Icke-linjär optimering

Icke-linjär optimering utan bivillkor handlar om att lösa problemet

$$\min_{x \in \Re^n} f(x),$$

där $f(x) : \Re^n \to \Re$ är en två gånger kontinuerligt deriverbar funktion.

Optimalitet, globala minimerare

Betrakta den n-dimensionella problemet

$$\min_{x \in \Re^n} f(x).$$

Om en punkt x_* uppfyller

$$f(x_*) \leq f(x) \ \forall x \in \Re^n$$

sägs x_* vara en *global minimerare* till f i $\Re^n$. Punkten x_* kallas ofta *lösningspunkt*.

Om

$$f(x_*) < f(x) \ \forall x \in \Re^n, x \neq x_*$$

kallas x_* *strikt global minimerare* och är unik.

Globala minimerare är önskvärda, men svåra att bestämma om inte f har speciella egenskaper.

Optimalitet, lokala minimerare

Många gånger får vi nöja oss med en *lokal minimerare*, dvs ett x_* sådant att

$$f(x_*) \leq f(x) \ \forall x \in \Re^n, \|x - x_*\| < \epsilon, \epsilon \text{ “litet”}.$$

På samma sätt definieras en *strikt lokal minimerare* x_* som

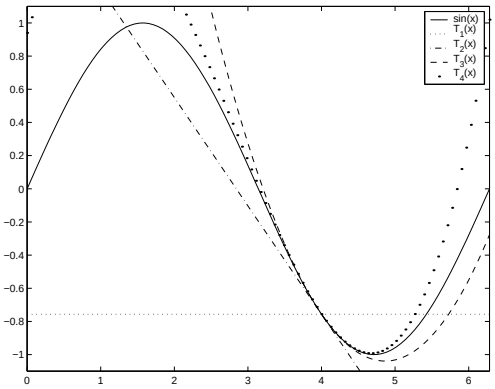
$$f(x_*) < f(x) \ \forall x \in \Re^n, x \neq x_*, \|x - x_*\| < \epsilon, \epsilon \text{ “litet”}.$$

Taylorserier

Taylorserier är ett verktyg för att approximera en funktion f nära en punkt x.

Definition: Låt x vara specificerad punkt och $f : \Re \to \Re$ med n kontinuerliga derivator. Då är taylorapproximationen av grad n:

$$f(x+p) \approx f(x) + pf'(x) + \frac{p^2}{2}f''(x) + \cdots + \frac{p^n}{n!}f^{(n)}(x).$$



Taylorapproximationen av $\sin x$ runt $x = 4$.
$T_1(x) = \sin 4$, $T_2(x) = \sin 4 + \cos 4 \cdot (x-4)$,
$T_3(x) = \sin 4 + \cos 4 \cdot (x-4) - \sin 4\frac{(x-4)^2}{2}$,
$T_4(x) = \sin 4 + \cos 4 \cdot (x-4) - \sin 4\frac{(x-4)^2}{2} - \cos 4\frac{(x-4)^3}{3!}$.

Flerdimensionella taylorserier

Det går också att härleda flerdimensionella taylorserier:

1-variabelfallet:

$f(x_0 + p) = f(x_0) + pf'(x_0) + \frac{1}{2}pf''(x_0)p + \cdots$.

n-variabelfallet:

$f(x_0 + p) = f(x_0) + p^T\nabla f(x_0) + \frac{1}{2}p^T\nabla^2 f(x_0)p + \cdots$.

$\nabla f(x_0)$ betecknar *gradienten* till f i x_0, dvs en kolumnvektor med element $\frac{\partial f}{\partial x_i}(x_0)$ och $\nabla^2 f(x_0)$ betecknar *hessianen* till f i x_0, dvs en matris med element $\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)$.

Villkor på minimerare

Antag x_* lokal minimerare till f. Studera f runt x_*:

$$f(x_* + p) = f(x_*) + \nabla f(x_*)^T p + \frac{1}{2}p^T\nabla^2 f(\xi)p$$

Om x_* lokal minimerare så finns ingen tillåten nedförsriktning, dvs

$$\nabla f(x_*)^T p \geq 0$$

för alla tillåtna riktningar p.

För problem utan bivillkor medför detta att

$$\nabla f(x_*) = 0 \qquad (1)$$

En punkt som uppfyller (1) kallas *stationär punkt* till f.

Villkoret (1) kallas för ett *nödvändigt* villkor av första graden för en optimerare.

Villkor på minimerare, forts.

Studera f runt stationär punkt x_* som antas vara en lokal minimerare:

$$f(x) = f(x_*+p) = f(x_*) + \underbrace{\nabla f(x_*)^T p}_{=0} + \frac{1}{2}p^T\nabla^2 f(\xi)p.$$

För x nära x_* kommer $\nabla^2 f(\xi)$ vara nära $\nabla^2 f(x_*)$.

Om $\nabla^2 f(x_*)$ ej positivt semi-definit finns v sådan att

$$v^T\nabla^2 f(x_*)v < 0.$$

Då finns p nära v sådan att

$$p^T\nabla^2 f(\xi)p < 0$$

Villkor på minimerare, forts.

vilket medför att

$$\begin{aligned} f(x) &= f(x_* + p) = f(x_*) + \underbrace{\frac{1}{2}p^T\nabla^2 f(\xi)p}_{<0} \\ f(x) &< f(x_*) \end{aligned}$$

vilket är en motsägelse, ty x_* antogs vara en minimerare.

Alltså måste $\nabla^2 f(x_*)$ vara positivt semi-definit för att en stationär punkt x_* ska vara en lokal minimerare.

Detta kallas för ett *nödvändigt* minimeringsvillkor av andra graden.

Begränsningar hos villkoren

Informationen att $\nabla f(x_*) = 0$ och $\nabla^2 f(x_*)$ är positivt semi-definit räcker inte till för att avgöra om en punkt är en minimerare.

Exempel:

$f(x)$	$\nabla f(x)$	$\nabla^2 f(x)$	$\nabla^2 f(x_*)$	$x_* = 0$
x^2	$2x$	2	$2 > 0$	min
x^3	$3x^2$	$6x$	$0 \geq 0$	ej min, ej max
x^4	$4x^3$	$12x^2$	$0 \geq 0$	min
$-x^4$	$-4x^3$	$-12x^2$	$0 \geq 0$	max

Villkor på minimerare, forts.

Studera f runt x_* då $\nabla f(x_*) = 0$ och $\nabla^2 f(x_*)$ positivt definit.

Då är

$$f(x) = f(x_*+p) = f(x_*) + \underbrace{\nabla f(x_*)^T p}_{=0} + \frac{1}{2} p^T \nabla^2 f(\xi) p.$$

För x tillräckligt nära x_* kommer $\nabla^2 f(\xi)$ också att vara positivt definit, dvs

$$f(x) = f(x_* + p) = f(x_*) + \underbrace{\frac{1}{2} p^T \nabla^2 f(\xi) p}_{>0 \ \forall p \neq 0}$$

$$f(x) > f(x_*) \quad \forall p$$

Vilket innebär att x_* är en strikt minimerare till f.

Detta är ett *tillräckligt* minimeringsvillkor av andra graden.

Newtons klassiska metod för minimering

För att använda Newtons metod till att finna en minimerare applicerar vi första ordningens nödvändiga villkor på en funktion f:

$$\nabla f(x) = 0 \qquad (f'(x) = 0)$$

Detta ger newtonsekvensen

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \quad (x_{k+1} = x_k - f'(x)/f''(x))$$

Detta skrivs ofta som $x_{k+1} = x_k + p_k$, där p_k är lösningen till Newtons ekvation:

$$\nabla^2 f(x_k) p_k = -\nabla f(x_k),$$

vilket tydliggör att det i varje steg löses ett linjärt ekvationssystem (och inte beräknas en invers).

Newtons klassiska metod, forts.

Notera att approximationen av det icke-linjära funktionen $\nabla f(x)$ med

$$\nabla f(x_k + p) \approx \nabla f(x_k) + \nabla^2 f(x_k) p$$

motsvarar att approximera den icke-linjära funktionen $f(x)$ med den kvadratiska funktionen

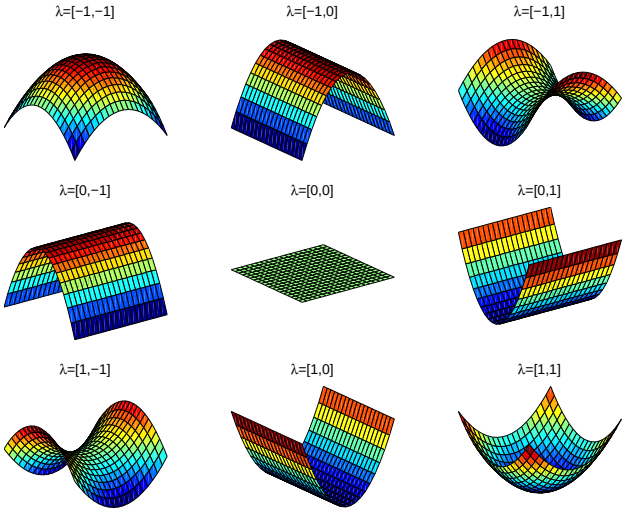
$$Q(p) \equiv f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p,$$

dvs de tre första termerna i taylorutvecklingen av f runt x_k.

En användbar tolkning av Newtons metod blir att vi i varje iterationssteg approximerar f med en kvadratisk funktion Q och beräknar x_{k+1} som minimerare av Q.

Krökningar och egenvärden

Om hessianen $\nabla^2 f(x)$ är positivt definit har den enbart positiva egenvärden och den kvadratiska approximationen Q kröker "uppåt" i alla dimensioner. Pss för andra kombinationer av egenvärden.



Alla kombinationer av egenvärden för $\nabla^2 f(x)$ och motsvarande approximationsytor.

Egenskaper hos Newtons metod

Fördel:

- Den konvergerar typiskt kvadratiskt mot en stationär punkt.

Nackdelar:

- Den konvergerar ej nödvändigtvis mot en min-punkt.
- Den kan divergera om startpunkten ligger för långt från lösningen.
- Den kan misslyckas om $\nabla^2 f(x_k)$ ej inverterbar för något k.
- Den behöver andraderivateinformation i $\nabla^2 f(x_k)$.
  - ”Jobbigt” ta fram funktionerna. Kan bli fel.
  - Kräver beräkning och lagring av n^2 element.
  - Lösningen av Newtons ekvationer kräver n^3 operationer.

I praktiken används Newtons metod i sin klassiska formulering aldrig. Däremot är den grunden för många andra algoritmer som söker minska dess negativa egenskaper på olika sätt och samtidigt behålla metodens positiva egenskaper.

Garantera nedåtriktning

Enligt vår generella optimeringsalgoritm bestäms den nya approximationen på formen

$$x_{k+1} = x_k + \alpha p, \quad \alpha > 0, \quad f(x_{k+1}) < f(x_k).$$

Detta är möjligt endast om p är en nedförsriktning, dvs

$$p^T \nabla f(x_k) < 0.$$

I Newtons metod bestäms sökriktningen som

$$p = -\nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

Om p ska vara en nedåtriktning i punkten x_k måste

$$p^T \nabla f(x_k) = -\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) < 0$$

eller

$$\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) > 0$$

Detta uppfylls om $\nabla^2 f(x_k)^{-1}$ är positivt definit.

Garantera minskning av objektfunktionens värde

Att p är en nedåtriktning innebär att det finns ett $\alpha > 0$ sådant att $f(x_k + \alpha p) < f(x_k)$. För att finna ett sådant α används *linjesökning*.

Låt x_k vara nuvarande approximation av en minimerare till f och låt p_k vara sökriktningen i x_k. Den nya uppskattningen definieras som

$$x_{k+1} = x_k + \alpha_k p_k,$$

där steglängden α_k väljs så att

$$f(x_{k+1}) < f(x_k),$$

dvs varje steg minskar funktionsvärdet och tar oss närmare lösningen.

Global konvergens

För att garantera *global konvergens*, dvs konvergens mot ett lokalt minimum från varje startpunkt ställs två ytterligare krav på sökriktningen p och två krav på steglängden α.

- Sökriktningen p får ej vara godtyckligt nära ortogonal med negativa gradienten $-\nabla f(x_k)$.
- Sökriktningen p får ej vara godtyckligt kort relativt gradienten.

Dessa två uppfylls av Newtons metod om $\nabla^2 f(x_k)$ positivt definit och ej godtyckligt nära semi-definit.

- Steglängden α får ej producera en godtyckligt liten minskning av objektfunktionen.
- Steglängden α får ej vara godtyckligt kort.

Armijos linjesökning med backtracking

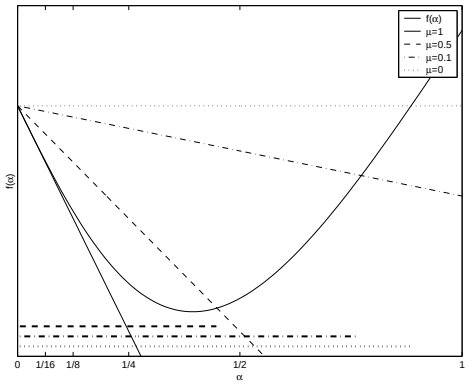
För att uppfylla kravet på tillräcklig minskning av objektfunktionen utgår vi från en linjär taylorapproximation av objektfunktionen

$$f(x_k + \alpha p_k) \approx f(x_k) + \alpha p_k^T \nabla f(x_k).$$

och kräver att minskningen av objektfunktionen är en fraktion av den som predikteras av den linjära taylorapproximationen, dvs

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \mu \alpha_k p_k^T \nabla f(x_k),$$

där $0 < \mu < 1$. Detta kallas ibland för *Armijos villkor*.



För att undvika att steglängden blir för kort accepteras steglängden α_k som det första elementet i sekvensen $1, \frac{1}{2}, \frac{1}{4}, \ldots, 2^{-i}$ som uppfyller Armijos villkor. Detta kallas *backtracking*.

Icke-linjära minsta-kvadratproblem

Ett icke-linjärt minsta-kvadrat-problem är ett optimeringsproblem utan bivillkor på formen

$$\min_x f(x) = \sum_{i=1}^{m} f_i(x)^2,$$

där objektfunktionen definieras via externa funktioner $\{f_i(x)\}$.

Problemet kallas ”minsta-kvadrat” därför att vi minimerar summan av kvadraten på funktionerna.

Icke-linjär minsta-kvadrat parameterestimering

I ett parameterestimeringsproblem motsvarar funktionerna $f_i(x)$ skillnaden (residualen) mellan en modellfunktion och ett mätvärde. Studera exempelvis datamängden

t_i	:	1	2	4	5	8
y_i	:	3	4	6	11	20

där t_i är tiden i år och y_i hundratal (antiloper). Om vi antar att dessa mätvärden följer ett exponentialförlopp skulle basfunktionen t ex bli $g(t) = x_1 e^{x_2 t}$ och residualerna

$$f_i(x) = g(t_i) - y_i = x_1 e^{x_2 t_i} - y_i.$$

Icke-linjär minsta-kvadrat-parameterestimering, forts.

Vi kommer att skriva problemet som

$$\min_x f(x) = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2 \equiv \frac{1}{2}F(x)^T F(x),$$

där F är en vektorvärd funktion

$$F(x) = [f_1(x)\ f_2(x)\ \ldots\ f_m(x)]^T.$$

Gradienten $\nabla f(x)$ går att härleda m h a kedjeregeln

$$\nabla f(x) \ = \ \nabla F(x)F(x) = J(x)^T F(x)$$

där $J(x) = \nabla F(x)^T$ är *jacobianen* till $F(x)$.

Detsamma gäller hessianen

$$\begin{array}{rcccc} \nabla^2 f(x) & = & \nabla F(x)\nabla F(x)^T & + & \sum_{i=1}^m f_i(x)\nabla^2 f_i(x) \\ & = & J(x)^T J(x) & + & Q(x). \end{array}$$

Icke-linjär minsta-kvadrat-parameterestimering, forts.

Notera att gradienten $\nabla f(x_*) = 0$ i ett av två fall:

- Då $F(x_*) = 0 \Rightarrow f(x_*) = 0$, vilket kallas att problemet har *nollresidual*. Då är dessutom hessianen
$$\nabla^2 f(x_*) = J(x_*)^T J(x_*)$$
positivt semi-definit. Om $J(x_*)$ har full rang är dessutom $\nabla^2 f(x_*)$ positivt definit. Skulle $J(x_*)$ vara rangdefekt sägs problemet vara *överparameteriserat*.

- Då $J(x)^T F(x) = 0$, dvs residualvektorn $F(x)$ är ortogonal mot gradientplanet som spänns upp av $J(x)$.

Icke-linjär minsta-kvadrat-parameterestimering, forts.

Notera att hessianen

$$\begin{array}{rcccc} \nabla^2 f(x) & = & \nabla F(x)\nabla F(x)^T & + & \sum_{i=1}^m f_i(x)\nabla^2 f_i(x) \\ & = & J(x)^T J(x) & + & Q(x). \end{array}$$

är en summa av två komponeneter, en med förstaderivateinformation och en med andraderivateinformation.

Om problemet har nollresidual kommer termen $Q(x)$ att vara nära noll nära lösningen. En metod som använder approximationen $Q(x) = 0$ kallas *Gauss-Newtons* metod och bestämmer sökriktningen som lösningen till Newtons ekvation

$$\nabla^2 f(x)p = -\nabla f(x)$$

med hessianen approximerad med $J(x)^T J(x)$:

$$J(x)^T J(x)p = -J(x)^T F(x).$$

Annan härledning av Gauss-Newtons metod

Gauss-Newtons metod går att härleda på ett sätt till:

Om vi studerar det (typiskt överbestämda) linjära problemet

$$\min_p \|J(x)p + F(x)\|^2 = \|J(x)p - (-F(x))\|^2$$

så blir lösningen p lika med lösningen till normalekvationerna

$$p = -(J(x)^T J(x))^{-1} J(x)^T F(x),$$

vilket är sökriktningen i Gauss-Newtons metod. Denna beräknas naturligtvis med hjälp av lämplig faktorisering (QR eller SVD) *utan* att matrisen $J(x)^T J(x)$ bildas.

Konvergens för Gauss-Newtons metod

Om $F(x_*) = 0$ är approximationen $Q(x) \approx 0$ bra, och Gauss-Newtons metod kommer att uppföra sig som Newtons metod nära lösningen, dvs konvergera snabbt om $J(x_*)$ har full rang. Detta utan kostnaden att beräkna andraderivatorna $\nabla^2 f_i(x)$.

Om $F(x_*)$ är ”stor” och/eller krökningarna $\nabla^2 f_i(x)$ är stora, blir approximationen $Q(x) \approx 0$ dålig, och Gauss-Newtons metod kommer att konvergera långsammare än Newtons metod.

Statistisk tolkning av residualerna

Om residualerna tolkas statistiskt som ”fel”, dvs vi har en modell

$$y_i = x_1 e^{x_2 t_i} + \varepsilon_i$$

och felen ε_i antas vara oberoende normalfördelade $\in N(0, \sigma^2)$ blir våra uppskattade parameterar de bästa tänkbara (”maximum likelihood”) givet våra mätvärden. Detta gör det möjligt att säga saker om konfidensintervall, göra hypotesprövning, etc.

Den informationen fås ur varians-kovariansmatrisen

$$D = \sigma^2 (\nabla^2 f(x_*))^{-1},$$

där diagonalelementen D_{ii} motsvarar variansen hos x_i och D_{ij} motsvarar kovariansen mellan x_i och x_j.

Ofta används approximationen $Q(x_*) = 0$, vilket gör att konfidensintervall m.m. också blir approximativa.

Ortogonal regression

När vi löser

$$\min_x \sum_{i=1}^{m} f_i(x)^2, \qquad (2)$$

där $f_i(x) = g(t_i) - y_i$ är skillnaden mellan vår modell och våra mätvärden, så minimerar vi kvadraten på det *vertikala* avståndet. I andra sammanhang, t ex om vi antar att vi har fel även i den oberoende variabeln t_i, kan det vara rimligt att i stället minimera det *ortogonala* avståndet mellan vår modellfunktion och våra mätvärden.

Ortogonal regression, forts.

Det kan beskrivas som att vi löser

$$\min_{x,\delta} f(x) = \sum_{i=1}^{m} f_i(x; t_i + \delta_i)^2 + \|\delta\|^2, \qquad (3)$$

där δ_i är felet i t_i och $f_i(x; t_i + \delta_i) = g_i(t_i + \delta_i) - y_i$.

Detta kräver dock att vi skriver om våra optimeringsalgoritmer för att lösa problem (3) i stället för problem (2).

Ortogonal regression, forts.

Det är dock möjligt att använda våra algoritmer för problem (2) om vi i stället utökar om våra delfunktioner f_i:

För vårt testexempel

$$y = g(t) = x_1 e^{x_2 t}$$

introducerar vi en punkt $(s_i, g(s_i))$ på kurvan för varje mätpunkt (t_i, y_i).

Om vi låter vår delfunktion f_i bli

$$f_i = \begin{bmatrix} g(t_i) - y_i \\ s_i - t_i \end{bmatrix},$$

och formulerar om vårt minimeringsproblem

$$\begin{aligned} \min_x f(x) &= \sum_{i=1}^{m} f_i(x)^2 \\ &\Downarrow \\ \min_x f(x) &= \sum_{i=1}^{m} f_i^T(x) f_i(x) \end{aligned}$$

så får vi samma lösning.