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

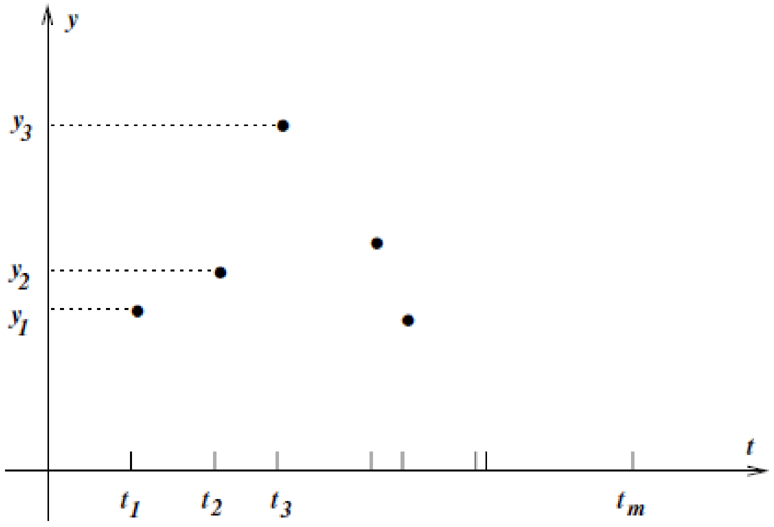
$$\underset{\boldsymbol{x}}{\text{کم}}\ f(\boldsymbol{x})$$

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

$n \geq ۱$

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

مقدمه – مثال



یافتن منحنی که چند داده را تقریب بزند

نمایش نمایی و نوسانی

- $\phi(t;x) = x_۱ + x_۲ e^{-(x_۳ - t)^۲ / x_۴} + x_۵ \cos(x_۶ t)$

x_i–ها ضرائب مدل

- مجهول‌ها
- $\boldsymbol{x} = (x_۱, x_۲, x_۳, x_۴, x_۵, x_۶)^T$

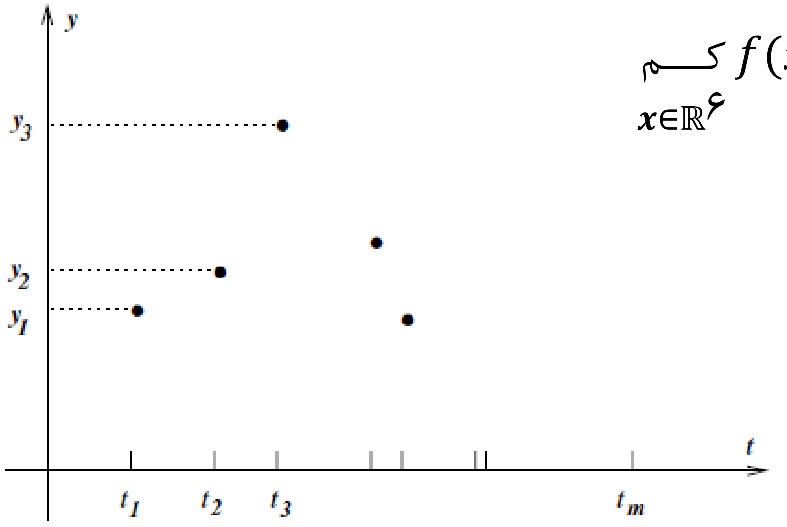
تعریف باقی‌مانده

- نمایشگر تفاوت بین مدل و داده مشاهده شده
- $r_j(\boldsymbol{x}) = y_j - \phi(t_j; \boldsymbol{x}), j = ۱, ۲, \ldots, m$

$$\underset{\boldsymbol{x} \in \mathbb{R}^۶}{\text{کم}}\ f(x) = r_۱^۲(\boldsymbol{x}) + r_۲^۲(\boldsymbol{x}) + \cdots + r_m^۲(\boldsymbol{x})$$

- مسئلهٔ برازش کمترین مربعات غیرخطی

مقدمه – مثال - ادامه



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

- مسئلهٔ برازش کمترین مربعات غیرخطی
  - از انواع بهینه‌سازی نامقید

فرض: پاسخ $\boldsymbol{x}^* = (1.1, 0.01, 1.2, 1.5, 2.0, 1.5)^T$

مقدار کمینه $\mathrm{f}(\boldsymbol{x}^*) = 0.34$

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 $f(x^*) \leq f(x)\ \forall x$
- 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 $\mathcal{H}$) $f(x^*) \leq f(x),\ x \in \mathcal{H}$
  - معروف به کمینه‌ساز محلی ضعیف
- کمینه‌ساز محلی اکید
  - $f(x^*) < f(x),\ x \in \mathcal{H}, x \neq x^*$

تشخیص کمینه محلی

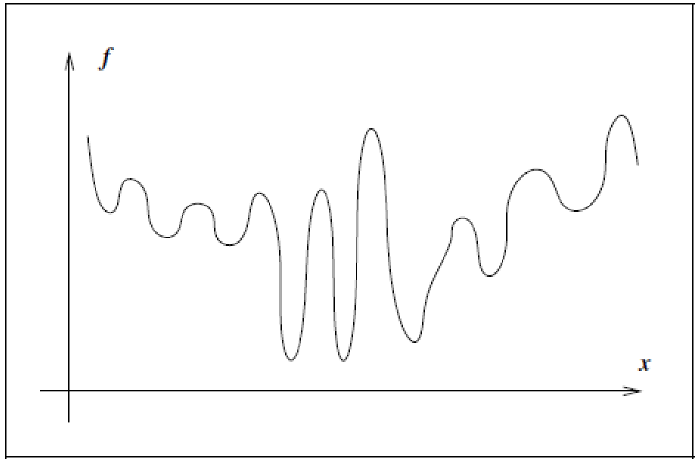
مثال

- $f(x) = ۲$

مثال ۲

- $f(x) = (x - ۲)^{۴}$

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

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

$t \in (0,۱)$.

هم‌چنین، اگر مشتق‌پذیر مرتبه دوم پیوسته باشد، آن‌گاه

$$\nabla f(\boldsymbol{x} + \boldsymbol{p}) = \nabla f(\boldsymbol{x}) + \int_0^۱ \nabla^۲ f(\boldsymbol{x} + t\boldsymbol{p})\boldsymbol{p}\,dt$$

$$f(\boldsymbol{x} + \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T\boldsymbol{p} + \frac{۱}{۲}\boldsymbol{p}^T \nabla^۲ f(\boldsymbol{x} + t\boldsymbol{p})\boldsymbol{p}$$

$t \in (0,۱)$

قضیهٔ شروط لازم مرتبهٔ اول

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

اثبات

برهان خلف- فرض $\nabla f(\boldsymbol{x}^*) \neq 0$

تعریف $\mathbf{p} = -\nabla f(\boldsymbol{x}^*)$ در نتیجه $\mathbf{p}^{\mathrm{T}} \nabla f(\boldsymbol{x}^*) = -\|\nabla f(\boldsymbol{x}^*)\|^{۲}$

به دلیل پویستگی گردایان حول نقطه کمینه،وجود $T > 0$ به طوری که
$\mathbf{p}^{\mathrm{T}} \nabla f(\boldsymbol{x}^* + t\boldsymbol{p}) < 0, t \in [0, T]$

استفاده از قضیهٔ تیلور
$f(\boldsymbol{x}^* + \bar{t}\boldsymbol{p}) = f(\boldsymbol{x}^*) + \bar{t}\mathbf{p}^{\mathrm{T}} \nabla f(\boldsymbol{x}^* + t\boldsymbol{p}) \Rightarrow f(\boldsymbol{x}^* + \bar{t}\boldsymbol{p}) < f(\boldsymbol{x}^*)$

قضیهٔ شروط لازم مرتبهٔ اول

$\boldsymbol{x}^*$ کمینه‌ساز 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^* کمینه‌ساز محلی f و $\nabla^۲ f$ موجود و پیوسته همسایهٔ نقطهٔ مذکور، آن‌گاه $\nabla f(x^*) = 0$ و $\nabla^۲ f(x^*)$ مثبت نیمه‌معین است.

اثبات برهان خلف مشتق دوم مثبت معین نباشد.

$$f\ (\boldsymbol{x}\ + \bar{t}\boldsymbol{p}) = f\ (\boldsymbol{x}) + \bar{t}\nabla f(\boldsymbol{x})^T\boldsymbol{p} + \frac{۱}{۲}\bar{t}^۲\boldsymbol{p}^T\nabla^۲ f(\boldsymbol{x} + t\boldsymbol{p})\boldsymbol{p}$$

قضیهٔ شروط کافی مرتبهٔ دوم

$\nabla^{۲} f$ 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- تمرین

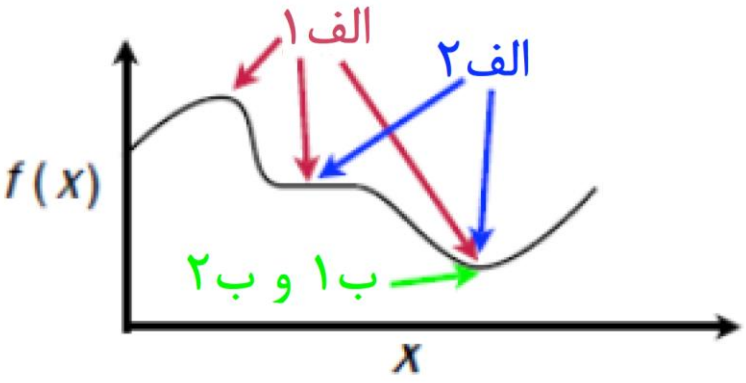
شرایط بهینگی کمینه‌سازی چندمتغیره

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

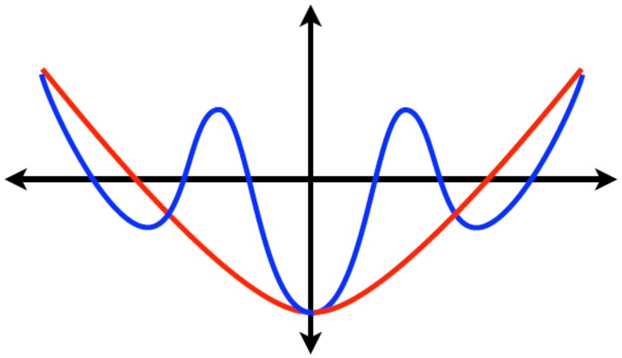
f کوژ باشد، هر x^* کمینه‌ساز محلی کمینه‌ساز سراسری f است. هم‌چنین اگر f مشتق‌پذیر باشد، آن‌گاه هر نقطه مانای x^* کمینه‌ساز سراسری f خواهد بود.

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

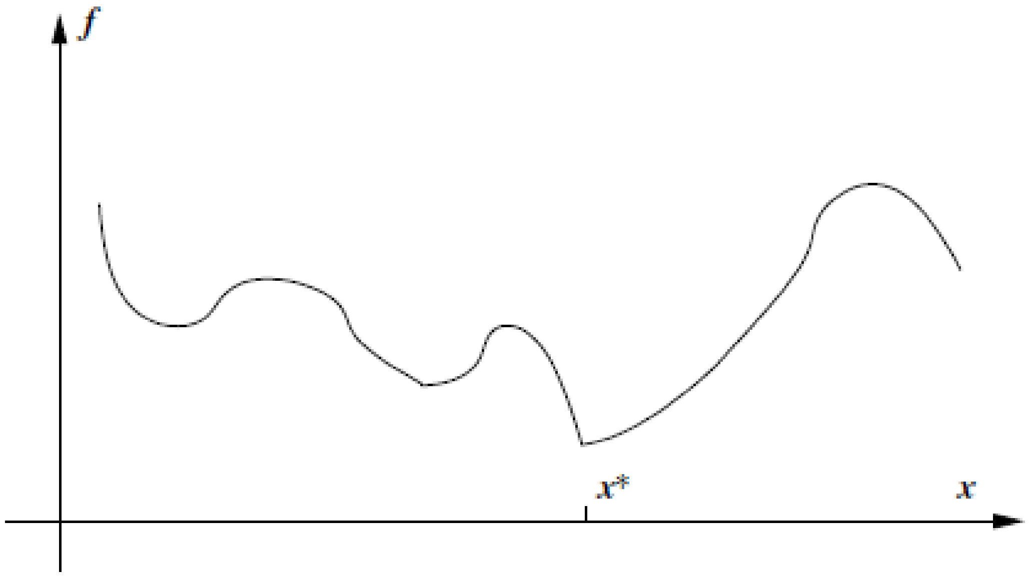
توابع ناهموار و ناپیوسته

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 $f\ (x_i)\ <\ f\ (x_{i-m})$

دو استراتژی اساسی جهت حرکت از نقطه فعلی x_i به نقطهٔ x_{i+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 $i = 0$

تازمان همگرا نشدن

{

انتخاب جهت $\boldsymbol{p}_i$ و اندازه قدم α_i

$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \alpha_i \boldsymbol{p}_i$

$i \leftarrow i + 1$

}

دو استراتژی- جستجو خط و منطقه اعتماد

استراتژی جستجو خط

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

$$\underset{\alpha>0}{\text{کم}} f(\boldsymbol{x_i} + \alpha \boldsymbol{p}_i)$$

استراتژی منطقه اعتماد

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

$$\underset{\boldsymbol{p}}{\text{کم}} \boldsymbol{m_i}(\boldsymbol{x}_i + \boldsymbol{p})$$

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

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

دو استراتژی – ادامه

استراتژی منطقه اعتماد

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

$$\boldsymbol{m}_i(\boldsymbol{x}_i + \boldsymbol{p}) = f_i + \boldsymbol{p}^T \nabla f_i + \frac{١}{٢} p^T B_i p$$

B_i یا تابع هسی $\nabla^2 f_i$ یا تخمینی از آن

سیاست نخست- جستجو خط

انتخاب $\boldsymbol{p}_i$ 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 مرتبهٔ دوم، همگرائی ابرخطی
- با حل مسئله کمینه‌سازی یک‌بعدی زیر جهت یافتن α

انتخاب طول قدم α_i برآورده‌گر شروط وولف

- کروشه‌گذاری: یافتن بازه‌ای شامل طول قدم مناسب
- نیمه‌سازی/درون‌یابی: محاسبهٔ قدم مناسب در بازهٔ فعلی

شدیدترین نزول

شبیه پائین‌آمدن از کوه

انتخاب سریع‌ترین جهت سرازیری

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

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 - ادامه

حل با استفاده از شدیدترین نزول

$$f(x) = x^۲ - ۲x + ۲ = (x - ۱)^۲ + ۱$$

$$\underset{x}{\text{کمینه}}\, f(x)$$

مقدار بهینه را نمی‌دانیم

- شروع از مقدار تصادفی $x = ۳$

قدم اول: مشتق‌گیری

$$\frac{df}{dx} = ۲\mathrm{x} - ۲$$

شدیدترین نزول- مثال - ادامه

قدم دوم

- مطالعۀ مشتق در نقطۀ داده شده

$$f'(۳) = ۲(۳) - ۲ = ۴$$

- مشتق در کمینه باید صفر باشد
- مقدار مثبت مشتق
  - نمایشگر اینکه مقدار تابع افزایشی است. و باید عقب رفت
- اگر مقدار $x = -۱$ به عنوان حدس اولیه انتخاب می‌شد
  - آن‌گاه مشتق $f' = -۴$
  - 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

$$x_{i+۱} = x_i + \alpha(-f'(x_i))$$

$x_{i+۱}$ حدس بعدی

طول قدم $\alpha = 0.۲$

$x_0 = ۳$

$$x_{i+۱} = x_i + 0.۲(-f'(x_i))$$
$$x_۱ = ۳ - \alpha f'(۳) = ۳ + 0.۲(-۴) = ۲.۲$$
$$x_۲ = ۲.۲ - \alpha f'(۲.۲) = ۲.۲ + 0.۲(-۲.۴) = ۱.۷۲$$

ادامه محتملا به پاسخ می‌رسد.

- ?

چرا شدیدترین نزول به جای روش تحلیلی و حسابان

نوشتن برنامه

شدیدترین نزول- مثال ۲

x	y
۱	۱
۲	۱
۲	۲
۳	۲

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

$$e_۱ = \hat{y}_۱ - y_۱$$
$$e_۲ = \hat{y}_۲ - y_۲$$

نیاز به کل خطا

- یک روش:جمع تمامی مقادیر خطا $e_{کل} = e_۱ + e_۲ + e_۳ + e_۴$
  - غلطانداز
- روش بهتر $e_{کل} = e_۱^۲ + e_۲^۲ + e_۳^۲ + e_۴^۲$

شدیدترین نزول - مثال ۲ - ادامه

x	y
۱	۱
۲	۱
۲	۲
۳	۲

روش بهتر $e_{\text{کل}} = e_۱^۲ + e_۲^۲ + e_۳^۲ + e_۴^۲$

$$e_۱ = \hat{y}_۱ - y_۱ \Rightarrow e_۱ = [\alpha x_۱ + \beta] - y_۱$$
$$e_۲ = \hat{y}_۲ - y_۲ \Rightarrow e_۲ = [\alpha x_۲ + \beta] - y_۲$$
$$e_۳ = \hat{y}_۳ - y_۳ \Rightarrow e_۳ = [\alpha x_۳ + \beta] - y_۳$$
$$e_۴ = \hat{y}_۴ - y_۴ \Rightarrow e_۴ = [\alpha x_۴ + \beta] - y_۴$$

خطای کل

$$e_{\text{کل}} = \sum_{i=۱}^{۴} e_i^۲$$

$$= \sum_{i=۱}^{۴} ([\alpha x_i + \beta] - y_i)^۲$$

شدیدترین نزول- مثال ۲ – ادامه

x	y
۱	۱
۲	۱
۲	۲
۳	۲

به دنبال یافتن خطی با کمترین خطای ممکن

$$e_{\text{کل}} = \sum_{i=۱}^{۴} e_i^۲ = \sum_{i=۱}^{۴} \left([\alpha x_i + \beta] - y_i \right)^۲$$

$$\underset{\alpha,\beta}{\text{کـــــــم}} \sum_{i=۱}^{۴} \left([\alpha x_i + \beta] - y_i \right)^۲$$

استفاده از شدیدترین نزول

$$\begin{bmatrix} \alpha_{\text{ب}} \\ \beta_{\text{ب}} \end{bmatrix} = \begin{bmatrix} \alpha_{\text{ق}} \\ \beta_{\text{ق}} \end{bmatrix} - 0.۲ \begin{bmatrix} \frac{\partial f}{\partial \alpha}(\alpha_{\text{ق}}, \beta_{\text{ق}}) \\ \frac{\partial f}{\partial \beta}(\alpha_{\text{ق}}, \beta_{\text{ق}}) \end{bmatrix}$$

شدیدترین نزول- مثال ۲ – ادامه

x	y
۱	۱
۲	۱
۲	۲
۳	۲

$$\begin{bmatrix}\alpha_{\text{ب}}\\ \beta_{\text{ب}}\end{bmatrix} = \begin{bmatrix}\alpha_{\text{ق}}\\ \beta_{\text{ق}}\end{bmatrix} - 0.۲\begin{bmatrix}\sum_{i=۱}^{۴} ۲([\alpha x_i + \beta] - y_i)x_i\\ \sum_{i=۱}^{۴} ۲([\alpha x_i + \beta] - y_i)\end{bmatrix}$$

مقدار اولیه

$$\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix} = \begin{bmatrix}۱\\ ۱\end{bmatrix}$$

شدیدترین نزول- مثال ۲ - ادامه

x	y
۱	۱
۲	۱
۲	۲
۳	۲

تحریر محل نزاع

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

$$\begin{bmatrix} \alpha_{\text{ب}} \\ \beta_{\text{ب}} \end{bmatrix} = \begin{bmatrix} \alpha_{\text{ق}} \\ \beta_{\text{ق}} \end{bmatrix} - 0.۲ \begin{bmatrix} \sum_{i=۱}^{۴} ۲(\,[\alpha x_i + \beta] - y_i)x_i \\ \sum_{i=۱}^{۴} ۲(\,[\alpha x_i + \beta] - y_i) \end{bmatrix}$$

شدیدترین نزول- مثال ۲ - ادامه

راه‌حل

- گرادیان نزولی تصادفی

$$\begin{bmatrix}\alpha_{j+1}\\ \beta_{j+1}\end{bmatrix} = \begin{bmatrix}\alpha_{j}\\ \beta_{j}\end{bmatrix} - 0.۲\begin{bmatrix}2(\,[\alpha_j x_۱ + \beta_j] - y_1)x_1\\ 2(\,[\alpha_j x_1 + \beta_j] - y_1)\end{bmatrix}$$

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

- 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

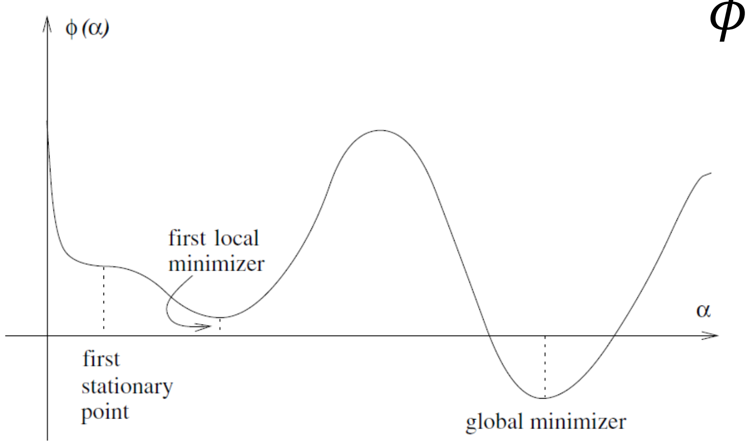
طول قدم

- 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

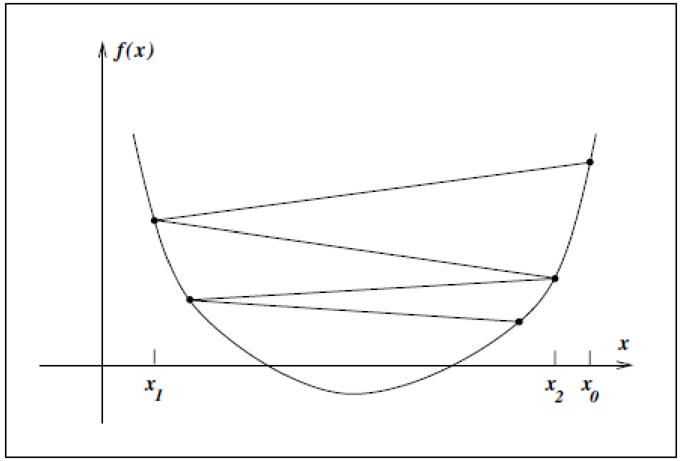


طول قدم α

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



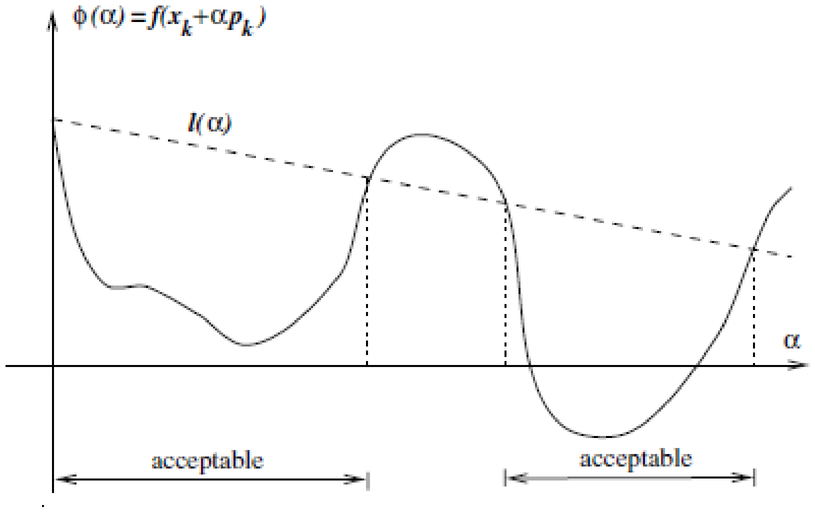
سادهترین شرط

- 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 $\nabla f_i^T \boldsymbol{p}_i$

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

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

اما «شرط کافی» مکفی نیست!

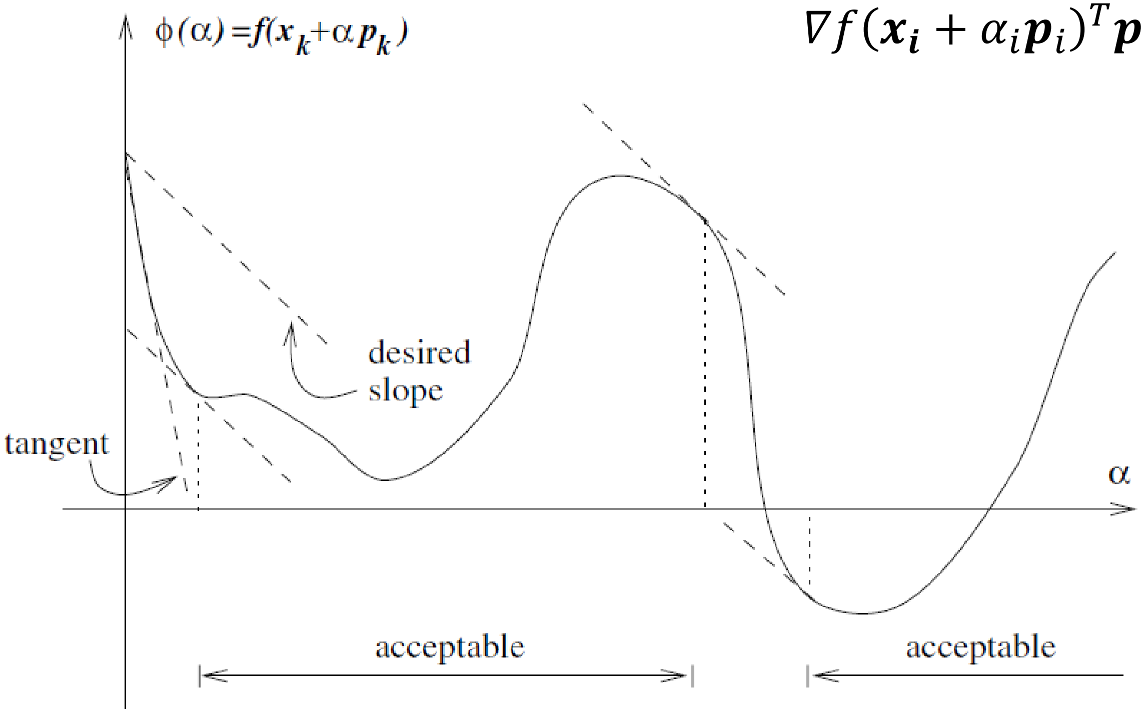
شروط وولف

جهت حذف مقادیر اقدام کوچک

- شرط انحنا

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

- $c_2 \in (c_1, 1)$
- $\phi'(\alpha_i) = \nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i$
  - 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

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

- $c_2 \in (c_1, 1)$
- $\phi'(\alpha_i) = \nabla f(\boldsymbol{x_i} + \alpha_i \boldsymbol{p}_i)^T \boldsymbol{p}_i$
  - 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 وولف

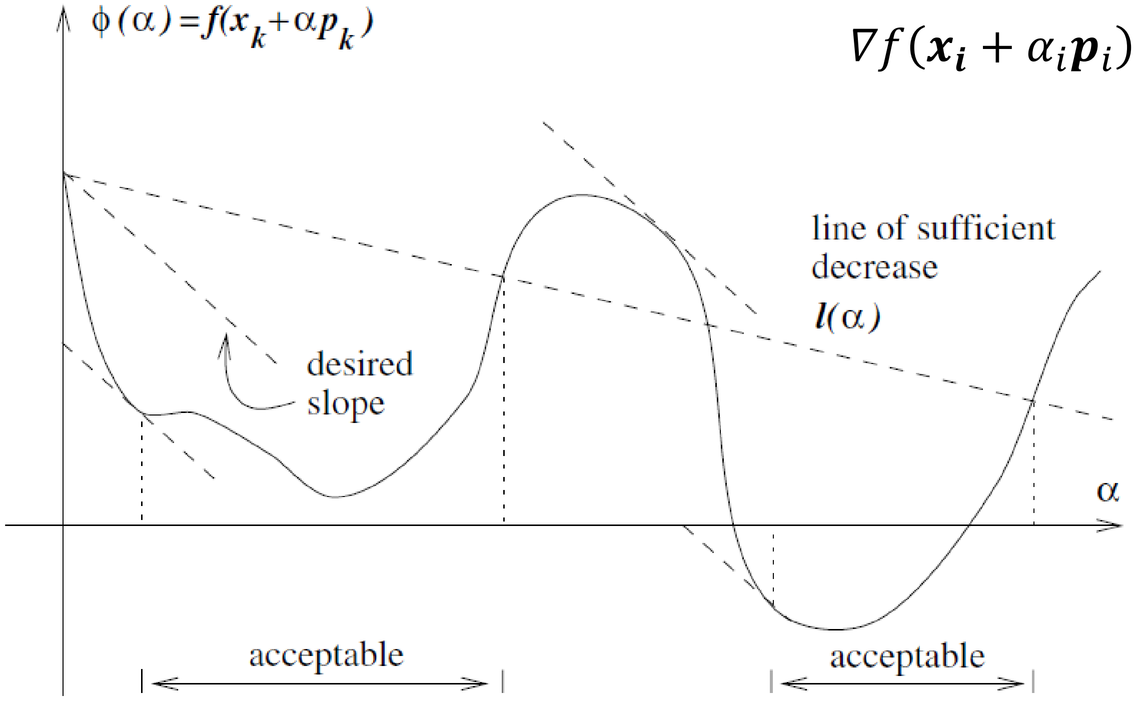
- شرط ارمیخو

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

- شرط انحنا

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

- $0 < c_1 < c_2 < 1$



شروط قوی وولف

امکان وجود شرایط وولف بدون اینکه طول مناسبی باشد

شروط قوی وولف

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

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

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

شروط گلدشتین

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

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

$$0 < c < 0.5$$

همان شرط کاهش کافی

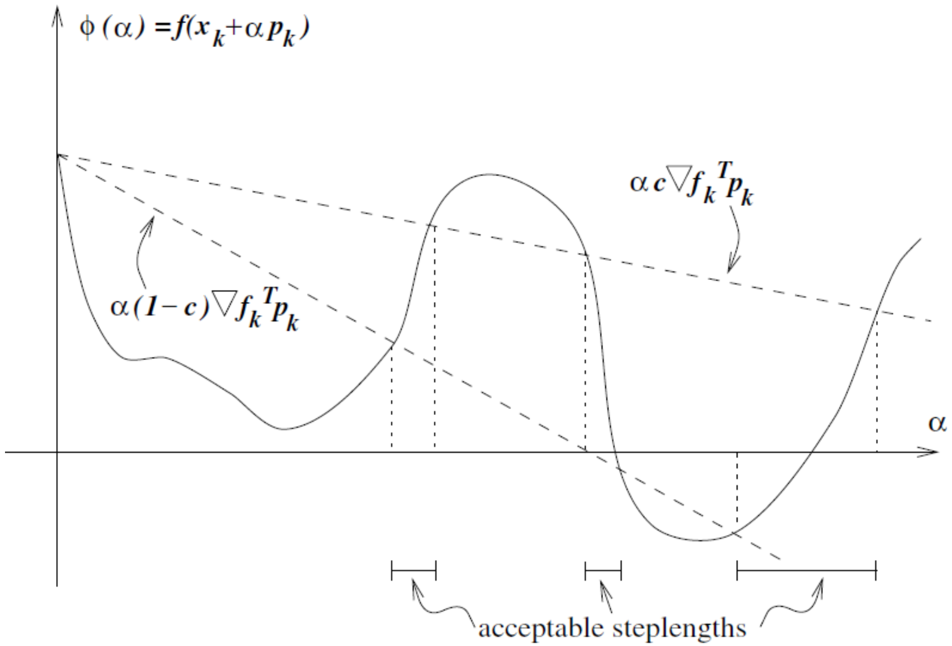
شروط گلدشتین

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

کاهش کافی با پس‌روی

صرفا استفاده از شرط کافی و نه شرط دوم

الگوریتم جستجو خط با پس‌روی

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

روش نیوتن و شبه‌نیوتن

- $\bar{\alpha} = 1$

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

$s = 0$

تازمان همگرا نشدن

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

$$s = s + ||\nabla f(\boldsymbol{x}_i)||^۲$$

$$\boldsymbol{x}_{i+۱} = \boldsymbol{x}_i + \frac{\alpha}{\sqrt{s+\epsilon}}\boldsymbol{g}$$

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
- ار-ام-اس-پراپ RMSPROP
  - جلوگیری از کوچک‌شدن بی‌اندازهٔ مقدار طول قدم
  - کاهش طول و استفاده از قبلی‌ها
  - با استفاده از پنجره از چند مقدار قبلی

$$\boldsymbol{s} = 0$$

تازمان همگرا نشدن

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

$$\boldsymbol{s} = \gamma \boldsymbol{s} + (۱ - \gamma)\nabla f(\boldsymbol{x}_i)^۲$$

$$\boldsymbol{x}_{i+۱} = \boldsymbol{x}_i + \alpha \frac{\boldsymbol{g}}{\sqrt{\boldsymbol{s}+\epsilon}}$$

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

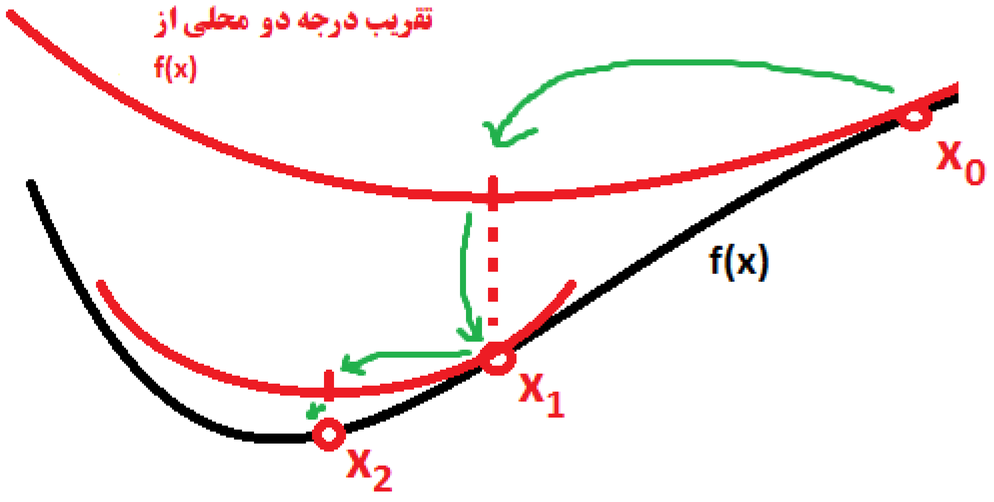
روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

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

روشی مبتنی بر مرتبه دوم

- ماتریس هسی



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

روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

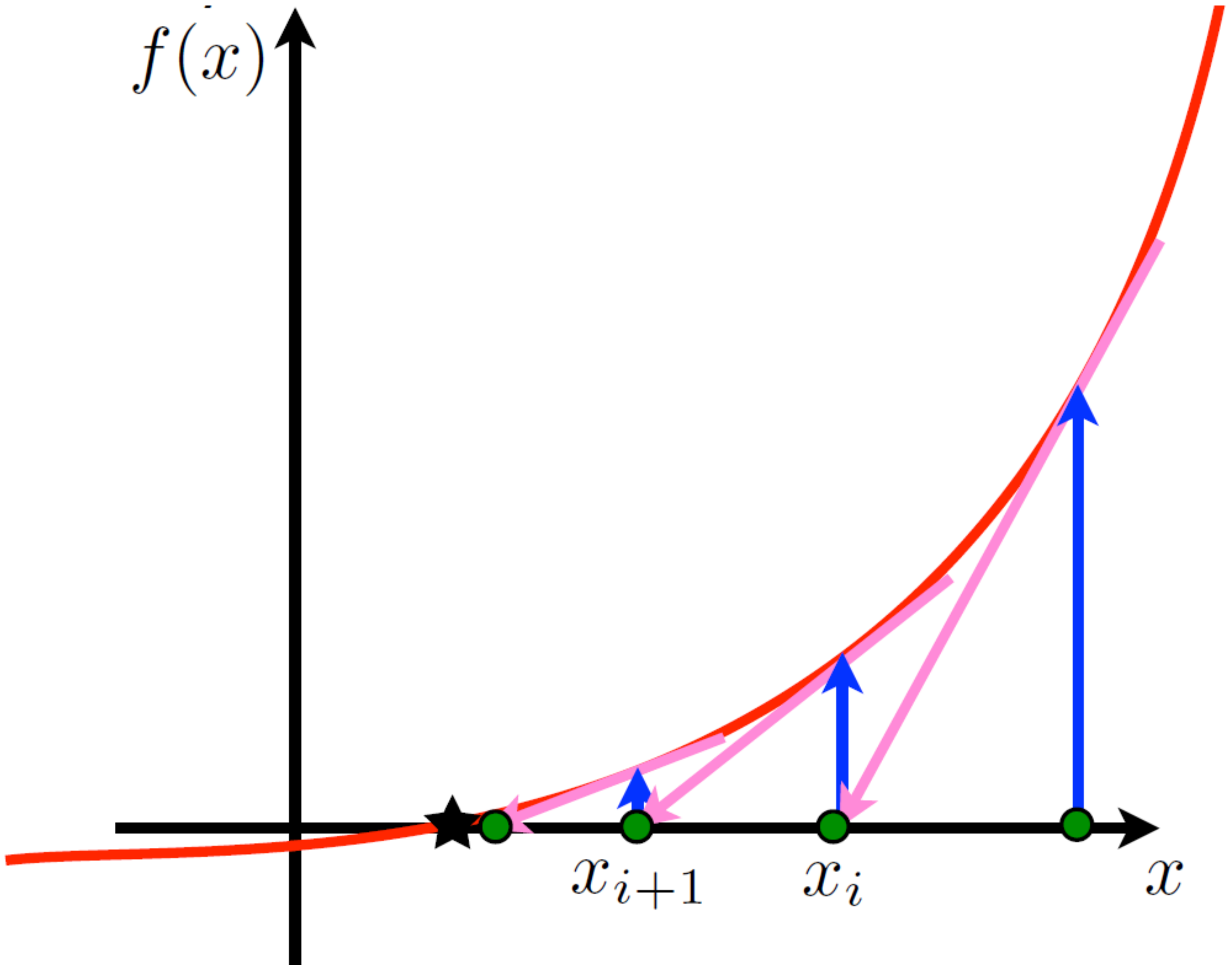
گن $\Leftarrow$ $x_{i+1} = x_i - \alpha_i\big(\nabla f(x_i)\big)$

روشی مبتنی بر مرتبه دوم (استفاده از $g_k = \nabla f(\boldsymbol{x}_k)$ و $H_k = \nabla^2 f(\boldsymbol{x}_k)$)

- ماتریس هسی

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

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



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

روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

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

روشی مبتنی بر مرتبه دوم

- ماتریس هسی

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

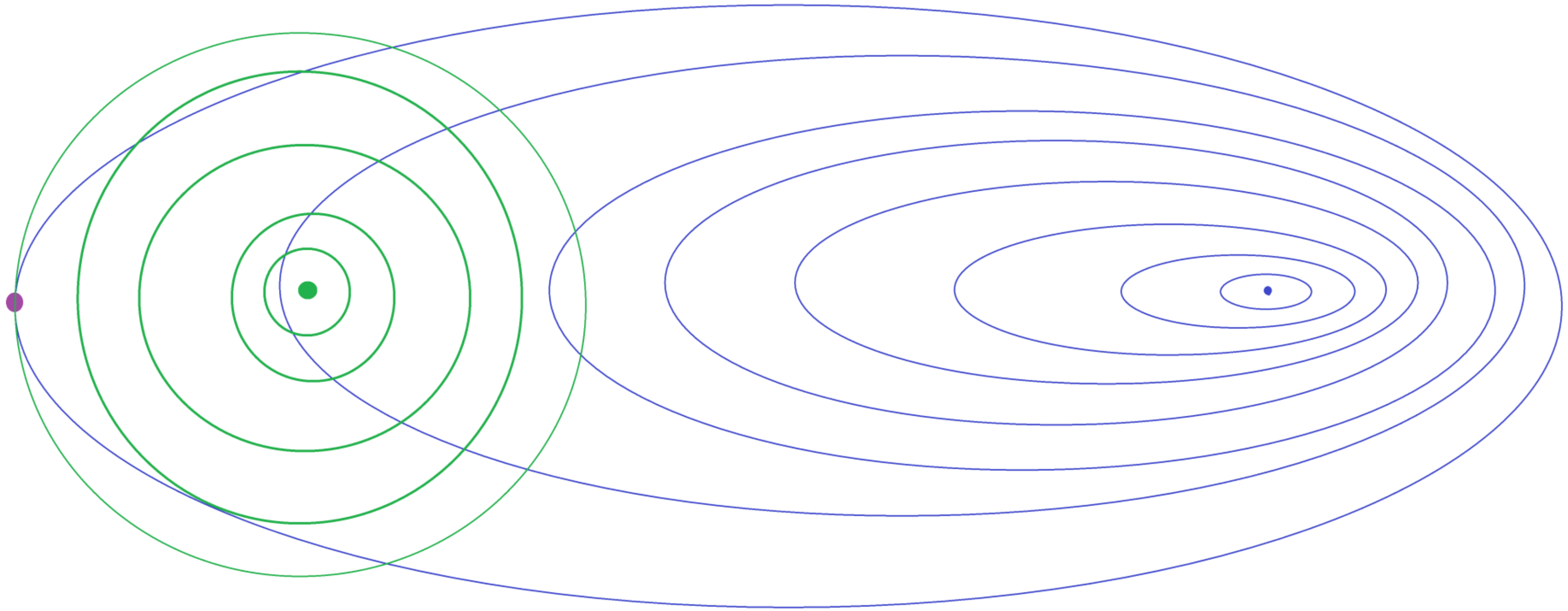
$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

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

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

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

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

$$\boldsymbol{p_k} = -\boldsymbol{H}_{\boldsymbol{k}}^{-\boldsymbol{1}}\boldsymbol{g_k}$$

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p_k}$ 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 $\boldsymbol{g_i}(\boldsymbol{x}) = 0$

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

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

روش نیوتن – ادامه

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

برآورده‌پذیر در صورت مثبت معین بودن H

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

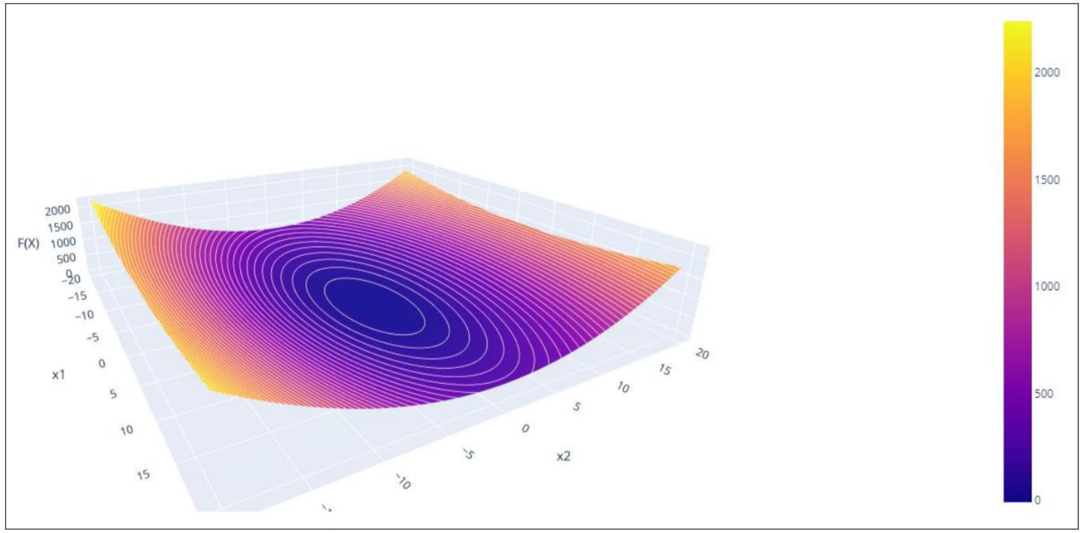
$$H_k + \Delta_k > \epsilon I \Leftrightarrow \lambda_i(H_k) > \epsilon$$

روش نیوتن تغییریافته

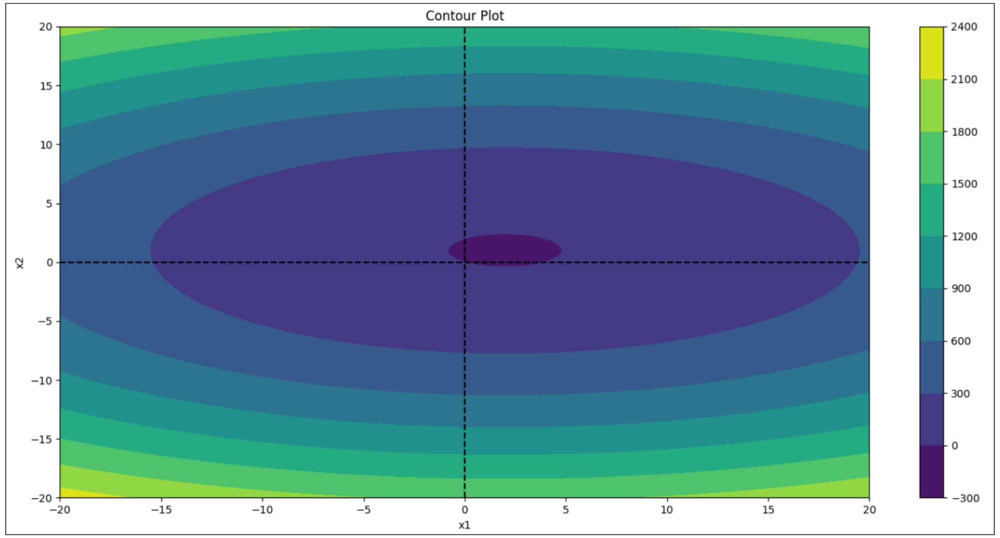
$$\boldsymbol{p}_k = (H_k + \Delta_k)^{-\mathbf{1}}\boldsymbol{g}_k$$

روش تندترین نزول در مقابل روش نیوتن

$$f(\boldsymbol{x}) = x_۱^۲ + ۴x_۲^۲ - ۴x_۱ - ۸x_۲$$



شکل-۱. نمایش ۳-بعدی تابع f به همراه سطح ترازهای آن.

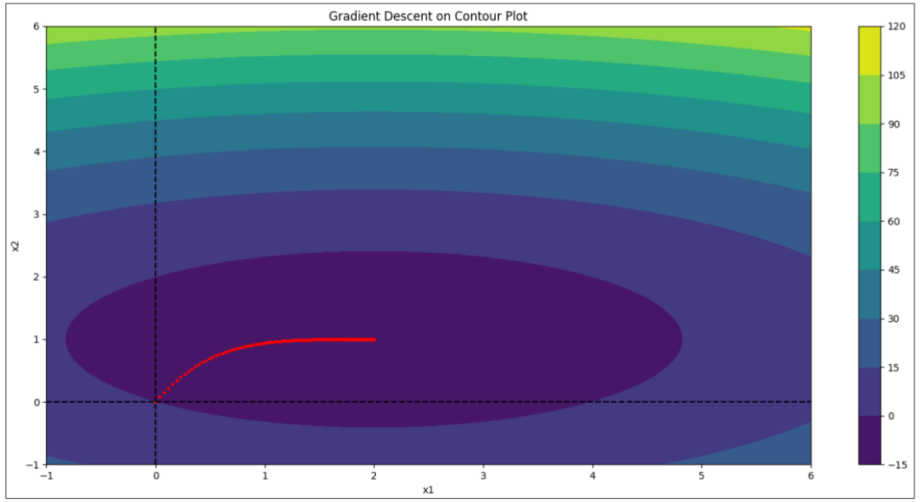


شکل-۲. رسم سطح ترازهای تابع f .

روش تندترین نزول در مقابل روش نیوتن

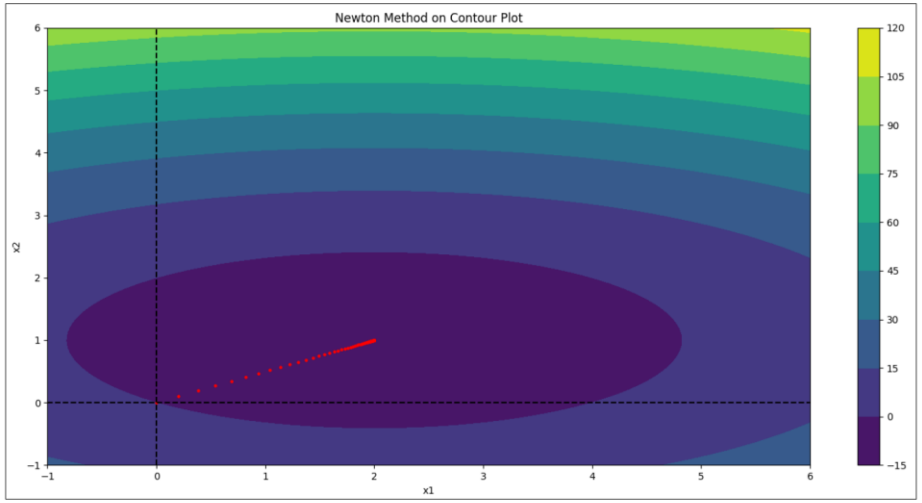
$$f(\boldsymbol{x}) = x_۱^۲ + ۴x_۲^۲ - ۴x_۱ - ۸x_۲$$

با شروع از $\boldsymbol{x_0} = [0,0]^T$ و سرعت یادگیری (طول قدم) $\alpha_i = 0.۱$



شکل-۳. نتیجهٔ اجرای الگوریتم تندترین نزول روی تابع f. نقاط قرمز رنگ کوچکی که در تصویر قابل مشاهده است توسط الگوریتم تندترین بدست آمده است.

تندترین نزول با ۱۶۳۶ قدم

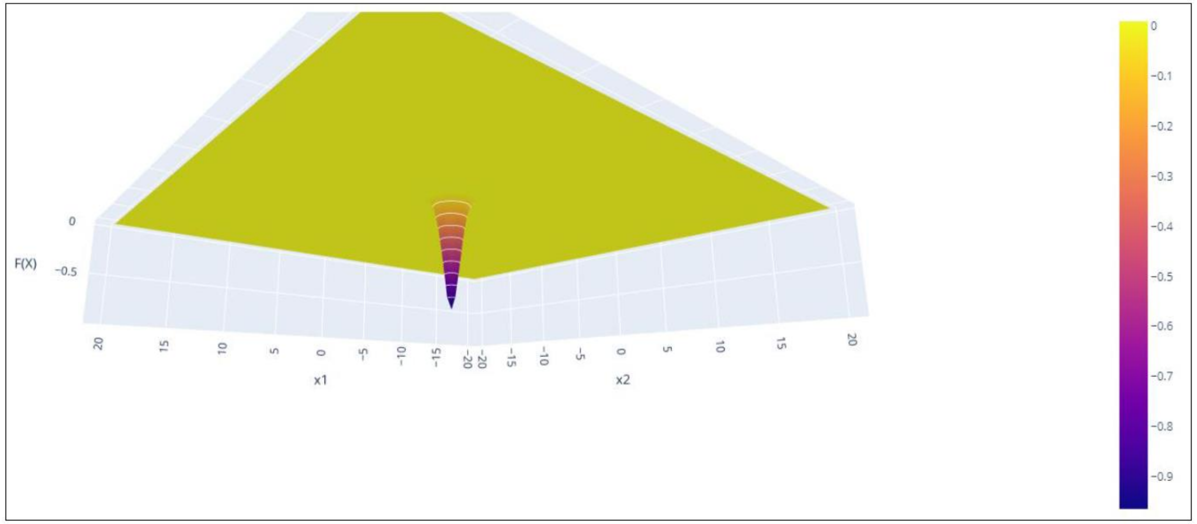


شکل-۴. نتیجهٔ اجرای الگوریتم نیوتن روی تابع f. نقاط قرمز رنگ کوچکی که در تصویر قابل مشاهده است توسط الگوریتم نیوتن بدست آمده است.

نیوتن با ۳۲۴ قدم

روش تندترین نزول در مقابل روش نیوتن

$$f(\boldsymbol{x}) = -\cos x_1 \cos x_2 \, e^{[-(x_1-\pi)^2-(x_2-\pi)^2]}$$

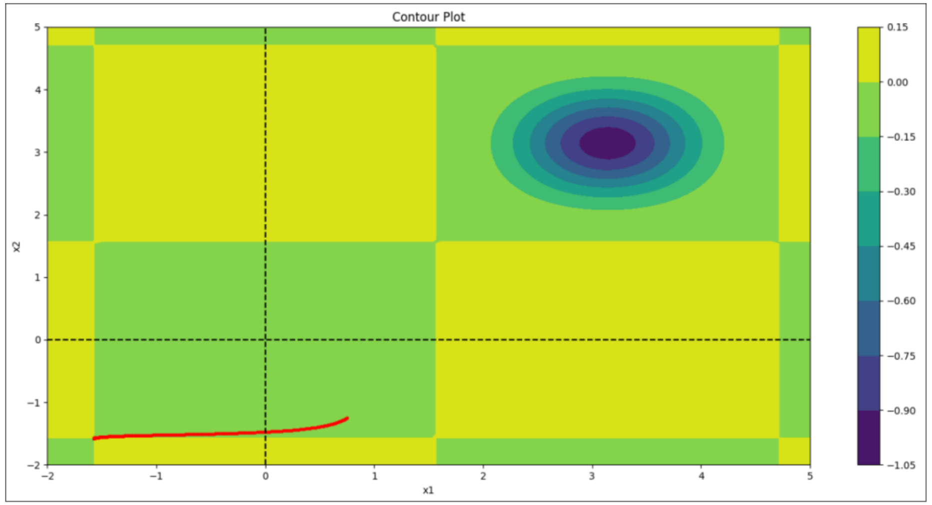


شکل-۵. نمایش ۳-بعدی تابع f به همراه سطح ترازهای آن.

روش تندترین نزول در مقابل روش نیوتن

$$f(\boldsymbol{x}) = -\cos x_۱ \cos x_۲\, e^{\left[-(x_۱-\pi)^۲-(x_۲-\pi)^۲\right]}$$

با شروع از $\boldsymbol{x_0} = [0.۷۵, -۱.۲۵]^T$ و سرعت یادگیری (طول قدم) $\alpha_i = 0.۱$

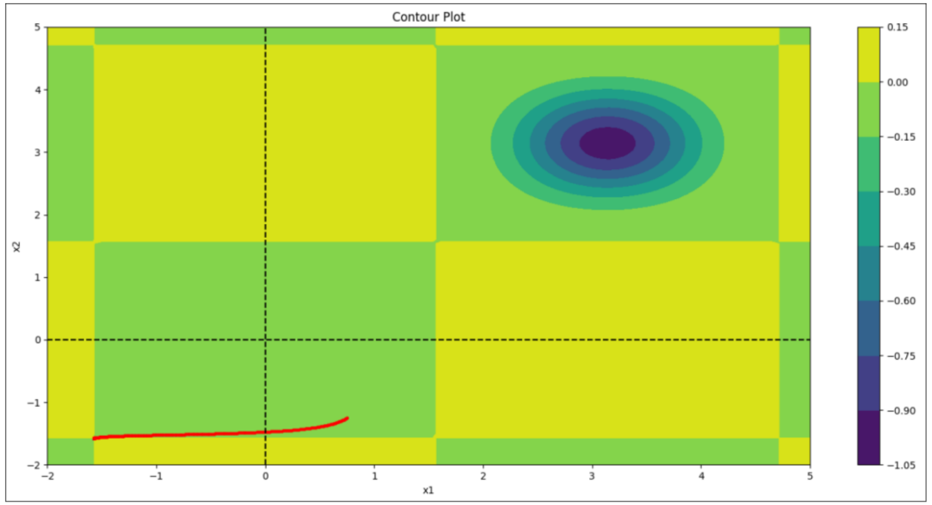


شکل-٦. نتیجهٔ اجرای الگوریتم نیوتن روی تابع f. نقاط قرمز رنگ کوچکی که در تصویر قابل مشاهده است توسط الگوریتم نیوتن بدست آمده است.

روش تندترین نزول در مقابل روش نیوتن

$$f(\boldsymbol{x}) = -\cos x_{۱} \cos x_{۲}\, e^{\left[-(x_{۱}-\pi)^{۲}-(x_{۲}-\pi)^{۲}\right]}$$

با شروع از $\boldsymbol{x_0} = [0.۷۵, -۱.۲۵]^T$ و سرعت یادگیری (طول قدم) $\alpha_i = 0.۱$



شکل-۶. نتیجهٔ اجرای الگوریتم نیوتن روی تابع f . نقاط قرمز رنگ کوچکی که در تصویر قابل مشاهده است توسط الگوریتم نیوتن بدست آمده است.

همگرائی نیوتن با ۵۳۶ قدم، اما نرسیدن به کمینه سراسری

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 کردن روش هنگام نزدیکی به کمینه
  - 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