_i$$

$$\rightarrow\ 2\left(f_{best}^{(k)} - f(x^*)\right)\sum_{i=1}^{k}\alpha_i \leq R^2 + \sum_{i=1}^{k}\alpha_i^2 \|g^{(i)}\|_2^2$$

$$\left(f_{best}^{(k)} - f(x^*)\right) \leq \frac{R^2 + \sum_{i=1}^{k}\alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^{k}\alpha_i}$$

& now $\|x^{(1)} - x^*\|_2 \leq R$, $\forall_i\ \|g^{(i)}\| \leq G$ by assumption

$$\therefore \sum_{i=1}^{k}\alpha_i^2 \|g^{(i)}\|_2^2 \leq \sum_{i=1}^{k}\alpha_i^2 G^2 = G^2 \sum \alpha_i$$

$$\boxed{\forall_{k \in \{1,\dots\}}\ 0 \leq f_{best}^{(k)} - f(x^*) \leq \frac{R^2 + \sum_{i=1}^{k}\alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^{k}\alpha_i} \leq \frac{R^2 + G^2\sum \alpha_i^2}{2\sum_{i=1}^{k}\alpha_i}}$$

eq: inequality of subgradient convergence

Optimality bound for different step sizes   Contents of this page:

* now set constant step size $\alpha_i = \alpha$ then the right hand side of the inequality becomes:

$$\frac{R^2 + G^2 k \alpha^2}{2 k \alpha}$$

∴ constant step size:

$$\underbrace{\lim_{k\to\infty} 0}_{0} \leq \lim_{k\to\infty} f_{best}^{(k)} - f(x^*) \leq \lim_{k\to\infty} \frac{\cancel{k}\left(\frac{R^2}{k} + G^2\alpha^2\right)}{2\alpha \cancel{k}} = \frac{G^2 \cancel{\alpha}\alpha}{2\cancel{\alpha}} = \frac{G^2\alpha}{2}\ ✓$$

* constant step length: $\alpha_i = \dfrac{\gamma}{\|g^{(i)}\|_2}$,

then Left R.H.S becomes

$$\frac{R^2 + \sum_{i=1}^{k}\frac{\gamma^2}{\cancel{\|g^{(i)}\|_2^2}}\cdot\cancel{\|g^{(i)}\|_2^2}}{2\sum_{i=1}^{k}\frac{\gamma}{\|g^{(i)}\|_2}} = \frac{R^2 + k\gamma^2}{2\gamma\sum_{i=1}^{k}\frac{1}{\|g^{(i)}\|_2}} = \frac{R^2 + k\gamma^2}{2\gamma}\cdot\frac{1}{\sum_{i=1}^{k}\frac{1}{\|g^{(i)}\|_2}} \leq \frac{R^2 + k\gamma^2}{2\gamma}\cdot\frac{G}{k} = \frac{k\left(\frac{R^2}{k} + \gamma^2\right)\cdot G}{2\gamma\cdot\cancel{k}} = \frac{\left(\frac{R^2}{k} + \gamma^2\right)G}{2\gamma}$$

$$\#\ \|g^{(i)}\|_2 \leq G \rightarrow \frac{1}{\|g^{(i)}\|_2} \geq \frac{1}{G} \rightarrow \sum_{i=1}^{k}\frac{1}{\|g^{(i)}\|_2} \geq \sum_{i=1}^{k}\frac{1}{G} = \frac{k}{G} \rightarrow \frac{1}{\sum_{i=1}^{k}\frac{1}{\|g^{(i)}\|_2}} \leq \frac{G}{k}$$

$$\therefore \lim_{k\to\infty} 0 \le \lim_{k\to\infty} \left(f_{best}^{(k)} - f^*\right) \le \lim_{k\to\infty} \frac{\left(\frac{R^2}{k} + \gamma^2\right)G}{2\gamma} = \frac{G\gamma^2}{2\gamma} = \frac{G\gamma}{2}$$

$$\therefore 0 \le \lim_{k\to\infty} f_{best}^{(k)} - f^* \le \frac{G\gamma}{2} \quad \# \ \frac{G\gamma}{2} \text{ optimal}$$

④ diminishing step size:

obviously $\lim_{k\to\infty} \sum_{i=1}^{k} \alpha_i = \infty$, $\lim_{k\to\infty} \sum_{i=1}^{k} \alpha_i^2 = \text{finite}$ (say $\alpha_i = \frac{1}{i}$) 같은 series

select 하면

$$\lim_{k\to\infty} f_{best}^{(k)} - f^* = 0 \text{ 된다 !}$$

3.3 A bound on the suboptimality bound

It's interesting to ask the question, what sequence of step sizes minimizes the righthand side of (3)? In other words, how do we choose positive $\alpha_1, \dots, \alpha_k$ so that

$$\frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}$$

(which is an upper bound on $f_{best}^{(k)} - f^*$) is minimized? This is a convex and symmetric function of $\alpha_1, \dots, \alpha_k$, so we conclude the optimal occurs when all α_i are equal (to, say, α). This reduces our suboptimality bound to

$$\frac{R^2 + G^2 k \alpha^2}{2k\alpha}$$

(eq:cnvgncSbgrdnt)

which is minimized by $\alpha = (R/G)/\sqrt{k}$.

In other words, the choice of $\alpha_1, \dots, \alpha_k$ that minimizes the suboptimality bound (3) is given by

$$\alpha_i = (R/G)/\sqrt{k}, \quad i = 1, \dots, k.$$

8

3.3 A bound on the suboptimality: eq: inequality of subgradient convergence equation의 bound 는 finite number of step 과 tightest 하지 않다

(단) step size select 하면되나?

$$\min_{(\alpha_i)_{i=1}^{k}} \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} = ?$$

$$\frac{R^2 + G^2 \alpha^T \alpha}{2 \mathbf{1}^T \alpha} = \frac{R^2}{2} \frac{1}{\mathbf{1}^T\alpha} + \frac{G^2}{2} \cdot \frac{\|\alpha\|_2^2}{\mathbf{1}^T\alpha}$$

$\therefore$ ( ) convex in α as this is a nonnegative weighted sum of two convex functions.

convex in α: proof: $f(x) = (c^T x)^{-1}$, $c^T x > 0$

$$\nabla f(x) = \left(\frac{\partial}{\partial x_i} \cdot \frac{1}{\sum_{j=1}^{n} c_j x_j}\right)_{i=1}^{n} = \left(-\left(\sum_{j=1}^{n} c_j x_j\right)^{-2} \cdot c_i\right)$$

$$\nabla^2 f(x) = \left[\frac{\partial}{\partial x_j}\frac{\partial}{\partial x_i} f(x)\right] = \left[\frac{\partial}{\partial x_j}\left[-c_i \left(\sum_{k=1}^{n} c_k x_k\right)^{-2}\right]\right]$$

$$= -c_i \left[-2 \cdot \left(\sum_{k=1}^{n} c_k x_k\right)^{-3} c_j\right] = \frac{2 c_i c_j}{\left(\sum_{k=1}^{n} c_k x_k\right)^3} \quad \therefore \forall x \in \operatorname{dom} f,\ \nabla^2 f(x) = \frac{2}{(c^T x)^3}\left(\left[c_i c_j\right]_{i=1,j=1}^{n\ \ n}\right)$$

$$= \frac{2}{(c^T x)^3} c c^T \succeq 0 \quad \therefore \frac{1}{c^T x} \ \{\text{convex}_{x : c^T x > 0}\}$$

$(c^Tx)^3 > 0$; cc^T: Cholesky decomposition $\Leftrightarrow$ positive semidefinite

$$\therefore \frac{1}{\mathbf{1}^T \alpha} \ \{\text{convex}_{\alpha : \mathbf{1}^T\alpha > 0}\}$$

Alternative proof: Scalar composition rule $f(x) = h(g(x))$ is convex now $g(x) = c^T x + d$ is affine so concave

h: convex, $\uparrow$; g: concave; $\operatorname{dom} f = \{x \in \operatorname{dom} g \mid g(x) \in \operatorname{dom} h\}$; $\operatorname{dom} g = \mathbb{R}^n$; $\mathbb{R}^n \to \mathbb{R}$

$h(x) = \frac{1}{x}$ is convex $\{\because x^{a}: a \ge 1 \vee a \le 0 : \text{convex on } \mathbb{R}_{++}\}$

also for $a \le 0$, $(x^a : (\text{convex}, \downarrow)_{\mathbb{R}_{++}})$ $\therefore \operatorname{dom} h = \mathbb{R}_{++}$

$x^{a}: a \le 0$

$\mathbb{R}_{++}$

$$\therefore f(x) = h(g(x)) = \left(\left(\frac{1}{\square}\right) \circ \left(c^T \square + d\right)\right)(x) = \left(\frac{1}{\square}\right)\left(c^T x + d\right) = \frac{1}{c^T x + d} \text{ is convex on } \{x \in \mathbb{R}^n \mid c^T x + d > 0\}$$

convex, $\downarrow$ on $\mathbb{R}_{++}$; $\square$

$\operatorname{dom} f = \{x \in \operatorname{dom} g \mid g(x) \in \operatorname{dom} h\} = \{x \in \mathbb{R}^n \mid g(x) = c^T x + d \in \operatorname{dom} h = \mathbb{R}_{++}\} = \{x \in \mathbb{R}^n \mid c^T x + d > 0\}$

$$\# \ c c^T = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} [c_1\ c_2 \cdots c_n] = \begin{bmatrix} c_1 [c_1\ c_2 \cdots c_n] \\ c_2 [c_1\ c_2 \cdots c_n] \\ \vdots \\ c_n [c_1\ c_2 \cdots c_n] \end{bmatrix} = \begin{bmatrix} c_1^2 & c_1 c_2 & \cdots & c_1 c_n \\ c_2 c_1 & c_2^2 & \cdots & c_2 c_n \\ \vdots & & & \vdots \\ c_n c_1 & c_n c_2 & \cdots & c_n^2 \end{bmatrix} = [c_i c_j]_{i=1, j=1}^{n\ \ n}$$

convex in α

proof: say $f(x) = x^T x$ convex in x

perspective

$$g(x) = (c^T x + d)\, f\!\left(\frac{Ax+b}{c^T x + d}\right) \{\text{convex}_x\} \quad \text{// Boyd Example 3.20 // page 90}$$

$c^Tx+d > 0$

$$= (c^T x + d) \frac{(Ax+b)^T}{(c^T x + d)} \cdot \frac{(Ax+b)}{(c^T x + d)}$$

$$= \frac{\|Ax+b\|_2^2}{(c^T x + d)} \{\text{convex}_x\} \quad c^Tx + d > 0$$

$$\therefore \frac{\|\alpha\|_2^2}{\mathbf{1}^T\alpha} \{\text{convex}_\alpha\}$$

This choice of constant step size yields the suboptimality bound

(eq:cnvgncSbgrdnt)

$$f_{best}^{(k)} - f^* \le RG/\sqrt{k}.$$

Put another way, we can say that for *any* choice of step sizes, the suboptimality bound (3) must be at least as large as $RG/\sqrt{k}$. If we use (3) as our stopping criterion, then the number of steps to achieve a guaranteed accuracy of ϵ will be at least $(RG/\epsilon)^2$, no matter what step sizes we use. (It will be this number if we use the step size $\alpha_k = (R/G)/\sqrt{k}$).

Note that RG has a simple interpretation as an initial bound on $f(x^{(1)}) - f^*$, based on $\|x^{(1)} - x^*\|_2 \le R$ and the Lipschitz constant G for f. Thus $(RG)/\epsilon$ is the ratio of initial uncertainty in f^* to final uncertainty in f^*. If we square this number, we get the minimum number of steps it will take to achieve this reduction in uncertainty. This tells us that the

So,

$$\phi(\alpha) = \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{\dots} \text{ is a convex function which is}$$

Note that RG has a simple interpretation as an initial bound on $f(x^{(1)}) - f^\star$, based on $\|x^{(1)} - x^\star\|_2 \le R$ and the Lipschitz constant G for f. Thus $(RG)/\epsilon$ is the ratio of initial uncertainty in $f^\star$ to final uncertainty in $f^\star$. If we square this number, we get the minimum number of steps it will take to achieve this reduction in uncertainty. This tells us that the subgradient method is going to be very slow, if we use (3) as our stopping criterion. To reduce the initial uncertainty by a factor of 1000, say, it will require at least 10^6 iterations.

$1^T\alpha$

$$\phi(\alpha) = \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}$$

is a convex function which is symmetric in α as well, so at optimal α^*, $\alpha_1^* = \alpha_2^* = \dots = \alpha_n^*$ that can be found by setting $\alpha_1^* = \alpha_2^* = \dots = \alpha_n^* = \alpha^*$ in

$$\alpha = \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha^2}{2G^2 \sum_{i=1}^{k} \alpha} = \frac{R^2 + G^2 k\alpha^2}{2G^2 k\alpha}$$

$\rightarrow 2G^2 k\alpha^2 = R^2 + G^2 k\alpha^2$

$\rightarrow G^2 k\alpha^2 = R^2 \rightarrow \alpha = \frac{R}{G} \cdot \frac{1}{\sqrt{k}} = \alpha_1 = \dots = \alpha_n$

$\therefore \forall_j \; \alpha_j = \frac{R}{G} \cdot \frac{1}{\sqrt{k}}$

// rigorous explanation: $\nabla_\alpha \phi(\alpha) = 0 \leftrightarrow \forall_j \; \frac{\partial}{\partial \alpha_j} \phi(\alpha) = 0$

$$\frac{\partial}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j}\left(\frac{R^2 + G^2\|\alpha\|_2^2}{1^T\alpha}\right) = \frac{(1^T\alpha)(G^2 2\alpha_j) - (R^2 + G^2\|\alpha\|_2^2)\,1}{\underbrace{(1^T\alpha)^2}_{\neq 0}} = 0 \rightarrow 2G^2(1^T\alpha)\,\alpha_j = R^2 + G^2\|\alpha\|_2^2$$

$\therefore \forall_j \; \alpha_j = \frac{R^2 + G^2\|\alpha\|_2^2}{2G^2(1^T\alpha)}$

$\forall_j \; \alpha_j^* = \frac{R^2 + G^2\|\alpha^*\|_2^2}{2G^2\, 1^T\alpha^*}$

$\alpha = \alpha_1^* = \alpha_2^* = \dots = \alpha_k^* = \frac{R^2 + G^2\|\alpha^*\|_2^2}{2G^2(1^T\alpha)}$

f(x) : convex in (x,x) <=> f(x) : convex in x

If a function f(x) is convex in (x,x) <=> f(x) is convex in x

// rigorous proof:

// $f(x)_{D(x,x)} \leftrightarrow f(x)_{D_x}$ proof: ($\leftarrow$) trivial ∎

($\rightarrow$) say $f(x)_{D(x,x)} \leftrightarrow \forall \; (x_1,x_1) \in \mathrm{dom} f \times \mathrm{dom} f, \; (x_2,x_2) \in \dots \times \dots$

$$f\left(\theta \begin{bmatrix} x_1 \\ x_1 \end{bmatrix} + (1-\theta)\begin{bmatrix} x_2 \\ x_2 \end{bmatrix}\right) \le \theta f\left(\begin{bmatrix} x_1 \\ x_1 \end{bmatrix}\right) + (1-\theta) f\left(\begin{bmatrix} x_2 \\ x_2 \end{bmatrix}\right)$$

but per absurdum let,

$\exists \; \bar{x}_1 \in \mathrm{dom} f, \; \bar{x}_2 \in \mathrm{dom} f$

$$f(\theta \bar{x}_1 + (1-\theta)\bar{x}_2) > \theta f(\bar{x}_1) + (1-\theta) f(\bar{x}_2) \leftrightarrow f\left(\theta \begin{bmatrix} \bar{x}_1 \\ \bar{x}_1 \end{bmatrix} + (1-\theta)\begin{bmatrix} \bar{x}_2 \\ \bar{x}_2 \end{bmatrix}\right) > \theta f\begin{bmatrix} \bar{x}_1 \\ \bar{x}_1 \end{bmatrix} + (1-\theta) f\begin{bmatrix} \bar{x}_2 \\ \bar{x}_2 \end{bmatrix}$$

but, $f(\bar{x}_1) = f\begin{bmatrix} \bar{x}_1 \\ \bar{x}_1 \end{bmatrix}$, $f(\bar{x}_2) = f\begin{bmatrix} \bar{x}_2 \\ \bar{x}_2 \end{bmatrix}$

contradiction! ∎

Stopping Criterion for Subgradient

Contents of this page:

* A stopping criterion for subgradient:

From equation 1

$$0 \le R^2 + \sum_{i=1}^{k} \alpha_i \|g^{(i)}\|_2^2 - 2\underbrace{\sum_{i=1}^{k} \alpha_i \left(f(x^{(i)}) - f(x^*)\right)}$$

$$-2\sum_{i=1}^{k} \alpha_i f(x^{(i)}) + 2\sum_{i=1}^{k} \alpha_i f^*$$

$$-2\sum_{i=1}^{k}\alpha_i f(x^{(i)}) + 2\sum_{i=1}^{k}\alpha_i f^*$$

$$\rightarrow \quad 0 \leq R^2 + \sum_{i=1}^{k}\alpha_i \|g^{(i)}\|_2^2 - 2\sum_{i=1}^{k}\alpha_i f(x^{(i)}) + 2f^*\sum_{i=1}^{k}\alpha_i$$

$$\rightarrow \quad -R^2 - \sum_{i=1}^{k}\alpha_i \|g^{(i)}\|_2^2 + 2\sum_{i=1}^{k}\alpha_i f(x^{(i)}) \leq 2f^*\sum_{i=1}^{k}\alpha_i$$

$$\rightarrow \quad f^* \geq \frac{2\sum_{i=1}^{k}\alpha_i f(x^{(i)}) - R^2 - \sum_{i=1}^{k}\alpha_i \|g^{(i)}\|_2^2}{2\sum_{i=1}^{k}\alpha_i} = l_k \rightarrow \forall_k \; f^* \geq l_k \rightarrow f^* \geq \max_k \{l_k\} = l_{best}^{(k)} \quad \therefore \; 0 \leq f^* - l_{best}^{(k)}$$

l_k can be computed at each step, in best case we want $\lim_{k\to\infty} l_k = f_k$ and l_k increase as iteration progresses. but as $f(x^{(i)})$ itself fluctuates and $\|g^{(i)}\|$ can fluctuate, so l_k may not be increasing, so we keep track of best lower bound found so far:

$$l_{best}^{(k)} = \max\{l_1, \ldots, l_k\} \quad \text{this is increasing}$$

Stopping criterion:

some small number

$$0 \leq f_{best}^{(k)} - l_{best}^{(k)} \leq \varepsilon$$

* Speeding up subgradient:

- Heavy ball method:

$$x^{(k+1)} = \underbrace{x^{(k)} - \alpha_k g^{(k)}}_{\text{basic}} + \underbrace{\beta_k (x^{(k)} - x^{(k-1)})}_{\text{memory of past steps}}$$

- Filtered subgradient:

hybrid subgradient at (k-1) th step

$$\tilde{g}^{(k)} = (1-\beta)g^{(k)} + \beta\tilde{g}^{(k-1)}, \quad \beta \in [0,1)$$

\# So, a convex combination of hybrid subgradient

\# and subgradient is taken at each step

hybrid subgradient at kth step

$g^{(k)} \in \partial f(x^{(k)})$

- An Italian subgradient:

$$\tilde{g}^{(k)} = g^{(k)} + \beta_k \tilde{g}^{(k-1)}$$

$$\beta_k = \max\{0, -\gamma_k \tilde{g}^{(k-1)T} g^{(k)} / \|\tilde{g}^{(k-1)}\|_2^2\}$$

$\gamma_k \in [0,2)$