بهینه سازی
مبانی بهینه سازی نامقید
روش گرادیان نزولی، روش نیوتن

محسن هوشمند
دانشکده تکنولوژی اطلاعات و علم رایانه
دانشگاه تحصیلات تکمیلی علوم پایه زنجان

مقدمه

یافتن کمینهٔ تابع هدف

مبتنی بر متغیرهای حقیقی بدون قید

از تمامی فضای اعداد حقیقی

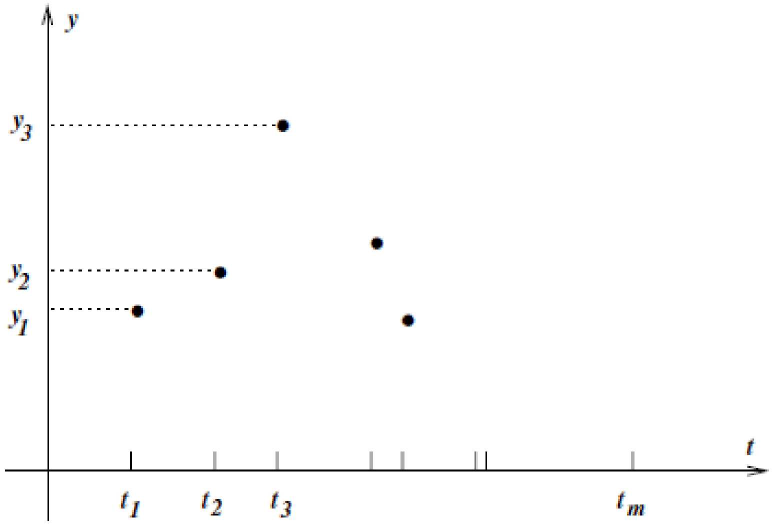
$$\underset{\boldsymbol{x}}{\text{کم}}\, f(\boldsymbol{x})$$

$\boldsymbol{x} \in \mathbb{R}^n$

$n \geq 1$

$f: \mathbb{R}^n \rightarrow \mathbb{R}$ تابعی هموار

مقدمه – مثال



یافتن منحنی که چند داده را تقریب بزند

نمایش نمایی و نوسانی $\phi(t;x) = x_1 + x_2 e^{-(x_3-t)^2/x_4} + x_5 \cos(x_6 t)$

x_i-ها ضرائب مدل

- مجهول‌ها $x = (x_1, x_2, \ldots, x_6)^T$

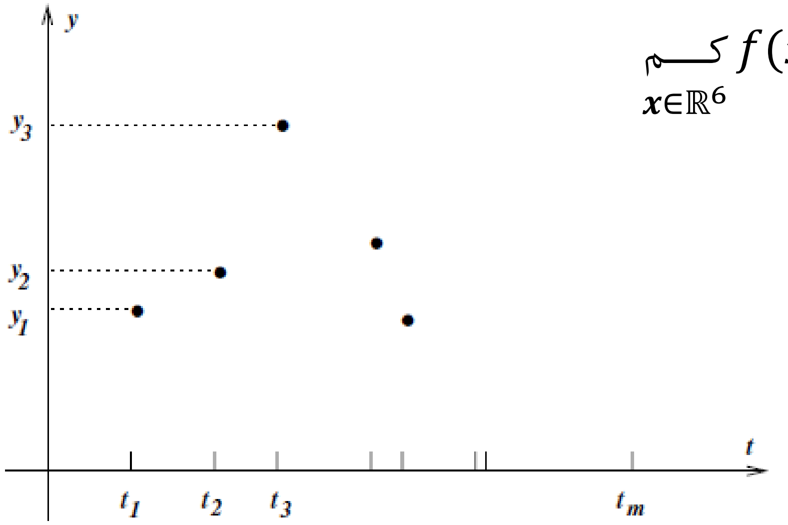
تعریف باقی‌مانده $r_j(x) = y_j - \phi(t_j;x), \qquad j = 1, 2, \ldots, m$

- نمایشگر تفاوت بین مدل و داده مشاهده شده

$$\underset{\boldsymbol{x}\in\mathbb{R}^6}{\text{کمـــم}}\; f(x) = r_1^2(x) + r_2^2(x) + \cdots + r_m^2(x)$$

- مسئلهٔ برازش کمترین مربعات غیرخطی

مقدمه – مثال - ادامه



$$\underset{x\in\mathbb{R}^6}{\text{کم}} f(x) = r_1^2(x) + r_2^2(x) + \cdots + r_m^2(x)$$

- مسئلهٔ برازش کمترین مربعات غیرخطی
  - از انواع بهینه‌سازی نامقید

فرض: پاسخ $x^* = (1.1, 0.01, 1.2, 1.5, 2.0, 1.5)$

مقدار کمینه $f(x^*) = 0.34$

وجود تفاوت بین مقدار واقعی و مقدار بدست‌آمده

- چگونه **x*** را کمینه‌ساز **f** می‌دانیم
  - نیاز به تعریف «راه‌حل»

تشخیص کمینه محلی

کمینه‌ساز سراسری f

- نقطهٔ بدست‌دهندهٔ کمترین مقدار تابع $f(x^*) \le f(x)\ \forall x$
- مشکل در یافتن
- اطلاع صرفا از دور و بر هر نقطه و نه بیشتر
- عدم‌اطمینان از وجود شکاف عمیق در داده
- بیشتر الگوریتم‌ها یابندهٔ کمینه‌ساز محلی

کمینه‌ساز محلی x^*

- اگر در همسایگی x^* (نمایش با $\mathcal{H}$) $f(x^*) \le f(x),\ x \in \mathcal{H}$
  - معروف به کمینه‌ساز محلی ضعیف
- کمینه‌ساز محلی اکید
  - $f(x^*) < f(x),\ x \in \mathcal{H}, x \neq x^*$

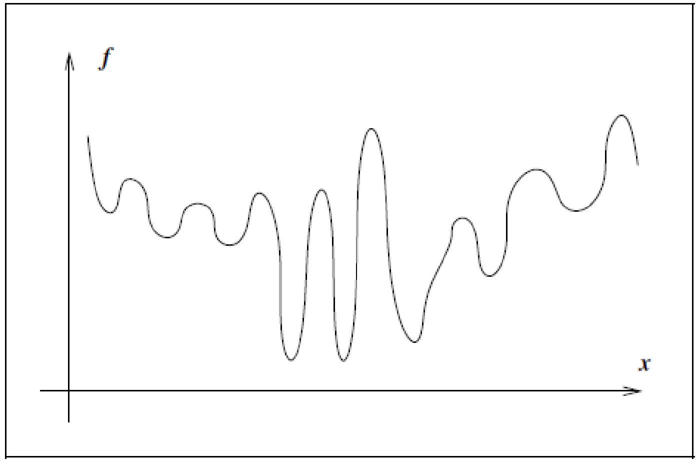
تشخیص کمینه محلی

مثال

- $f(x) = 2$

مثال ۲

- $f(x) = (x - 2)^4$



کمینه‌ساز محلی تک‌افتاده

یافتن کمینهٔ محلی

جستجو همسایگی؟

تابع هموار

- روش‌های کاراتر و عملی‌تر

تابع مشتق‌پذیر مرتبه اول و دوم

- وجود گرادیان و هسی
- محتملا بتوان نقطه‌ای را کمینه محلی (کمینه محلی قوی) نامید
- ابزار ریاضی مطالعهٔ کمینه‌سازها
  - قضیهٔ تیلور

قضیهٔ تیلور

قضیهٔ تیلور

$f:\mathbb{R}^n \to \mathbb{R}$ تابعی مشتق‌پذیر پیوسته و $\boldsymbol{p} \in \mathbb{R}^n$، آن‌گاه

$$f(\boldsymbol{x}+\boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}+t\boldsymbol{p})^T \boldsymbol{p}$$

$t \in (0,1)$.

هم‌چنین، اگر مشتق‌پذیر مرتبه دوم پیوسته باشد، آن‌گاه

$$\nabla f(\boldsymbol{x}+\boldsymbol{p}) = \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(\boldsymbol{x}+t\boldsymbol{p})\boldsymbol{p}\,dt$$

$$f(\boldsymbol{x}+\boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\nabla^2 f(\boldsymbol{x}+t\boldsymbol{p})\boldsymbol{p}$$

$t \in (0,1)$

قضیهٔ شروط لازم مرتبهٔ اول

$\boldsymbol{x}^*$ کمینهٔ محلی و f در همسایگی نقطهٔ مذکور مشتق‌پذیر و پیوسته است، آن‌گاه $\nabla f(\boldsymbol{x}^*) = 0$

اثبات

$\boldsymbol{x}^*$ نقطه ماناست اگر $\nabla f(\boldsymbol{x}^*) = 0$

نتیجهٔ قضیه: هر کمینهٔ محلی، نقطهٔ ماناست.

قضیهٔ شروط لازم مرتبهٔ دوم

x^* کمینهٔ محلی f و $\nabla^2 f$ موجود و پیوسته همسایهٔ نقطهٔ مذکور، آن‌گاه $\nabla f(x^*) = 0$ و $\nabla^2 f(x^*)$ مثبت نیمه‌معین است.

اثبات

قضیهٔ شروط کافی مرتبهٔ دوم

$\nabla^2 f$ موجود و در همسایگی x^* پیوسته و $\nabla f(x^*) = 0$ و $\nabla^2 f(x^*)$ مثبت معین، آن‌گاه نقطه مذکور کمینهٔ محلی اکید تابع **f** است.

اثبات

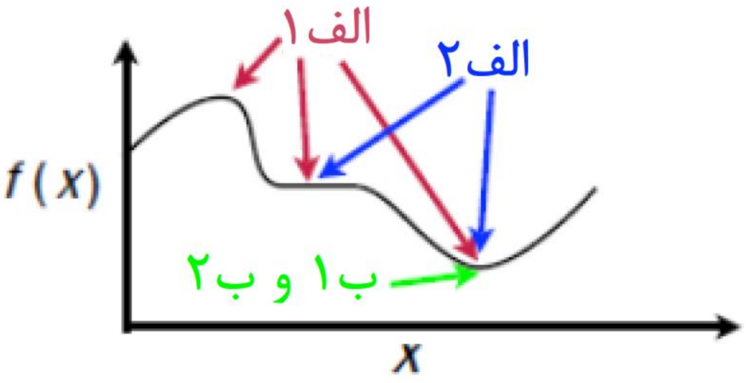
شرایط بهینگی کمینه‌سازی چندمتغیره

قضیهٔ شروط لازم کمینه محلی ضعیف

- الف۱: $\nabla f(x^*) = 0$ نقطه مانا
- الف۲: $\nabla^2 f(x^*)$ مثبت نیمه‌معین

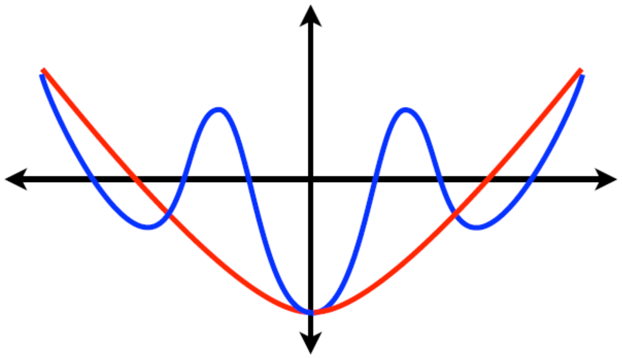
قضیهٔ شروط کافی کمینه محلی قوی

- ب۱: $\nabla f(x^*) = 0$
- ب۲: $\nabla^2 f(x^*)$ مثبت معین



قضیهٔ کوژ

f کوژ باشد، هر x^* کمینه‌ساز محلی کمینه‌ساز سراسری f است. هم‌چنین اگر f مشتق‌پذیر باشد، آن‌گاه هر نقطه مانای x^* کمینه‌ساز سراسری f خواهد بود.



کوژی

- ماتریس هسی مثبت نیمه‌معین

اکیدا کوژ

- ماتریس هسی مثبت معین

حل

نتایج بدست آمده از حسابان ساده و مقدماتی

بنیادی جهت الگوریتم‌های بهینه‌سازی نامقید

هر الگوریتم با یکی از روش‌ها به دنبال یافتن نقطه‌ای که گرادیان f ناپدید می‌شود

مسائل ناهموار

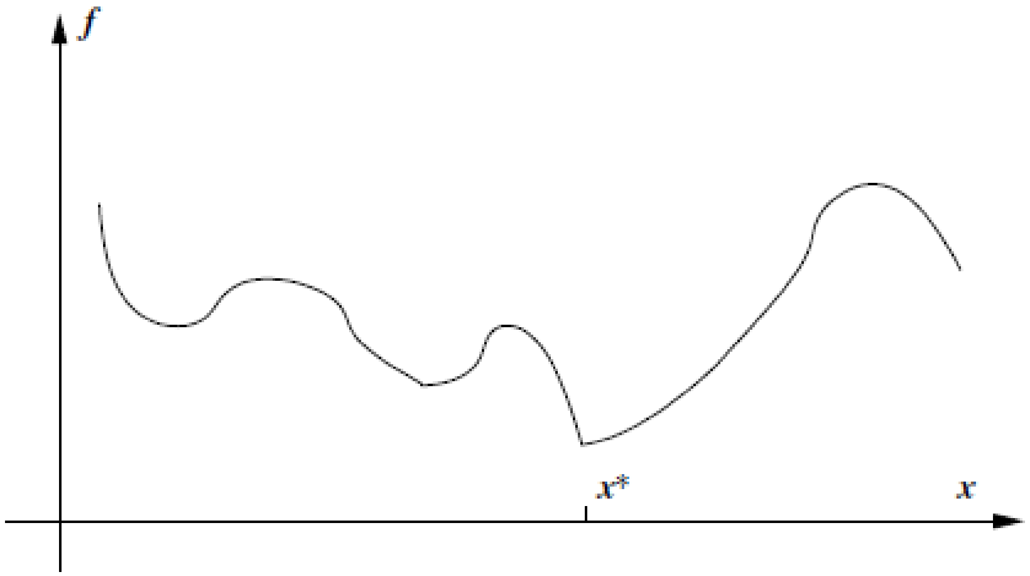
توابع ناهموار و ناپیوسته

روشی عمومی برای حل وجود ندارد

در صورت اتصال چند قطعهٔ هموار با ناپیوستگی بین قطعه‌ها

- امکان یافتن کمینه‌ساز با کمینه‌کردن جداگانهٔ هر قطعه

از روش‌های زیرگرادیان یا گرادیان تعمیمی



الگوریتم‌های بهینه‌سازی نامقید

شصت سال اخیر

جملگی با شروع از نقطهٔ آغاز

تکرار دنباله‌ای از مراحل

پایان

- بهبود غیرممکن
- رسیدن به تخمین مناسبی از پاسخ

نزولی $f\ (x_k)\ <\ f\ (x_{k-m})$

دو استراتژی اساسی جهت حرکت از نقطه فعلی x_k به نقطهٔ x_{k+1}

- جستجو خط
- منطقه اعتماد

انتخاب جهت حرکت و اندازهٔ حرکت

الگوریتم‌های بهینه‌سازی نامقید

روش‌های گرادیان-محور

- **انتخاب جهت حرکت و اندازهٔ حرکت**

الگوریتم گرادیان-محور

انتخاب مقدار اولیهٔ $\boldsymbol{x}_0$ و $k=0$

تازمان همگرا نشدن

}

انتخاب جهت $\boldsymbol{p}_k$ و اندازه قدم α_k

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$$

$$k \leftarrow k+1$$

{

دو استراتژی- جستجو خط و منطقه اعتماد

استراتژی جستجو خط

- انتخاب جهت $\boldsymbol{p}_k$
- جستجو در راستای جهت یافت شده از $\boldsymbol{x}_k$ فعلی به تکرار جدیدی با مقدار f کمتر
- با حل مسئله کمینه‌سازی یک‌بعدی زیر جهت یافتن α

$$\underset{\alpha>0}{\text{کم}}\, f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k)$$

استراتژی منطقه اعتماد

- تعریف شعاعی $\Delta > 0$ اطراف $\boldsymbol{x}_k$ به عنوان منطقهٔ اعتماد
- جمع‌آوری اطلاعات جهت ایجاد مدلی (تخمینی) از تابع f
  - تابع تخمین با نام $\boldsymbol{m}_k$
- دارای رفتار مشابه در نزدیکی نقطه $\boldsymbol{x}_k$

$$\underset{\boldsymbol{p}}{\text{کم}}\, \boldsymbol{m}_k(\boldsymbol{x}_k + \boldsymbol{p})$$

$\boldsymbol{x}_k + \boldsymbol{p}$ در داخل منطقه اعتماد

- $\|\boldsymbol{p}\|_2 \leq \Delta$

دو استراتژی - ادامه

استراتژی منطقه اعتماد

- تابع تخمین $\boldsymbol{m}_k$ معمولا برابر با

$$\boldsymbol{m}_k(\boldsymbol{x}_k + \boldsymbol{p}) = f_k + p^T \nabla f_k + p^T B_k p$$

B_k یا تابع هسی $\nabla^2 f_k$ یا تخمینی از آن

دو استراتژی- جستجو خط

انتخاب $\boldsymbol{p}_k$ در جهت کاهش

- بیشترین نزول (گرادیان نزولی): مرتبه اول، همگرائی خطی
- روش گرادیان مزدوج: مرتبه اول، همگرائی (سریعتر) خطی
- روش نیوتن: مرتبه دوم، همگرائی درجه دو
- روش شبه نیوتن: مرتبه اول تا تخمین مرتبۀ دوم، همگرائی ابرخطی
- با حل مسئله کمینه‌سازی یک‌بعدی زیر جهت یافتن α

انتخاب طول قدم α_k برآورده‌گر شروط وولف

- کروشه‌گذاری: یافتن بازه‌ای شامل طول قدم مناسب
- نیمه‌سازی/درون‌یابی: محاسبۀ قدم مناسب در بازۀ فعلی

شدیدترین نزول

شبیه پائین‌آمدن از کوه

انتخاب سریعترین جهت سرازیری

$$\boldsymbol{p}_k = -\nabla f$$

کاربرد

- بهینه‌سازی
- شبکه عصبی
- یادگیری ماشین
- شبکه عمیق

مزایا

- صرفا استفاده از اطلاع مرتبه اول
- همیشه جهت نزول
- حافظه کم

معایب

- کند در مسائل سخت
- حساس به مقیاس‌گذاری

شدیدترین نزول- مثال

حل تحلیلی

$$f(x) = x^2 - 2x + 2 = (x-1)^2 + 1$$

$$\underset{x}{\text{کــــم}}\, f(x)$$

$$\frac{df}{dx} = 2\mathrm{x} - 2$$

$$\mathrm{x} = 1$$

شدیدترین نزول- مثال - ادامه

حل با استفاده از شدیدترین نزول

$$f(x) = x^2 - 2x + 2 = (x-1)^2 + 1$$

$$\underset{x}{\text{کـــــــم}}\, f(x)$$

مقدار بهینه را نمی‌دانیم

- شروع از مقدار تصادفی x=3

قدم اول: مشتق‌گیری

$$\frac{df}{dx} = 2\mathrm{x} - 2$$

شدیدترین نزول- مثال - ادامه

قدم دوم

- مطالعۀ مشتق در نقطۀ داده شده

$$f'(3) = 2(3) - 2 = 4$$

- مشتق در کمینه باید صفر باشد
- مقدار مثبت مشتق
  - نمایشگر اینکه مقدار تابع افزایشی است. و باید عقب رفت
- اگر مقدار x=-1 به عنوان حدس اولیه انتخاب می‌شد
  - آن‌گاه مشتق f’=-4
  - نمایشگر نزولی بودن مقدار و در نتیجه نیاز به جلو رفتن

مقدار فعلی مشتق نشانگر مسیر

- نزدیک شدن به یا دور شدن از کمینه

شدیدترین نزول- مثال - ادامه

$$x_{i+1} = x_i + \alpha(-f'(x_i))$$

x_{i+1} حدس بعدی

طول قدم $\alpha = 0.2$

$x_0 = 3$

$$x_{i+1} = x_i + 0.2(-f'(x_i))$$
$$x_1 = 3 - \alpha f'(3) = 3 + 0.2(-4) = 2.2$$
$$x_2 = 2.2 - \alpha f'(2.2) = 2.2 + 0.2(-2.4) = 1.72$$

ادامه محتملا به پاسخ می‌رسد.

- ?

چرا شدیدترین نزول به جای روش تحلیلی و حسابان

نوشتن برنامه

شدیدترین نزول- مثال ۲

x	y
1	1
2	1
2	2
3	2

یافتن بهترین خط برازش

$$y = \alpha x + \beta$$

به دنبال یافتن α و β

ابتدا نیاز به تعریف خطا

- خطا تفاضل بین داده و مقدار بدست آمده از مدل

$$e_1 = \hat{y}_1 - y_1$$
$$e_2 = \hat{y}_2 - y_2$$

نیاز به کل خطا

- یک روش:جمع تمامی مقادیر خطا $e_{\text{کل}} = e_1 + e_2 + e_3 + e_4$
  - غلطانداز
- روش بهتر $e_{\text{کل}} = e_1^2 + e_2^2 + e_3^2 + e_4^2$

شدیدترین نزول- مثال ۲ - ادامه

x	y
1	1
2	1
2	2
3	2

روش بهتر $e_{\text{کل}} = e_1^2 + e_2^2 + e_3^2 + e_4^2$

$$e_1 = \hat{y}_1 - y_1 \Rightarrow e_1 = [\alpha x_1 + \beta] - y_1$$
$$e_2 = \hat{y}_2 - y_2 \Rightarrow e_2 = [\alpha x_1 + \beta] - y_2$$
$$e_3 = \hat{y}_3 - y_3 \Rightarrow e_3 = [\alpha x_1 + \beta] - y_3$$
$$e_4 = \hat{y}_4 - y_4 \Rightarrow e_4 = [\alpha x_1 + \beta] - y_4$$

خطای کل

$$e_{\text{کل}} = \sum_{i=1}^{4} e_i^2$$
$$= \sum_{i=1}^{4} ([\alpha x_i + \beta] - y_i)^2$$

شدیدترین نزول - مثال ۲ - ادامه

x	y
1	1
2	1
2	2
3	2

به دنبال یافتن خطی با کمترین خطای ممکن

$$e_{\text{کل}} = \sum_{i=1}^{4} e_i^2 = \sum_{i=1}^{4} \left([\alpha x_i + \beta] - y_i \right)^2$$

$$\underset{\alpha,\beta}{\text{کم}} \sum_{i=1}^{4} \left([\alpha x_i + \beta] - y_i \right)^2$$

استفاده از شدیدترین نزول

$$\begin{bmatrix} \alpha_{\text{ب}} \\ \beta_{\text{ب}} \end{bmatrix} = \begin{bmatrix} \alpha_{\text{ق}} \\ \beta_{\text{ق}} \end{bmatrix} - 0.2 \begin{bmatrix} \frac{\partial f}{\partial \alpha}(\alpha_{\text{ق}}, \beta_{\text{ق}}) \\ \frac{\partial f}{\partial \beta}(\alpha_{\text{ق}}, \beta_{\text{ق}}) \end{bmatrix}$$

شدیدترین نزول- مثال ۲ - ادامه

x	y
1	1
2	1
2	2
3	2

$$\mathrm{f}(\alpha,\beta) = \sum_{i=1}^{4}(\,[\alpha x_i + \beta] - y_i)^2$$

$$\frac{\partial f}{\partial \alpha} = \sum_{i=1}^{4} 2(\,[\alpha x_i + \beta] - y_i)x_i$$

$$\frac{\partial f}{\partial \beta} = \sum_{i=1}^{4} 2(\,[\alpha x_i + \beta] - y_i)$$

مقدار اولیه

$$\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix} = \begin{bmatrix}1\\ 1\end{bmatrix}$$

شدیدترین نزول - مثال ۲ - ادامه

x	y
1	1
2	1
2	2
3	2

$$\begin{bmatrix}\alpha_{\text{ب}}\\ \beta_{\text{ب}}\end{bmatrix} = \begin{bmatrix}\alpha_{\text{ق}}\\ \beta_{\text{ق}}\end{bmatrix} - 0.2\begin{bmatrix}\sum_{i=1}^{4} 2(\,[\alpha x_i + \beta] - y_i)x_i\\ \sum_{i=1}^{4} 2(\,[\alpha x_i + \beta] - y_i)\end{bmatrix}$$

مقدار اولیه

$$\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix} = \begin{bmatrix}1\\ 1\end{bmatrix}$$

شدیدترین نزول- مثال ۲ - ادامه

x	y
1	1
2	1
2	2
3	2

تحریر محل نزاع

- چهار داده
  - محاسبات فراوان
- حال اگر یک میلیون داده موجود باشد
- یک میلیارد داده چه؟
- راه‌حل؟

$$\begin{bmatrix} \alpha_{\text{ب}} \\ \beta_{\text{ب}} \end{bmatrix} = \begin{bmatrix} \alpha_{\text{ق}} \\ \beta_{\text{ق}} \end{bmatrix} - 0.2 \begin{bmatrix} \sum_{i=1}^{4} 2(\,[\alpha x_i + \beta] - y_i)x_i \\ \sum_{i=1}^{4} 2(\,[\alpha x_i + \beta] - y_i) \end{bmatrix}$$

شدیدترین نزول- مثال ۲ - ادامه

راه‌حل

- گرادیان نزولی تصادفی

$$\begin{bmatrix}\alpha_{j+1}\\ \beta_{j+1}\end{bmatrix} = \begin{bmatrix}\alpha_j\\ \beta_j\end{bmatrix} - 0.2\begin{bmatrix}2(\,[\alpha_j x_1 + \beta_j] - y_1)x_1\\ 2(\,[\alpha_j x_1 + \beta_j] - y_1)\end{bmatrix}$$

$$\begin{bmatrix}\alpha_{j+2}\\ \beta_{j+2}\end{bmatrix} = \begin{bmatrix}\alpha_{j+1}\\ \beta_{j+1}\end{bmatrix} - 0.2\begin{bmatrix}2(\,[\alpha_{j+1} x_2 + \beta_{j+1}] - y_2)x_2\\ 2(\,[\alpha_{j+1} x_2 + \beta_{j+1}] - y_2)\end{bmatrix}$$

- در هم کردن داده‌ها و خواندن از نمونه نخست
- ادامه تا خواندن همه نمونه‌ها و شروع کار با نمونه اول
- ادامه تا همگرائی

شدیدترین نزول- مثال ۲ - ادامه

گرادیان نزولی

- کندتر
- صحیح‌تر
- دسته‌ای
- استفاده از همه نمونه‌ها

گرادیان نزولی تصادفی

- سریعتر
- دقت کمتر
- استفاده از یک نمونه در هر زمان

گرادیان نزولی زیردسته‌ای

- بیشتر از یک نمونه در هرزمان و کمتر از تمامی نمونه‌ها

طول قدم α

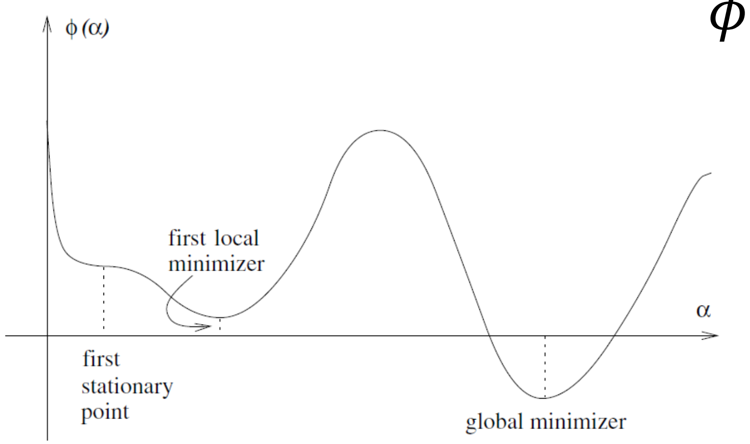
طول قدم

- نیاز به سبک‌سنگین‌کردن
- انتخاب طولی با کاهش مقدار قابل توجه **f**
- بدون سپری کردن زمان زیاد برای یافتن طول مناسب

تابعی بر اساس طول قدم

$$\phi(\alpha_i) = f(\boldsymbol{x}_i + \alpha_i \boldsymbol{p}_i), \alpha_i > 0$$

- تابعی تک متغیره
- کمینه‌ساز سراسری
- مشکل یافتن کمینه سراسری
- سیاست عملی‌تر
  - جستجو خط جزیی

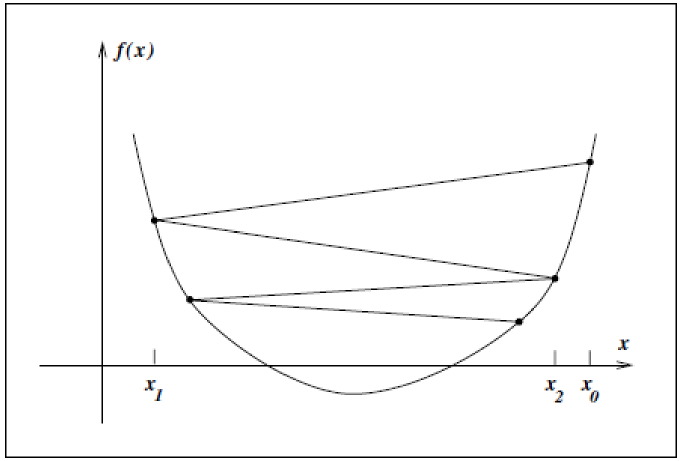


طول قدم α

امر غالب

- الگوریتم جستجو به دنبال یافتن مقداری برای طول قدم
- اتمام الگوریتم هنگام یافتن مقدار برآورده‌کنندهٔ چند شرط
- انجام جستجو خط در دو مرحله
  - کروشه‌گذاری
    - یافتن بازه مقادیر مطلوب
  - درون‌یابی/نیمه‌سازی
    - یافتن مقدار مناسب در بازه مذکور
- ابتدا بررسی شرط خاتمه

طول قدم α

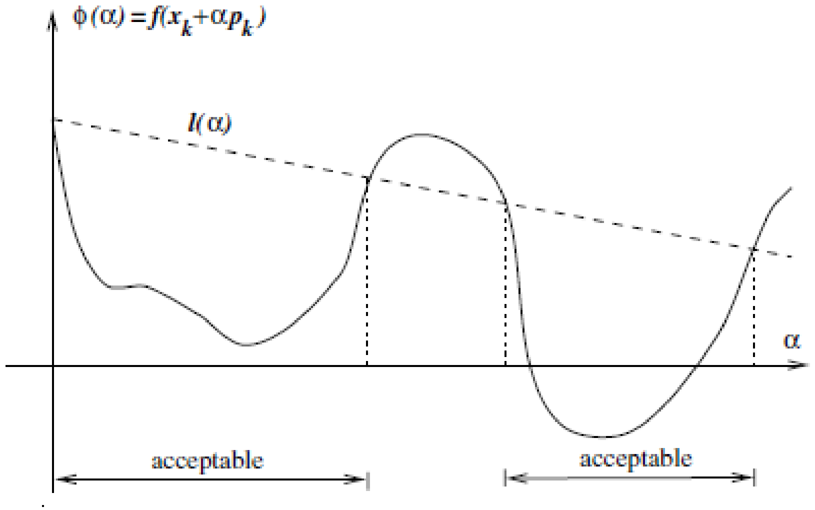


سادهترین شرط

- کاهش در مقدار تابع
- $f(\boldsymbol{x}_k + \alpha_i \boldsymbol{p}_k) < f(\boldsymbol{x}_k)$
- ناکافی جهت همگرایی به کمینهساز

نیاز به شرط کافی کاهش

شروط وولف



طول قدم در هر مرحله صادق در کاهش کافی تابع هدف

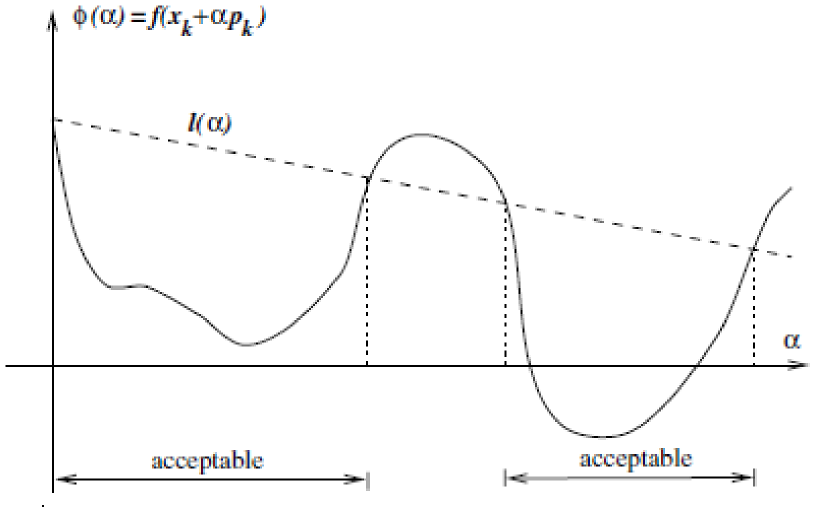
کاهش کافی

- $f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \leq f(\boldsymbol{x_i}) + c_1 \alpha_i \nabla f_i^T \boldsymbol{p}_i$
- $c_1 \in (0,1)$
- تناسب کاهش **f** با
  - طول قدم α_i
  - مشتق جهت‌دار $\nabla f_i^T \boldsymbol{p}_i$
- شهره به شرط ارمیخو
  - $\phi(\alpha_i) \leq l(\alpha_i)$
- اختصاص مقداری کوچک به c_1
  - 10^{-4}

$$l(\alpha_i) = f(\boldsymbol{x}_i) + c_1 \alpha_i \nabla f_i^T \boldsymbol{p}_i$$

شروط وولف

طول قدم در هر مرحله صادق در کاهش کافی تابع هدف



کاهش کافی

- $f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \leq f(\boldsymbol{x_i}) + c_1 \alpha_i \nabla f_i^T \boldsymbol{p}_i$
- $c_1 \in (0,1)$
- تناسب کاهش **f** با
  - طول قدم α_i
  - مشتق جهتدار $\nabla f_i^T \boldsymbol{p}_i$

$$l(\alpha_i) = f(\boldsymbol{x}_i) + c_1 \alpha_i \nabla f_i^T \boldsymbol{p}_i$$

- شهره به شرط ارمیخو
  - $\phi(\alpha_i) \leq l(\alpha_i)$
- اختصاص مقداری کوچک به c_1

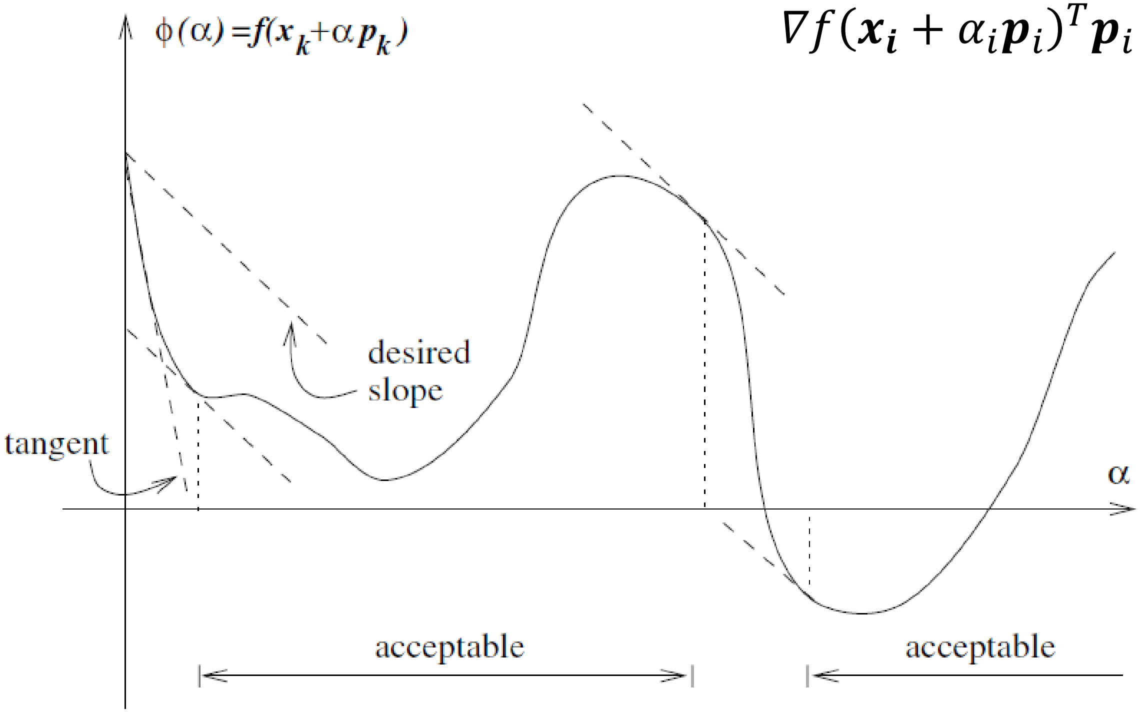
اما «شرط کافی» مکفی نیست!

شروط وولف

جهت حذف مقادیر اقدام کوچک

- شرط انحنا

$$\nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i \geq c_2 \alpha_i \nabla f(\boldsymbol{x}_i)^T \boldsymbol{p}_i$$



- $c_2 \in (c_1, 1)$
- $\phi'(\alpha_i) = \nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i$
  - شیب ϕ در α_i بزرگتر از مضربی از شیب ابتدایی $\phi'(0)$ (یا بزرگتر از $c_2\phi'(0)$)
  - شیب خیلی منفی
    - امکان کاهش بیشتر
  - شیب کمی منفی یا کمی مثبت
    - ته خط!
    - عدم انتظار کاهش بیشتر در مقدار تابع

شروط وولف

جهت حذف مقادیر اقدام کوچک

- شرط انحنا

$$\nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i \geq c_2 \alpha_i \nabla f(\boldsymbol{x}_i)^T\, \boldsymbol{p}_i$$

- $c_2 \in (c_1, 1)$
- $\phi'(\alpha_i) = \nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i$
  - شیب ϕ در α_i بزرگتر از مضربی از شیب ابتدایی $\phi'(0)$ (یا بزرگتر از $c_2\phi'(0)$)
  - شیب خیلی منفی
    - امکان کاهش بیشتر
  - شیب کمی منفی یا کمی مثبت
    - ته خط!
    - عدم انتظار کاهش بیشتر در مقدار تابع
- c_2
  - روش نیوتن یا شبه نیوتن معمولا برابر ۰٫۹
  - روش گرادیان مزدوج معمولا برابر ۰٫۱

شروط ولف

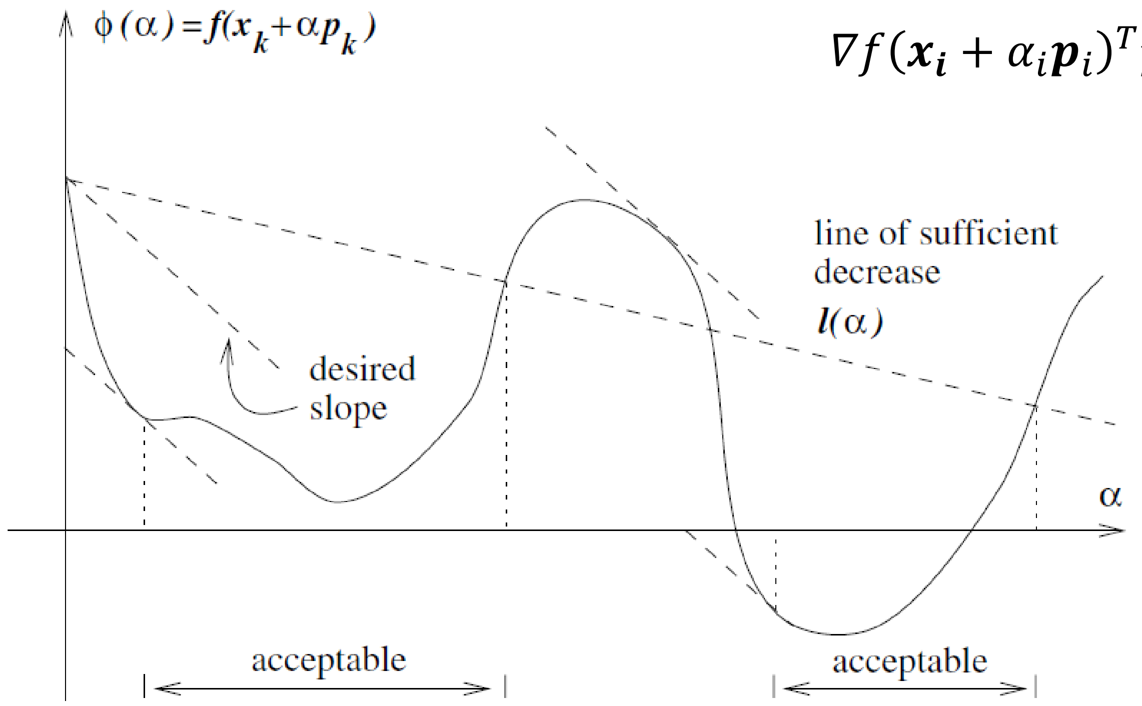
- شرط ارمیخو

$$f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \le f(\boldsymbol{x_i}) + c_1 \alpha_i \nabla f(\boldsymbol{x}_i)^T\, \boldsymbol{p}_i$$

- شرط انحنا

$$\nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i \ge c_2 \alpha_i \nabla f(\boldsymbol{x}_i)^T\, \boldsymbol{p}_i$$

- $0 < c_1 < c_2 < 1$



شروط قوی وولف

امکان وجود شرایط وولف بدون اینکه طول مناسبی باشد

شروط قوی وولف

$$f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \le f(\boldsymbol{x_i}) + c_1 \alpha_i \nabla f(\boldsymbol{x}_i)^T \, \boldsymbol{p}_i$$

$$\left|\nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i\right| \le c_2 \left|\alpha_i \nabla f(\boldsymbol{x}_i)^T \, \boldsymbol{p}_i\right|$$

- $0 < c_1 < c_2 < 1$

شروط قوی وولف

امکان وجود شرایط وولف بدون اینکه طول مناسبی باشد

شروط قوی وولف

$$f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \leq f(\boldsymbol{x_i}) + c_1 \alpha_i \nabla f(\boldsymbol{x}_i)^T \, \boldsymbol{p}_i$$

$$\left|\nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i\right| \leq c_2 \left|\alpha_i \nabla f(\boldsymbol{x}_i)^T \, \boldsymbol{p}_i\right|$$

- $0 < c_1 < c_2 < 1$
- جلوگیری از مقادیر خیلی مثبت

شروط گلدشتین

اطمینان از دستیابی طول به کاهش کافی و جلوگیری از کوتاه قدم‌برداشتن

$$f(\boldsymbol{x_i}) + (1-c)\alpha_i \nabla f(\boldsymbol{x}_i)^T \boldsymbol{p}_i \leq f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \leq f(\boldsymbol{x_i}) + c\alpha_i \nabla f(\boldsymbol{x}_i)^T \boldsymbol{p}_i$$

$$0 < c < 0.5$$

همان شرط کاهش کافی

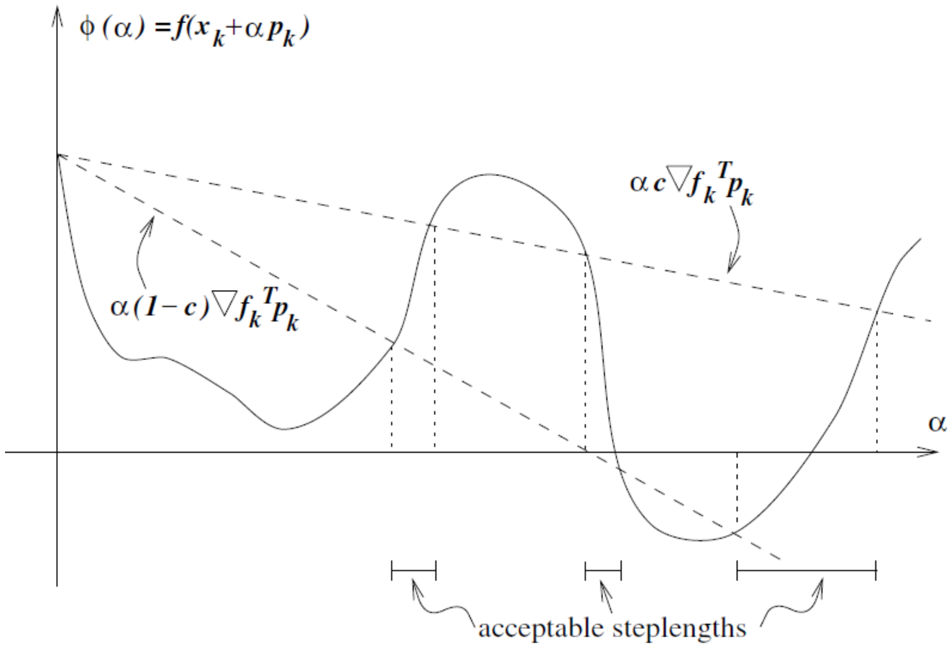
شروط گلدشتین

اطمینان از دستیابی طول به کاهش کافی و جلوگیری از کوتاه قدم‌برداشتن

$$f(\boldsymbol{x_i}) + (1-c)\alpha_i \nabla f(\boldsymbol{x}_i)^T \boldsymbol{p}_i \le f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i) \le f(\boldsymbol{x_i}) + c\alpha_i \nabla f(\boldsymbol{x}_i)^T \boldsymbol{p}_i$$

$$0 < c < 0.5$$

کنترل طول قدم از پائین



کاهش کافی با پس‌روی

صرفا استفاده از شرط کافی و نه شرط دوم

الگوریتم جستجو خط با پس‌روی

- انتخاب $\overline{\alpha} > 0$ و $\rho \in (0,1)$ و $c \in (0,1)$ و $\alpha = \overline{\alpha}$
- تازمان $f(\boldsymbol{x_i} + \alpha \boldsymbol{p}_i) \leq f(\boldsymbol{x_i}) + c\alpha \nabla f(\boldsymbol{x}_i)^T \boldsymbol{p}_i$
- }
    - $\alpha = \rho\alpha$
- {
- $\alpha_i = \alpha$

امکان تخصیص پویای ρ در هر مرحله

طول قدم α

طول قدم

اندازه هر پرش

- بزرگ
  - همگرا نشدن
- کوچک
  - کند

یکی از راه‌حل‌ها

- شروع با طول قدم بزرگ
  - مثلا ۱
- کاهش طول قدم هنگام اضافه رفتن $\alpha = 0.8\alpha$
- کاهش دوبارهٔ مقدار درصورت کافی نبودن کاهش مقدار قبلی
- ادامه تا یافتن طول قدم مناسب

راه دیگر $\alpha_i = \frac{c}{i}$

طول قدم α – ادامه

روش دیگر

- وابسته به شکل تابع
- در طول بهینه‌سازی
  - پیگیری گرادیان‌های بدست آمده در گذشته
- آداگراد

$$\boldsymbol{s} = 0$$

تازمان همگرا نشدن

$$\boldsymbol{g} = -\nabla f(\boldsymbol{x}_i)$$

$$\boldsymbol{s} = \boldsymbol{s} + [\![\nabla f(\boldsymbol{x}_i)]\!]\text{^}2$$

$$\boldsymbol{x_{i+1}} = \boldsymbol{x}_i + \alpha \frac{\boldsymbol{g}}{\sqrt{\boldsymbol{s}+\epsilon}}$$

John Duchi, Elad Hazan, Yoram Singer, "adaptive Subgradient Methods for Online Learning and Stochastic Optimization," JMLR, 12(61):2121–2159, 2011.

طول قدم α – ادامه

مشکل آداگراد

- جمع مربعات در مخرج
  - بی‌اندازه کوچک شدن طول قدم
- آدادلتا
  - جلوگیری از کوچک‌شدن بی‌اندازۀ مقدار طول قدم
  - کاهش طول و استفاده از قبلی‌ها
  - با استفاده از پنجره از چند مقدار قبلی

$$\boldsymbol{s} = 0$$

تازمان همگرا نشدن

$$\boldsymbol{g} = -\nabla f(\boldsymbol{x}_i)$$

$$\boldsymbol{s} = \gamma \boldsymbol{s} + (1-\gamma)[\nabla f(\boldsymbol{x}_i)]^2$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \alpha \frac{\boldsymbol{g}}{\sqrt{\boldsymbol{s}+\epsilon}}$$

D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method Matthew," 2012

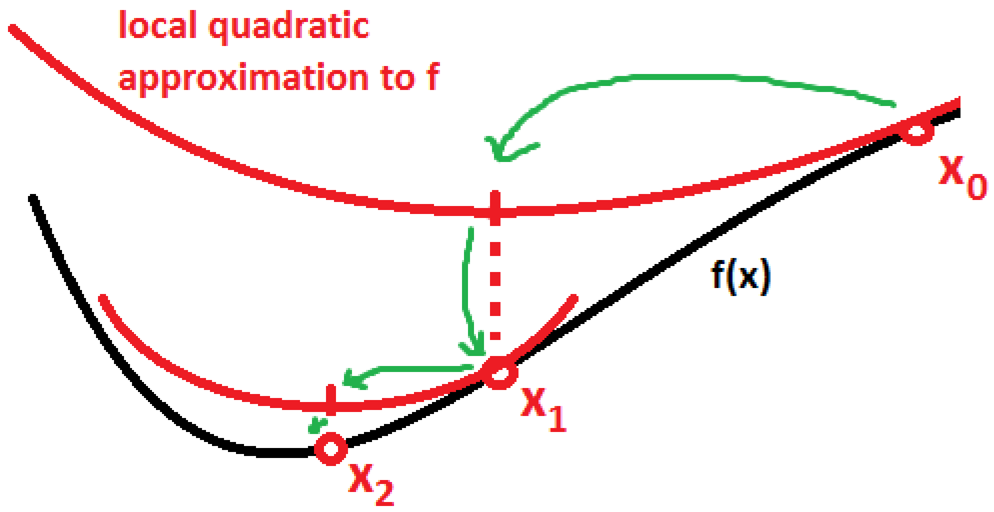
روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

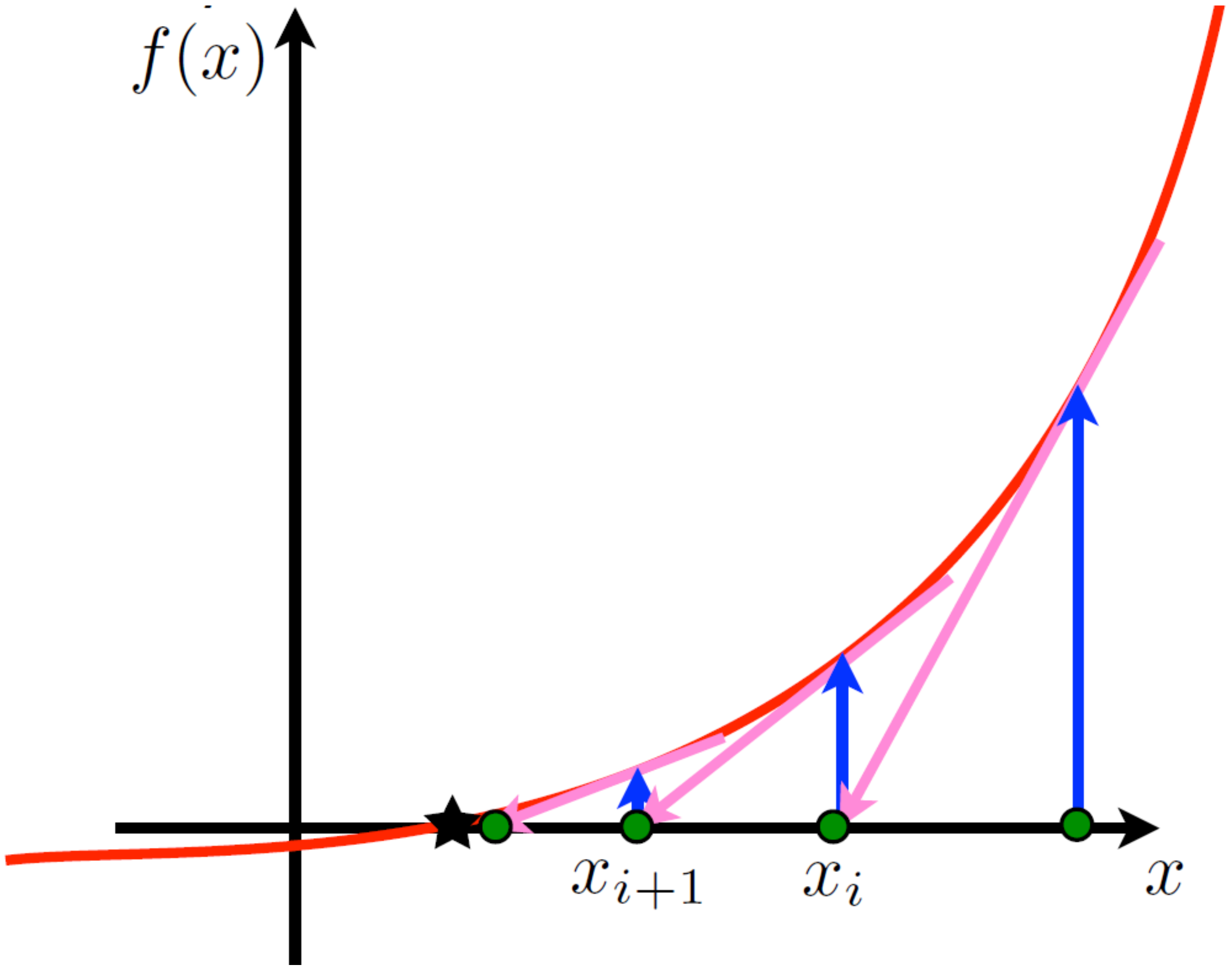
گن $\Leftarrow x_{i+1} = x_i - \alpha_i\big(\nabla f(x_i)\big)$

روشی مبتنی بر مرتبه دوم

- ماتریس هسی



https://math.stackexchange.com/questions/609680/newtons-method-intuition



$f(x)$
x_{i+1}
x_i
x

روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

گن $\Leftarrow x_{i+1} = x_i - \alpha_i\left(\nabla f(x_i)\right)$

روشی مبتنی بر مرتبه دوم

- ماتریس هسی

یادآوری جهت تقریب ذهن- برای یافتن صفر (یک بعدی)

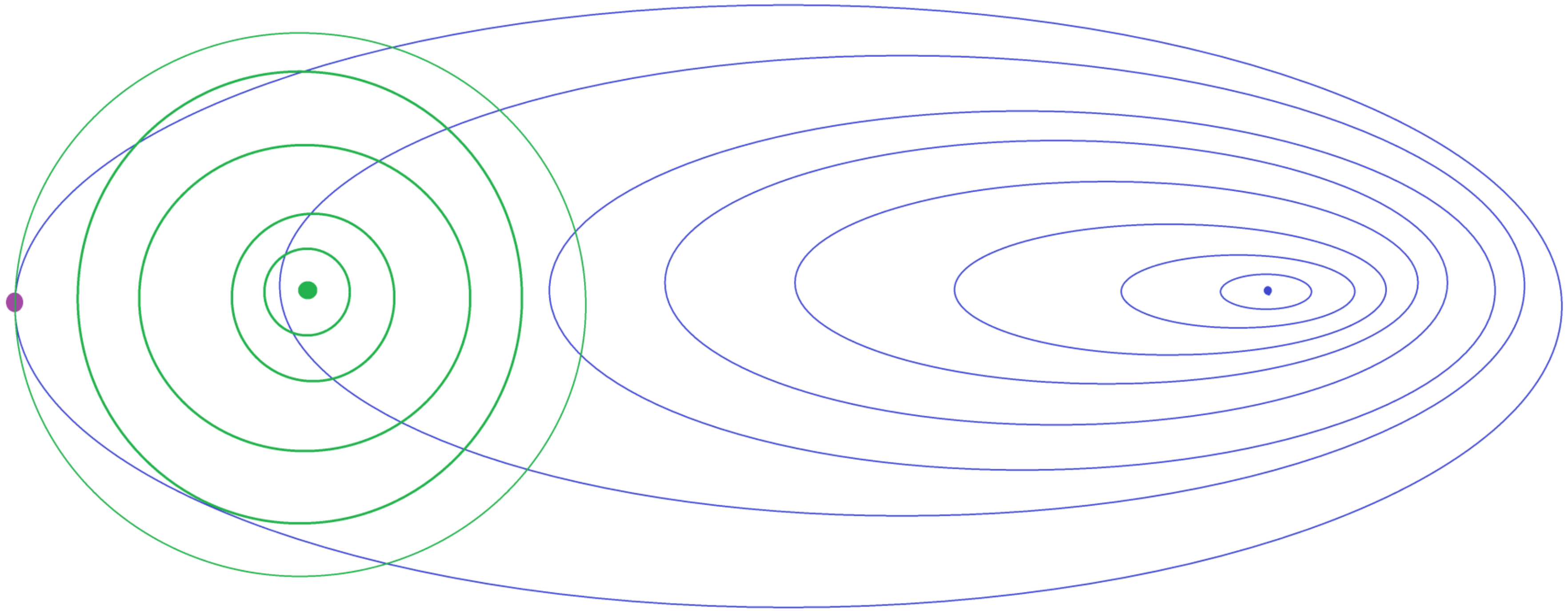
$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

کمینه یا بیشینه

- به دنبال اینکه مشتق صفر باشد

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

روش نیوتن - ادامه



روش نیوتن - ادامه

$f: \mathbb{R}^n \to \mathbb{R}$

- به دنبال یافتن کمینه

قضیهٔ تیلور جهت $\boldsymbol{x} \in \mathbb{R}^n$ نزدیک a

$$f(\boldsymbol{x}) \approx f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^T(\boldsymbol{x}-\boldsymbol{a}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^T \nabla^2 f(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a})$$

$$g = \nabla f(a)$$

$$H = \nabla^2 f(a) = \frac{\partial^2 y}{\partial x_i \partial x_j}, i, j \in \{1, \dots, n\}$$

دیگر روش نوشتن

$$f(\boldsymbol{x}+\boldsymbol{p}) \approx f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{p}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k + \boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T \nabla^2 f(\boldsymbol{x}_k) \boldsymbol{p}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k + \boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k + \boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{H}_k \boldsymbol{p}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k + \boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k + \boldsymbol{p}) \approx \underbrace{f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{H}_k \boldsymbol{p}}_{q(\boldsymbol{p})}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx \underbrace{f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{H}_k\boldsymbol{p}}_{q(\boldsymbol{p})}$$

روش نیوتن: $\boldsymbol{p}$ کاهنده تقریب درجهٔ دو $q(\boldsymbol{p})$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx \underbrace{f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{H}_k\boldsymbol{p}}_{q(\boldsymbol{p})}$$

روش نیوتن: $\boldsymbol{p}$ کاهنده تقریب درجهٔ دو $q(\boldsymbol{p})$

$$\nabla q(\boldsymbol{p}) = \boldsymbol{g} + H\boldsymbol{p} = \boldsymbol{0}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx \underbrace{f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{H}_k\boldsymbol{p}}_{q(\boldsymbol{p})}$$

روش نیوتن: $\boldsymbol{p}$ کاهنده تقریب درجهٔ دو $q(\boldsymbol{p})$

$$\nabla q(\boldsymbol{p}) = \boldsymbol{g} + H\boldsymbol{p} = \boldsymbol{0}$$

$$H\boldsymbol{p} = -\mathbf{g}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx \underbrace{f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{H}_k\boldsymbol{p}}_{q(\boldsymbol{p})}$$

روش نیوتن: $\boldsymbol{p}$ کاهنده تقریب درجهٔ دو $q(\boldsymbol{p})$

$$\nabla q(\boldsymbol{p}) = \boldsymbol{g} + H\boldsymbol{p} = \boldsymbol{0}$$

معادلهٔ نیوتن

$$H\boldsymbol{p} = -\mathbf{g}$$

روش نیوتن - ادامه

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$$

$$f(\boldsymbol{x}_k+\boldsymbol{p}) \approx \underbrace{f(\boldsymbol{x}_k) + \boldsymbol{g}_k{}^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{H}_k \boldsymbol{p}}_{q(\boldsymbol{p})}$$

روش نیوتن: $\boldsymbol{p}$ کاهنده تقریب درجهٔ دو $q(\boldsymbol{p})$

$$\nabla q(\boldsymbol{p}) = \boldsymbol{g} + H\boldsymbol{p} = \boldsymbol{0}$$

معادلهٔ نیوتن

$$H\boldsymbol{p} = -\mathbf{g}$$

$$\boldsymbol{p_k} = -\boldsymbol{H_k^{-1}} \boldsymbol{g_k}$$

$\boldsymbol{p_k}$ جهت نیوتن

روش نیوتن - ادامه

$$\boldsymbol{p}_{\boldsymbol{k}} = -\boldsymbol{H}_{\boldsymbol{k}}^{-1}\boldsymbol{g}_{\boldsymbol{k}}$$

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_{\boldsymbol{k}}$ با جستجو خط

روش نیوتن - ادامه

الگوریتم روش نیوتن

مقداردهی اولیه $\boldsymbol{x}_0 \in \mathbb{R}^n$

تکرار $x_{i+1} = x_i - H^{-1} g$

- $g = \nabla f(x_i)$
- $H = \nabla^2 f(x_i)$

روش نیوتن – حل معادلهٔ غیرخطی

$$\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$$

$$\nabla f(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}$$

نیاز به حل دستگاه n-معادلهٔ غیرخطی $\boldsymbol{g_i}(\boldsymbol{x}) = 0$

$$g(\boldsymbol{x}_k + \boldsymbol{p}) \approx g(\boldsymbol{x}_k) + H(\boldsymbol{x}_k)\boldsymbol{p}$$

روش نیوتن – حل معادلهٔ غیرخطی

$$\min_{x\in\mathbb{R}^n} f(\boldsymbol{x})$$

$$\nabla f(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}$$

نیاز به حل دستگاه n-معادلهٔ غیرخطی $\boldsymbol{g_i}(\boldsymbol{x}) = 0$

$$g(\boldsymbol{x}_k + \boldsymbol{p}) \approx g(\boldsymbol{x}_k) + H(\boldsymbol{x}_k)\boldsymbol{p} = \boldsymbol{0}$$

$$\boldsymbol{p} = H^{-\boldsymbol{1}}(\boldsymbol{x}_k) g(\boldsymbol{x}_k)$$

روش نیوتن –ادامه

$$\boldsymbol{p}_k = H^{\boldsymbol{-1}}(\boldsymbol{x}_k)g(\boldsymbol{x}_k)$$
$$f_p'(\boldsymbol{x}_k) = g_k^T\boldsymbol{p}_k = -g_k^T H_k^{-1} g_k < 0, g_k \neq 0$$

برآورده‌پذیر در صورت مثبت معین بودن H

در صورت مثبت معین نبودن H: افزودن قطر مثبت Δ_k به طوری که

$$H_k + \Delta_k > \epsilon I \Leftrightarrow \lambda_i(H_k) > \epsilon$$

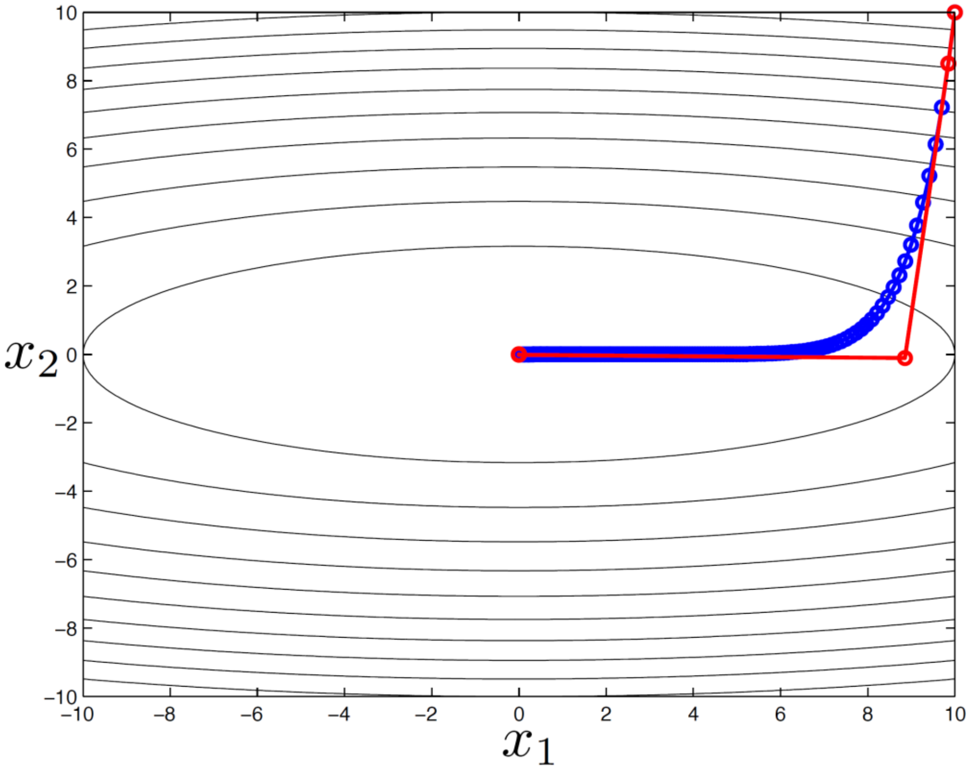
روش نیوتن تغییریافته

$$\boldsymbol{p}_k = (H_k + \Delta_k)^{\boldsymbol{-1}}\boldsymbol{g}_k$$

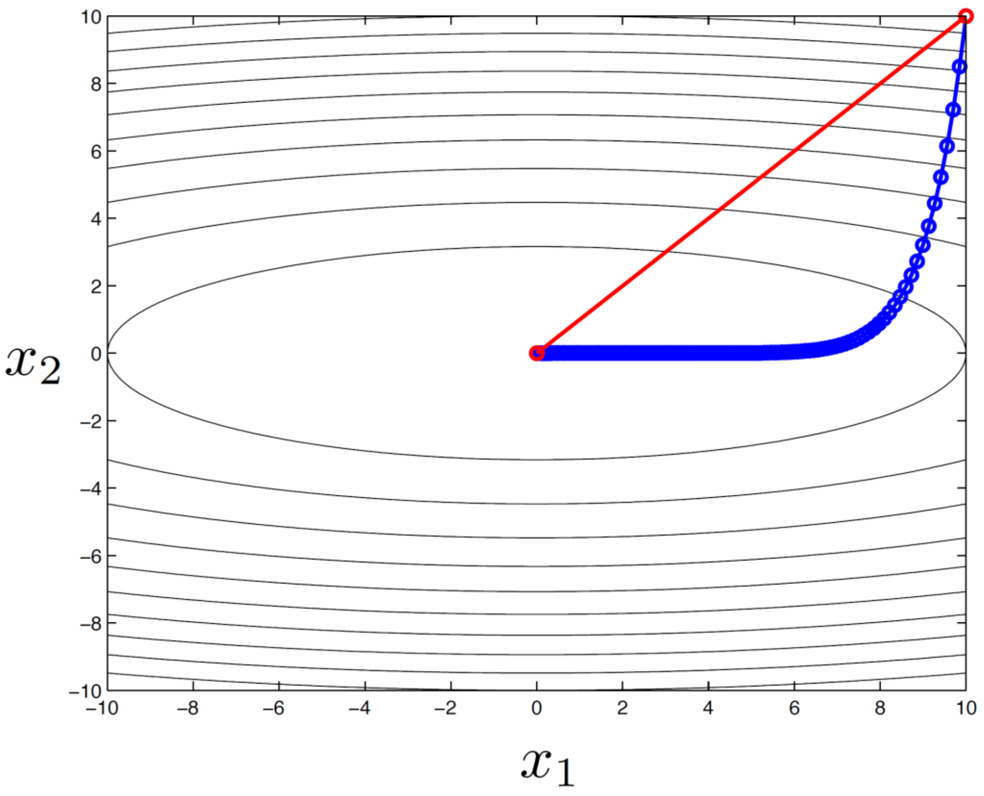
مثال- تندترین نزول در برابر گرادیان مزدوج

$$f(\mathbf{x}) = x_1^2 + 10x_2^2 \qquad \mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}, \mathbf{b} = 0$$

$$\alpha_i = 0.015$$



مثال



روش نیوتن - ادامه

مزایا

- بسیار سریع

معایب

- ماتریس هسی ممکن است مثبت معین نباشد
  - تغییر به گرادیان نزولی
  - به جای یافتن معکوس ماتریس هسی، حل Hy=g برای y
    - استفاده از $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \boldsymbol{y}$
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \alpha \boldsymbol{y}$
- بیشتر اوقات کار نمی‌کند
  - کار می‌کند وقتی که خیلی نزدیک کمینه باشد
  - پیشنهاد ترکیب گرادیان نزولی و سپس روش نیوتن

منابع

[نازهدل]

What are the differences between the different gradient-based numerical optimization methods?, https://scicomp.stackexchange.com/questions/26960/what-are-the-differences-between-the-different-gradient-based-numerical-optimiza

“An overview of gradient descent optimization algorithms”' https://ruder.io/optimizing-gradient-descent/, 2016