Lühiülevaade optimiseerimismeetoditest

Konstantin Tretjakov

3. märts 2004. a.

1 Gradient ja Hessiaan

Definitsioon Olgu $f : \mathbb{R}^n \to \mathbb{R}^m$. Funktsiooni f nimetatakse diferentseeruvaks punktis $\mathbf{x}_0$, kui leidub lineaarteisendus $\mathbf{A}(\mathbf{x}_0)$ selline, et

$$f(\mathbf{x}_0 + \mathbf{\Delta x}) = f(\mathbf{x}_0) + \mathbf{A}(\mathbf{x}_0)\mathbf{\Delta x} + o(\mathbf{\Delta x})$$

Funktsiooni f nimetatakse diferentseeruvaks hulgal $Q \subset \mathbb{R}^n$, kui ta on diferentseeruv igas hulga Q punktis. Kui f on diferentseeruv hulgal $\mathbb{R}^n$, siis ütleme lihtsalt, et f on diferentseeruv. Maatriksit $\mathbf{A}(\mathbf{x}_0)$ nimetatakse funktsiooni f *tuletiseks või Jakobi maatriksiks* (punktis $\mathbf{x}_0$) ning tähistatakse $\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{x}}$ või $f'(\mathbf{x}_0)$.

Teoreem 1.1

$$(\mathbf{A}(\mathbf{x}_0))_{ij} = \frac{\partial f_i(\mathbf{x}_0)}{\partial x_j}$$

Teoreem 1.2

$$(f(g(\mathbf{x}))' = f'(g(\mathbf{x}))g'(\mathbf{x})$$

Olgu $f : \mathbb{R}^n \to \mathbb{R}$ diferentseeruv funktsioon. Tema tuletis (punktis $\mathbf{x}_0$) on siis $1 \times n$ maatriks

$$\mathbf{A}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f(\mathbf{x}_0)}{\partial x_1} & \frac{\partial f(\mathbf{x}_0)}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x}_0)}{\partial x_n} \end{pmatrix}$$

Vektorit $\mathbf{A}(\mathbf{x}_0)^T$ nimetatakse funktsiooni f *gradiendiks* ja tähistatakse $\nabla f(\mathbf{x}_0)$.

Teoreem 1.3

$$\begin{aligned} \nabla(f(\mathbf{x})g(\mathbf{x})) &= \nabla f(\mathbf{x})g(\mathbf{x}) + f(\mathbf{x})\nabla g(\mathbf{x}) \\ \nabla(f(\mathbf{x})/g(\mathbf{x})) &= (\nabla f(\mathbf{x})g(\mathbf{x}) - f(\mathbf{x})\nabla g(\mathbf{x}))/g^2(\mathbf{x}) \\ \nabla(f(g(\mathbf{x}))) &= \nabla f(g(\mathbf{x}))\nabla g(\mathbf{x}) = f'(g(\mathbf{x}))\nabla g(\mathbf{x}) \end{aligned}$$

Olgu nüüd funktsiooni f argument — $m \times n$ maatriks:

$$f = f(\mathbf{X}) = f(x_{11}, x_{12}, \dots, x_{ij}, \dots, x_{mn})$$

Selle funktsiooni gradient on mn elementidega vektor, mida on mõnikord mugav ka $m \times n$ maatriksina kirja panna. Defineerime seega ka *gradiendi maatriksi järgi*:

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \dots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \dots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix}$$

Näited

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^T$$
$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A} + \mathbf{A}^T$$
$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}^T$$
$$\frac{\partial}{\partial \mathbf{X}} \det(\mathbf{X}) = (\mathbf{X}^T)^{-1} \det(\mathbf{X})$$

Definitsioon Olgu $f : \mathbb{R}^n \to \mathbb{R}$ diferentseeruv ja $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ diferentseeruv (s.t. f on kaks korda diferentseeruv). Funktsiooni ∇f tuletis punktis $\mathbf{x}_0$ on $n \times n$ maatriks $\mathbf{H}(\mathbf{x}_0)$, mida nimetatakse funktsiooni f *teiseks tuletiseks* ehk *Hessiaaniks* (punktis $\mathbf{x}_0$) ja tähistatakse $\frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 \mathbf{x}}$ või $\nabla^2 f(\mathbf{x}_0)$.

Teoreem 1.4

$$(\mathbf{H}(\mathbf{x}_0))_{ij} = \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j}$$

Teoreem 1.5 *Kui funktsiooni f osatuletised $\frac{\partial^2 f}{\partial x_i \partial x_j}$, $\frac{\partial^2 f}{\partial x_j \partial x_i}$ on pidevad punktis $\mathbf{x}_0$, siis nad on võrdsed selles punktis.*

Seega piisavalt sileda funktsiooni Hessiaan on sümmeetriline maatriks.

Teoreem 1.6 *Sümmeetrilise maatriksi kõik omaväärtused on reaalarvud.*

Teoreem 1.7 *Sümmeetrilise maatriksi omavektoritest saab koostada ortonormeeritud baasi.*

Teoreem 1.8

$$f(\mathbf{x}_0 + \mathbf{\Delta x}) = f(\mathbf{x}_0) + \frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{x}}\mathbf{\Delta x} + \frac{1}{2}\mathbf{\Delta x}^T \left(\frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 \mathbf{x}}\right) \mathbf{\Delta x} + o(\|\mathbf{\Delta x}\|^2)$$

2 Elementaarsed optimiseerimismeetodid

2.1 Lokaalse miinimumi mõiste

Tähistagu f funktsiooni $\mathbb{R}^n \to \mathbb{R}$.

Definitsioon Punkt $\mathbf{x}^*$ on funktsiooni f *lokaalne miinimum* (või lihtsalt miinimum), kui leidub selle punkti ümbrus $U(\mathbf{x}^*)$ niisugune, et $f(\mathbf{x}) \geq f(\mathbf{x}^*) \quad \forall \mathbf{x} \in U(\mathbf{x}^*)$.

Teoreem 2.1 (Fermat) *Olgu $\mathbf{x}^*$ funktsiooni f miinimumi punkt ning f on diferentseeruv selles punktis. Siis $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Tõestus Olgu $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Siis

$$\begin{aligned} f(\mathbf{x}^* - \eta\nabla f(\mathbf{x}^*)) &= f(\mathbf{x}^*) - \eta\|\nabla f(\mathbf{x}^*)\|^2 + o(\|\eta\nabla f(\mathbf{x}^*)\|) \\ &= f(\mathbf{x}^*) - \eta(\|\nabla f(\mathbf{x}^*)\|^2 - \eta^{-1}o(\eta)) < f(\mathbf{x}^*) \end{aligned}$$

piisavalt väikese η korral. See on vastuolus sellega et $\mathbf{x}^*$ on lokaalne miinimum. ∎

Üldiselt vastupidine väide ei kehti.

Teoreem 2.2 *Kui f on kumer funktsioon, s.t.*

$$\forall\ \mathbf{x}, \mathbf{y} \quad f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$$

siis tingimus $\nabla f(\mathbf{x}^*) = \mathbf{0}$ on samaväärne sellega, et $\mathbf{x}^*$ on funktsiooni f globaalne miinimum.

Teoreem 2.3 *Olgu $\mathbf{x}^*$ funktsiooni f miinimumi punkt, ning f on kaks korda diferentseeruv punktis $\mathbf{x}^*$. Siis $\nabla^2 f(\mathbf{x}^*) \geq 0$ (s.t. funktsiooni Hessiaan on mittenegatiivselt määratud selles punktis).*

Teoreem 2.4 *Kui f on kaks korda diferentseeruv punktis $\mathbf{x}^*$, $\nabla f(\mathbf{x}^*) = \mathbf{0}$ ning $\nabla^2 f(\mathbf{x}^*) > 0$, siis $\mathbf{x}^*$ on funktsiooni f miinimum.*

2.2 Kiirema languse meetod

Eeltoodud Fermat' teoreemi tõestus annab idee kuidas võib funktsiooni miinimumi otsida: alustame suvalisest punktist $\mathbf{x}_0$, ning teeme väikseid samme gradiendi vastasel suunal otsides koha, kus funktsiooni gradient on $\mathbf{0}$. Niimoodi konstrueerime lähendite jada

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k\nabla f(\mathbf{x}_k)$$

Seda meetodit nimetataksegi kiirema languse meetodiks *(steepest descent, gradient descent)*. Tasub panna tähele, et juhul kui iga k korral $\eta_k = \eta$, siis kiirema languse meetod on *hariliku iteratsiooni meetodi* erijuht võrrandi

$$\mathbf{x} = \mathbf{x} - \eta\nabla f(\mathbf{x})$$

lahendamiseks.

Teoreem 2.5 *Olgu f diferentseeruv, ∇f rahuldab Lipschitz'i tingimust*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

f on alt tõkestatud, iga k korral $\eta_k = \eta$ ning $0 < \eta < 2/L$. Siis kiirema languse meetodis gradient läheneb nullile:

$$\lim_{k\to\infty} \nabla f(\mathbf{x}_k) = 0$$

ja jada $f(\mathbf{x}_k)$ monotoonselt kahaneb.

Teoreem 2.6 *Olgu f kaks korda diferentseeruv, iga $\mathbf{x}$ korral*

$$l\mathbf{I} \leq \nabla^2 f(\mathbf{x}) \leq L\mathbf{I}, \quad l > 0$$

ning $0 < \eta < 2/L$. Siis kiirema languse meetod koondub unikaalseks lahendiks $\mathbf{x}^*$, kusjuures

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \|\mathbf{x}_0 - \mathbf{x}^*\|q^k, \quad q < 1$$

2.3 Newton'i meetod

Teine tihti kasutatav iteratiivne optimiseerimismeetod on nn. *Newton'i meetod*. Olgu f kaks korda diferentseeruv. Olgu $\mathbf{x}_k$ mingi lähend. Arendame teda Taylor'i ritta (Teoreem 1.8):

$$f(\mathbf{x}) \approx f_k(\mathbf{x}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \left(\nabla^2 f(\mathbf{x}_k)\right)(\mathbf{x} - \mathbf{x}_k)$$

Kui $\nabla^2 f(\mathbf{x}_k) > 0$, siis funktsioonil f_k on ainus globaalne miinimum punktis $\mathbf{x}_k^*$, mida saab otseselt arvutada. Selle punkti võtame järgmiseks lähendiks: $\mathbf{x}_{k+1} = \mathbf{x}_k^*$, ja kordame algoritmi kuni saame rahuldava tulemuse.

$$\mathbf{x}_{k+1} = \mathbf{x}_k^* = \mathbf{x}_k - \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

Tasub panna tähele, et see algoritm on "tavalise" (s.t. võrrandi $f(x) = 0$ lahendamiseks) Newton'i meetodi erijuht võrrandi $\nabla f(\mathbf{x}) = 0$ lahendamiseks.

Teoreem 2.7 *Olgu $f(\mathbf{x})$ kaks korda diferentseeruv, $\nabla^2 f(\mathbf{x})$ rahuldab Lipschitz'i tingimust konstandiga L, f on tugevalt kumer konstandiga l:*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - l\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2/2$$

ning alglähend $\mathbf{x}_0$ rahuldab

$$q = (Ll^{-2})\|\nabla f(\mathbf{x}_0)\| < 1$$

Siis Newton'i meetod koondub globaalseks miinimumiks $\mathbf{x}^*$ ruut kiirusega:

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq (2l/L) q^{2^k}$$

3 Kvasi-newton'i meetodid, naturaalne gradient

3.1 Kvasi-newton'i meetodid

Kvasi-newton'i optimiseerimismeetodid kombineerivad gradient-meetodi eelised (lihtsad arvutused) koos Newton'i meetodi omadega (kiir koonduvus). Neid on pakutud mitu ja kõik omavad üldist struktuuri:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{H}_k \nabla f(\mathbf{x}_k)$$

kus maatriks $\mathbf{H}_k$ arvutatakse rekurrentselt igal sammul kasutades eelmisel sammul saadud informatsiooni, nii et $(\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k)^{-1}) \to 0$. Piiril need meetodid lähenevad seega Newton'i meetodile.

Teoreem 3.1 *Olgu f alt tõkestatud, diferentseeruv, ∇f rahuldab Lipschitz'i tingimust ning*

$$m\mathbf{I} \leq \mathbf{H}_k \leq M\mathbf{I}, \quad m > 0.$$

Siis kvasi-newton'i meetodis kus $\eta_k = \eta > 0$ on piisavalt väike, $\nabla f(\mathbf{x}_k) \to 0$.

Teoreem 3.2 *Olgu $\mathbf{x}^*$ funktsiooni f miinimum. f on kaks korda diferentseeruv $\mathbf{x}^*$ ümbruses, $\nabla^2 f(\mathbf{x}^*) > 0$ ning*

$$\|\mathbf{H}_k - \nabla^2 f(\mathbf{x}^*)^{-1}\| \to 0$$

Siis kvasi-newton'i meetod, kus $\eta_k = 1$, lokaalselt koondub punktiks $\mathbf{x}^*$ kiiremini igast geomeetrilisest progressioonist.

Näiteks järgmine on maatriksi $\mathbf{H}_k$ arvutamise reegel *Davidson-Fletcher-Powell*'i meetodis:

$$\mathbf{H}_{k+1} = \mathbf{H}_k - \frac{\mathbf{H}_k \mathbf{y}_k \mathbf{y}_k^T \mathbf{H}_k}{\mathbf{y}_k^T \mathbf{H}_k \mathbf{y}_k} + \eta_k \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{y}_k}, \quad \mathbf{H}_0 > 0$$

kus $\mathbf{p}_k = -\mathbf{H}_k \nabla f(\mathbf{x}_k)$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ (?).

3.2 Naturaalne gradient (Natural Gradient)

Teine mõtekäik kvasi-newton meetoditeni jõudmiseks. Tavaline kiirema languse meetod sisuliselt leiab etteantud punktist fikseeritud kaugusel oleva punkti, mis minimiseeriks funktsiooni f lineariseeritud versiooni, s.t. seda meetodit saab kirja panna niimoodi:

$$\mathbf{x}_{k+1} = \underset{\|\mathbf{x}-\mathbf{x}_k\| \leq \varepsilon_k}{\operatorname{argmin}} \left(f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T (\mathbf{x} - \mathbf{x}_k) \right)$$

Mõnikord osutub kasulikum kasutada kauguse mõõduks mitte tavalist Eukleidilist kaugust, vaid nn. Riemann'i kaugust. Riemann'i geomeetria korral defineeritakse kaugust punktide $\mathbf{x}$ ja $\mathbf{x} + \mathbf{\Delta x}$ vahel kui:

$$d(\mathbf{x}, \mathbf{x} + \mathbf{\Delta x}) = \sqrt{\mathbf{\Delta x}^T \mathbf{G}(\mathbf{x}) \mathbf{\Delta x}}$$

kus $\mathbf{G}(\mathbf{x})$ on $n \times n$ positiivselt määratud maatriks *(Riemannian metric tensor)* ja $\mathbf{\Delta x}$ on väike. Osutub et sellises ruumis avaldub funktsiooni gradient kujul

$$\nabla_{\mathbf{G}} f(\mathbf{x}) = \mathbf{G}(\mathbf{x})^{-1} \nabla f(\mathbf{x})$$

Seega kiirema languse meetodi reegel on

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{G}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x})$$

4 Kiirema languse meetodi edasiarendused

4.1 "Raske kuuli" meetod

Lisame kiirema languse meetodile "inertsi", s.t. arvestame igal sammul ka eelmist sammu:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k) + \alpha_k (\mathbf{x}_k - \mathbf{x}_{k-1})$$

Kus $0 \leq \alpha_k < 1$ on nn. "inertsitegur" *(momentum term)*.

Teoreem 4.1 *Olgu $\mathbf{x}^*$ funktsiooni f miinimum, $l\mathbf{I} \leq \nabla^2 f(\mathbf{x}^*) \leq L\mathbf{I}$. Kui*

$$\alpha_k = \alpha, \quad \eta_k = \eta, \quad 0 \leq \alpha < 1, \quad 0 < \eta < 2(1+\alpha)/L$$

siis leidub $\varepsilon > 0$ selline, et iga $\mathbf{x}_0$, $\mathbf{x}_1$ jaoks punkti $\mathbf{x}^*$ ε-ümbrusest, koondub raske kuuli meetod miinimumiks geomeetrilise progressiooni kiirusega.

4.2 Kaasgradientide meetod (Conjugate Gradient)

Kaasgradientide meetod on "raske kuuli" meetod, kus igal sammul leiakse optimaalsed η ja α:

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k) + \alpha_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \\ \{\eta_k, \alpha_k\} &= \underset{\{\eta,\alpha\}}{\operatorname{argmin}} f(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k) + \alpha(\mathbf{x}_k - \mathbf{x}_{k-1}))\end{aligned}$$

Ruutfunktsiooni korral

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \quad \mathbf{A} > 0$$

saab leida lahendid:

$$\begin{aligned}\eta_k &= \frac{\|\mathbf{r}_k\|^2(\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k) - (\mathbf{r}_k^T\mathbf{p}_k)(\mathbf{p}_k^T\mathbf{A}\mathbf{r}_k)}{(\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k)(\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k) - (\mathbf{p}_k^T\mathbf{A}\mathbf{r}_k)^2} \\ \alpha_k &= \frac{\|\mathbf{r}_k\|^2(\mathbf{p}_k^T\mathbf{A}\mathbf{r}_k) - (\mathbf{r}_k^T\mathbf{p}_k)(\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k)}{(\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k)(\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k) - (\mathbf{p}_k\mathbf{A}\mathbf{r}_k)^2} \\ \mathbf{r}_k &= \nabla f(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{b}, \quad \mathbf{p}_k = \mathbf{x}_k - \mathbf{x}_{k-1}\end{aligned}$$

Teoreem 4.2 *Gradiendid $\mathbf{r}_k$ kaasgradientide meetodis on paarikaupa ortogonaalsed.*

Teoreem 4.3 *Ruutfunktsiooni korral koondub kaasgradientide meetod lahendiks ülimalt n sammuga.*

5 Stohhastiline kiirema languse meetod

Praktikas meil on tavaliselt vaja minimiseerida mingit hinnangu (riski) funktsiooni kujul $J(\mathbf{w}) = E\{g(\mathbf{w}, \mathbf{x})\}$. Siin $\mathbf{w}$ on mingi *kaalude vektor* mille suhtes me minimiseerime, $\mathbf{x}$ — juhuslik vektor tihedusfunktsiooniga f, E on keskväärtus $\mathbf{x}$ jaotuse järgi. Tihedusfunktsiooni f tavaliselt teada ei ole, on aga *näidete* komplekt $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_m$.
(Näide: selleks et leida $\mathbf{x}$ keskväärtust, peame me leidma $\mathbf{w}$, mis minimiseeriks funktsiooni $J(\mathbf{w}) = E\{\|\mathbf{w} - \mathbf{x}\|^2\}$).
Sel juhul kiirema languse meetod näeb välja niimoodi:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}} E\{g(\mathbf{w}_k, \mathbf{x})\}$$

kus

$$\nabla_{\mathbf{w}} E\{g(\mathbf{w}_k, \mathbf{x})\} = \nabla_{\mathbf{w}} \int g(\mathbf{w}_k, \mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int \nabla_{\mathbf{w}} g(\mathbf{w}_k, \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

kus viimane operatsioon on lubatud juhul kui g on kaks korda diferentseeruv $\mathbf{w}$ järgi. Praktikas me kasutame teoreetilise keskväärtuse arvutamise asemel näidete keskmist:

$$\nabla_{\mathbf{w}} E\{g(\mathbf{w}_k, \mathbf{x})\} \approx \frac{1}{m} \sum_{i=1}^{m} \nabla_{\mathbf{w}} g(\mathbf{w}_k, \mathbf{x}_i)$$

Sellist optimiseerimisreeglit nimetatakse *paketipõhiseks (batch)*. Mõnikord keskväärtuse arvutamine igal iteratsiooni sammul on liiga raske kas selle pärast et näiteid on liiga palju, või selle pärast et näited tulevad pidevalt väliskeskkonnast. Siis kasutatakse nn. *on-line* optimiseerimistehnikat, kus keskväärtust ei arvutata üldse, vaid igal iteratsiooni sammul võetakse mingi juhuslik näide $\mathbf{x}_i$:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}} g(\mathbf{w}_k, \mathbf{x}_i)$$

Osutub, et kui näited valida komplektist juhuslikult, koondub (teatud mõttes) ka selline meetod lahendiks, tõsi küll, koondumine on aeglasem ja juhusliku iseloomuga. Koondumise jaoks peab kehtima

$$\sum_{k=1}^{\infty} \eta_k = \infty, \qquad \sum_{k=1}^{\infty} \eta_k^2 < \infty$$

6 Tingimustega optimiseerimine

6.1 Põhimõisted

Definitsioon Olgu $Q \subset \mathbb{R}^n$. Funktsioonil f on punktis $\mathbf{x}^*$ tinglik (lokaalne) miinimum, kui $f(\mathbf{x}^*) \leq f(\mathbf{x})$ iga $\mathbf{x}$ korral hulgast $U(\mathbf{x}^*) \cap Q$ kus $U(\mathbf{x}^*)$ on punkti $\mathbf{x}^*$ mingi ümbrus.

Teoreem 6.1 *Olgu f diferentseeruv tema miinimumpunktis $\mathbf{x}^*$ ja $Q \subset \mathbb{R}^n$ kumer. Siis iga $\mathbf{x} \in Q$ korral*

$$\nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) \geq 0$$

Teoreem 6.2 *Olgu f diferentseeruv punktis $\mathbf{x}^* \in Q$, Q kumer ja iga $\mathbf{x} \in U(\mathbf{x}^*) \cap Q$ korral*

$$\nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) \geq \alpha \|\mathbf{x} - \mathbf{x}^*\|, \quad \alpha > 0$$

Siis $\mathbf{x}^*$ on funktsiooni f miinimum hulgal Q.

6.2 Gradiendi projektsiooni meetod

See on kiirema languse meetodi vahetu üldistus. Kasutame tavalist gradientmeetodit, aga kui mingil sammul saadud $\mathbf{x}_k$ on väljaspool hulka Q, proetseerime seda sinna tagasi. Seega saame meetodid:

$$\mathbf{x}_{k+1} = P_Q(\mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k))$$

Kus P_Q on projektsioon hulgale Q:

$$P_Q(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y} \in Q} \|\mathbf{x} - \mathbf{y}\|$$

Kinnise Q korral $P_Q(\mathbf{x})$ eksisteerib iga $\mathbf{x}$ korral, ning kumera Q korral ta on unikaalne.

Teoreem 6.3 *Olgu $Q \subset \mathbb{R}^n$ kumer ja kinnine. Olgu f kumer diferentseeruv funktsioon millel leidub hulgal Q miinimum $\mathbf{x}^*$. Rahuldagu ∇f hulgal Q Lipschitzi tingimust konstandiga L ning olgu $0 < \eta < 2/L$. Siis gradiendi projektsiooni meetodis $\mathbf{x}_k \to \mathbf{x}^*$.*

Näiteks selline meetod võiks rakendada siis, kui meil oleks vaja leida funktsiooni $f(\mathbf{x})$ miinimumi tingimusel $\|\mathbf{x}\| \leq 1$. Sel juhul kasutame me lahendi leidmiseks tavalist kiirema languse meetodit, ning kui mingil sammul $\|\mathbf{x}_k\| > 1$, normaliseerime vektorit.

6.3 Lagrange'i kordajate reegel

Olgu meil antud funktsioonid $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ ning me otsime funktsiooni f miinimumi $\mathbf{x}^*$ mis rahuldaks tingimusi $g_i(\mathbf{x}^*) = 0$. S.t. me otsime f miinimumi hulgal $Q = \{\mathbf{x} \mid g_i(\mathbf{x}) = 0,\ i = 1, \ldots, m\}$.

Teoreem 6.4 (Lagrange'i kordajate reegel) *Olgu $\mathbf{x}^*$ funktsiooni f miinimum hulgal Q, funktsioonid f, g_i pidevalt diferentseeruvad $\mathbf{x}^*$ ümbruses ning $\nabla g_i(\mathbf{x}^*)$ on lineaarselt sõltumatud. Siis leiduvad arvud $\lambda_1, \ldots, \lambda_m$, mitte kõik võrdsed nulliga, et*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) = 0$$

Funktsiooni

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

nimetatakse *Lagrange'i funktsiooniks*. Lagrange'i kordajate reeglit saab siis ümbersõnastada kui:

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}, \quad \nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}$$

Viited

[1] A. Hüvarinen, J. Karhunen, E. Oja.
Independent Component Analysis, pp. 57–76,
2001, John Wiley & Sons Inc.

[2] Scott C. Douglas, Shun-ichi Amari. *Natural Gradient Adaptation. Unsupervised Adaptive Filtering, vol. I*, pp. 13–62,
2000, John Wiley & Sons Inc.

[3] Б. Т. Поляк
Введение в оптимизацию, pp. 15–43, 63–93
1983, Москва «Наука»