



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 کمینه تابع هدف

مبتنی بر متغیرهای حقیقی بدون قید

از تمامی فضای اعداد حقیقی

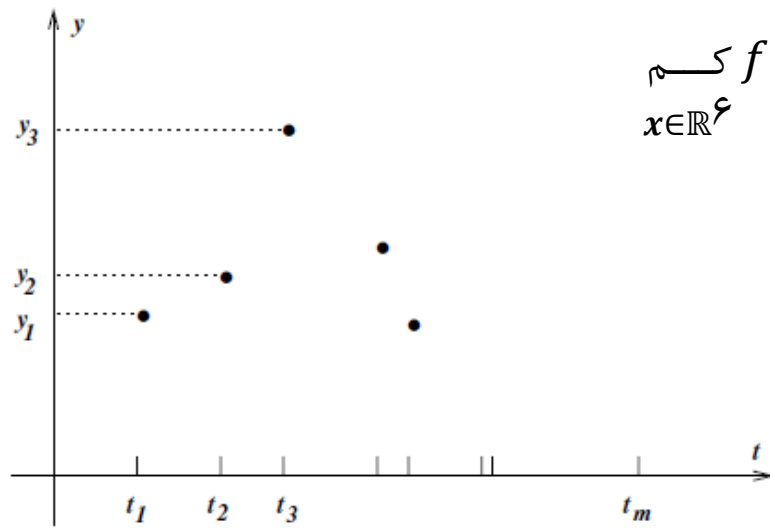
$$\min_x f(x)$$

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

$$n \geq 1$$

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

# مقدمه - مثال - ادامه



$$f(x) = r_1^2(x) + r_2^2(x) + \dots + r_m^2(x)$$

$x \in \mathbb{R}^6$

- مسئله برازش کمترین مربعات غیرخطی
- از انواع بهینه‌سازی نامقید

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

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  $f$

- نقطه بدست‌دهنده کمترین مقدار تابع  $f(x^*) \leq f(x) \forall x$
- مشکل در یافتن
- اطلاع صرفاً از دور و بر هر نقطه و نه بیشتر
- عدم اطمینان از وجود شکاف عمیق در داده
- بیشتر الگوریتم‌ها وابنده 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

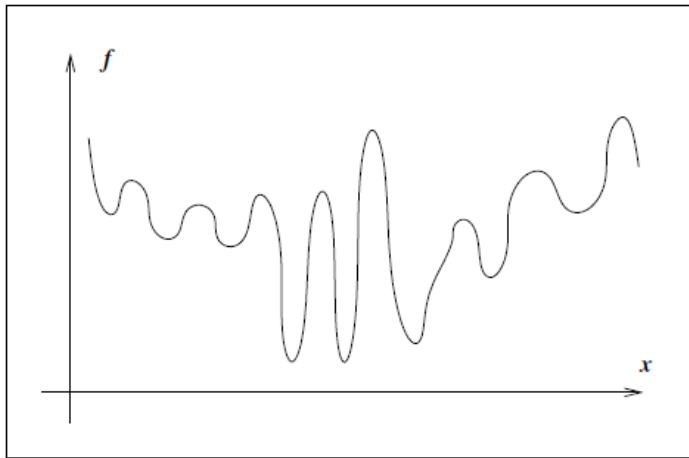
مثال

$$f(x) = 2 \cdot$$

مثال ۲

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

کمینه‌ساز محلی تک‌افتاده



# یافتن کمینه محلی

جستجو همسایگی؟

تابع هموار

▪ روش‌های کاراتر و عملی‌تر

تابع مشتق‌پذیر مرتبه اول و دوم

▪ وجود گرادیان و هسی

▪ محتملاً بتوان نقطه‌ای را کمینه محلی (کمینه محلی قوی) نامید

▪ ابزار ریاضی مطالعه کمینه‌سازها

▪ قضیه تیلور

قضیه تیلور

# قضیه تیلور

$f: \mathbb{R}^n \rightarrow \mathbb{R}$  تابعی مشتق پذیر پیوسته و  $\mathbf{p} \in \mathbb{R}^n$ ، آن گاه

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

$$t \in (0, 1)$$

همچنین، اگر مشتق پذیر مرتبه دوم پیوسته باشد، آن گاه

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

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

$$t \in (0, 1)$$

# قضیه شروط لازم مرتبه اول

$\mathbf{x}^*$  کمینه‌ساز محلی و  $f$  در همسایگی نقطه مذکور مشتق پذیر و پیوسته است، آن گاه  $\nabla f(\mathbf{x}^*) = 0$   
اثبات

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

تعریف  $\mathbf{p} = -\nabla f(\mathbf{x}^*)$  در نتیجه  $\mathbf{p}^T \nabla f(\mathbf{x}^*) = -\|\nabla f(\mathbf{x}^*)\|^2$

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

استفاده از قضیه تیلور

$$f(\mathbf{x}^* + \bar{t}\mathbf{p}) = f(\mathbf{x}^*) + \bar{t}\mathbf{p}^T \nabla f(\mathbf{x}^* + t\mathbf{p}) \Rightarrow f(\mathbf{x}^* + \bar{t}\mathbf{p}) < f(\mathbf{x}^*)$$

# قضیه شروط لازم مرتبه اول

$x^*$  کمینه‌ساز محلی و  $f$  در همسایگی نقطه مذکور مشتق پذیر و پیوسته است، آن گاه  $\nabla f(x^*) = 0$

اثبات

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

نتیجه قضیه: هر کمینه محلی، نقطه ماناست.

# قضیه شروط لازم مرتبه دوم

$x^*$  کمینه‌ساز محلی  $f$  و  $\nabla^2 f$  موجود و پیوسته همسایه نقطه مذکور، آن‌گاه  $\nabla f(x^*) = 0$  و  $\nabla^2 f(x^*)$  مثبت نیمه‌معین است.

اثبات برهان خلف مشتق دوم مثبت معین نباشد.

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

# قضیه شروط کافی مرتبه دوم

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

اثبات - تمرین

# شرایط بهینگی کمینه‌سازی چندمتغیره

قضیهٔ شروط لازم کمینه محلی ضعیف

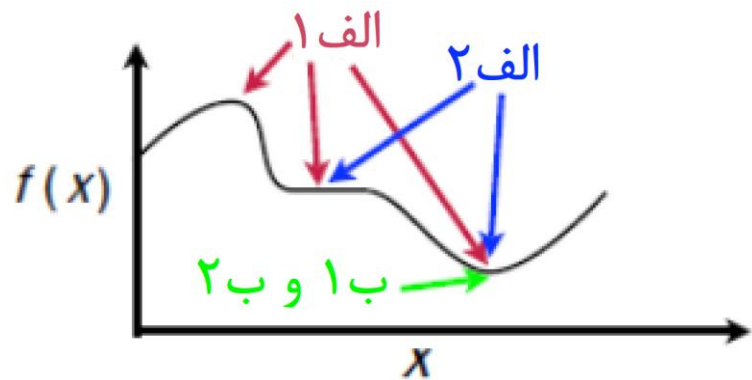
الف ۱:  $\nabla f(x^*) = 0$  نقطه مانا

الف ۲:  $\nabla^2 f(x^*)$  مثبت نیمه‌معین

قضیهٔ شروط کافی کمینه محلی قوی

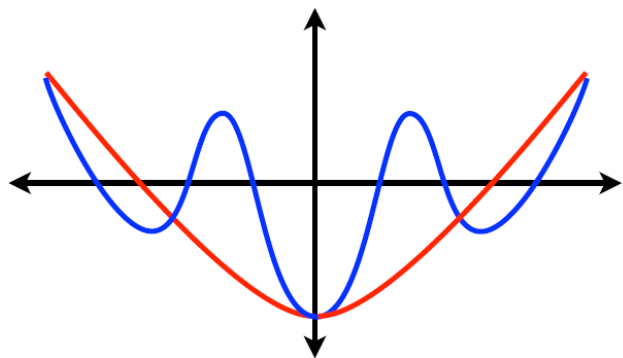
ب ۱:  $\nabla f(x^*) = 0$

ب ۲:  $\nabla^2 f(x^*)$  مثبت معین



# قضیه 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

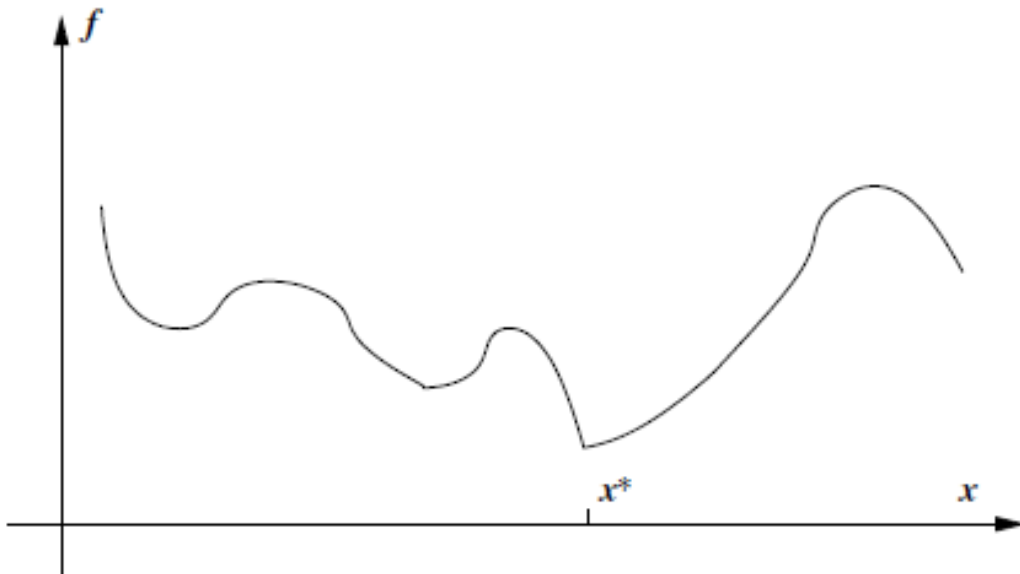
# مسائل ناهموار

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 نقطه 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 نقطه  $x_{i+1}$

- جستجو خط

- منطقه اعتماد

انتخاب جهت حرکت و اندازه حرکت

# الگوریتم‌های بهینه‌سازی نامقید

روش‌های گرادیان-محور  
▪ انتخاب جهت حرکت و اندازه حرکت

الگوریتم گرادیان-محور

انتخاب مقدار اولیه  $x_0$  و  $i = 0$   
تازمان همگرا نشدن  
}

انتخاب جهت  $p_i$  و اندازه قدم  $\alpha_i$   
$$x_{i+1} = x_i + \alpha_i p_i$$
$$i \leftarrow i + 1$$
  
{

# دو استراتژی - جستجو خط و منطقه اعتماد

## استراتژی جستجو خط

- انتخاب جهت  $p_i$
- جستجو در راستای جهت یافت شده از  $x_i$  فعلی به تکرار جدیدی با مقدار  $f$  کمتر
- با حل مسئله کمینه‌سازی یک‌بعدی زیر جهت یافتن  $\alpha$

$$\min_{\alpha > 0} f(x_i + \alpha p_i)$$

## استراتژی منطقه اعتماد

- تعریف شعاعی  $\Delta > 0$  اطراف  $x_i$  به عنوان منطقه اعتماد
- جمع‌آوری اطلاعات جهت ایجاد مدلی (تخمینی) از تابع  $f$
- تابع تخمین با نام  $m_i$
- دارای رفتار مشابه در نزدیکی نقطه  $x_i$

$$\min_p m_i(x_i + p)$$

$x_i + p$  در داخل منطقه اعتماد

$$\|p\|_2 \leq \Delta$$

# دو استراتژی - ادامه

استراتژی منطقه اعتماد

▪ تابع تخمین  $m_i$  معمولا برابر با

$$m_i(x_i + p) = f_i + p^T \nabla f_i + \frac{1}{2} p^T B_i p$$

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

# سیاست نخست - جستجو خط

## انتخاب $p_i$ در جهت کاهش

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

## انتخاب طول قدم $\alpha_i$ برآورده‌گر شروط وولف

- کروشه‌گذاری: یافتن بازه‌ای شامل طول قدم مناسب
- نیمه‌سازی/درون‌یابی: محاسبه قدم مناسب در بازه فعلی

# شدیدترین نزول

شبیه پائین آمدن از کوه

انتخاب سریعترین جهت سرازیری

$$p_i = -\nabla f$$

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 مقیاس گذاری

# شدیدترین نزول - مثال

حل تحلیلی

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

$$\frac{df}{dx}$$

$$= 2x - 2$$

$$x = 1$$

# شدیدترین نزول - مثال - $f(x)$ - ادامه

حل با استفاده از شدیدترین نزول

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

$$\frac{df(x)}{dx}$$

مقدار بهینه را نمی‌دانیم

▪ شروع از مقدار تصادفی  $x = 3$

قدم اول: مشتق‌گیری

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

# شدیدترین نزول - مثال - /د/امه

قدم دوم

▪ مطالعه مشتق در نقطه 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

$\alpha = 0.2$  طول قدم

$$x_0 = 3$$

$$x_{i+1} = x_i + 0.2(-f'(x_i))$$

$$x_1 = 3 - \alpha f'(3) = 3 + 0.2(-4) = 2.2$$

$$x_2 = 2.2 - \alpha f'(2.2) = 2.2 + 0.2(-2.4) = 1.72$$

$$x_3 = 1.72 - \alpha f'(1.72) = 1.72 + 0.2(-1.44) = 1.432$$

$$x_4 = 1.432 - \alpha f'(1.432) = 1.432 + 0.2(-2.864) = 1.1456$$

$$x_5 = 1.1456 - \alpha f'(1.1456) \approx 1.1456 + 0.2(-0.30) \approx 1.0956$$

ادامه محتملا به پاسخ می‌رسد.

▪ ?

چرا شدیدترین نزول به جای روش تحلیلی و حسابان

نوشتن برنامه

## شدیدترین نزول - مثال ۲

یافتن بهترین خط برازش

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

به دنبال یافتن  $\alpha$  و  $\beta$

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  $e_1 + e_2 + e_3 + e_4 = e_{\text{کل}}$

▪ غلطانداز

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

x	y
۱	۱
۲	۱
۲	۲
۳	۲

## شدیدترین نزول - مثال ۲ - ادامه

x	y
۱	۱
۲	۱
۲	۲
۳	۲

$$e_{\text{کل}} = e_1^2 + e_2^2 + e_3^2 + e_4^2 \quad \text{روش بهتر}$$
$$e_1 = \hat{y}_1 - y_1 \Rightarrow e_1 = [\alpha x_1 + \beta] - y_1$$
$$e_2 = \hat{y}_2 - y_2 \Rightarrow e_2 = [\alpha x_2 + \beta] - y_2$$
$$e_3 = \hat{y}_3 - y_3 \Rightarrow e_3 = [\alpha x_3 + \beta] - y_3$$
$$e_4 = \hat{y}_4 - y_4 \Rightarrow e_4 = [\alpha x_4 + \beta] - y_4$$

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

به دنبال یافتن خطی با کمترین خطای ممکن

x	y
۱	۱
۲	۱
۲	۲
۳	۲

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

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
۱	۱
۲	۱
۲	۲
۳	۲

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

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

مقدار اولیه

# 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؟
- راه حل؟

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

# شدیدترین نزول - مثال ۲ - ادامه

راه حل

▪ 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 همگرایی

# 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 هر زمان و کمتر از تمامی نمونه‌ها

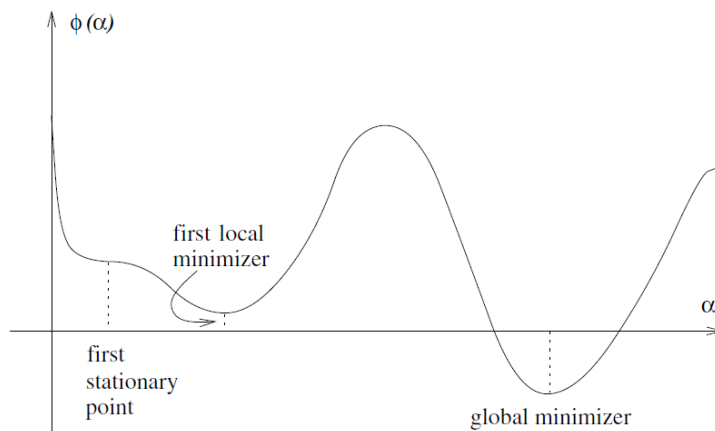
# طول قدم $\alpha$

## طول قدم

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

## تابعی بر اساس طول قدم

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



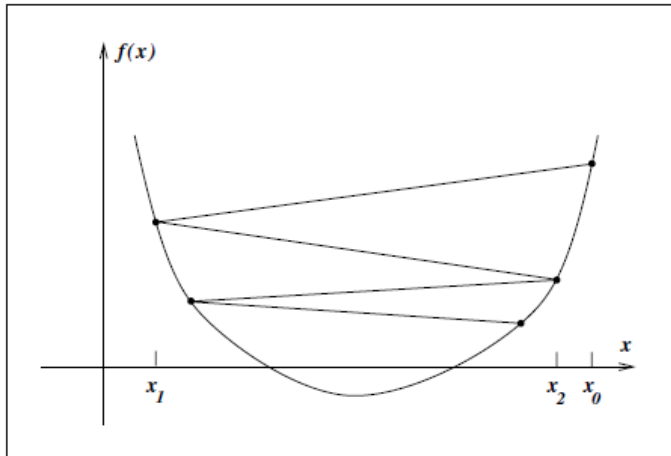
- تابعی تک متغیره
- کمینه‌ساز سراسری
- مشکل یافتن کمینه سراسری
- سیاست عملی‌تر
- جستجو خط‌جزیی

# طول قدم $\alpha$

امر غالب

- الگوریتم جستجو به دنبال یافتن مقداری برای طول قدم
- اتمام الگوریتم هنگام یافتن مقدار برآورده کننده چند شرط
- انجام جستجو خط در دو مرحله
  - کروشه گذاری
  - یافتن بازه مقادیر مطلوب
  - درون یابی/نیمه سازی
  - یافتن مقدار مناسب در بازه مذکور
- ابتدا بررسی شرط خاتمه

# طول قدم $\alpha$



ساده‌ترین شرط

▪ کاهش در مقدار تابع

$$f(x_i + \alpha_i p_i) < f(x_i) \quad \square$$

▪ 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

$$f(\mathbf{x}_i + \alpha_i \mathbf{p}_i) \leq f(\mathbf{x}_i) + c_1 \alpha_i \nabla f_i^T \mathbf{p}_i \quad \blacksquare$$

$$c_1 \in (0,1) \quad \blacksquare$$

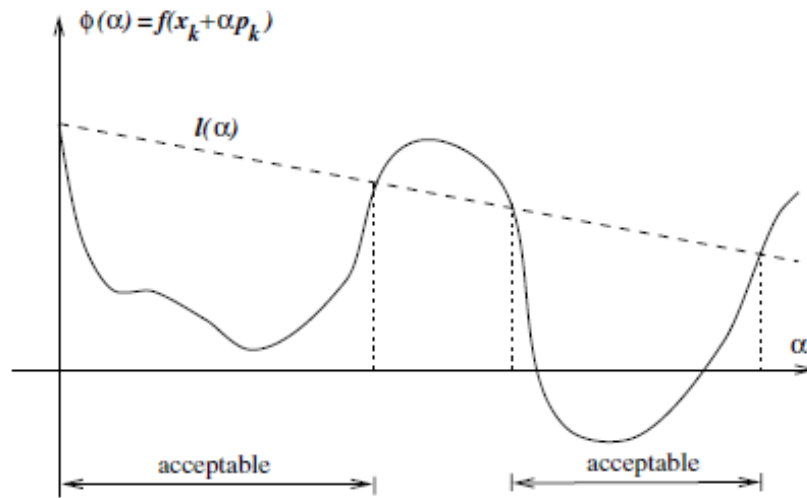
▪ تناسب کاهش  $f$  با

▪ طول قدم  $\alpha_i$

▪ مشتق جهت‌دار  $\nabla f_i^T \mathbf{p}_i$

# شروط وولف

طول قدم در هر مرحله صادق در کاهش کافی تابع هدف



کاهش کافی

$$f(x_i + \alpha_i p_i) \leq f(x_i) + c_1 \alpha_i \nabla f_i^T p_i$$

$$c_1 \in (0, 1)$$

تناسب کاهش  $f$  با

طول قدم  $\alpha_i$

مشتق جهت‌دار  $\nabla f_i^T p_i$

$$l(\alpha_i) = f(x_i) + c_1 \alpha_i \nabla f_i^T p_i$$

شهره به شرط ارمینخو

$$\phi(\alpha_i) \leq l(\alpha_i)$$

اختصاص مقداری کوچک به  $c_1$

اما «شرط کافی» مکفی نیست!

# شروط وولف

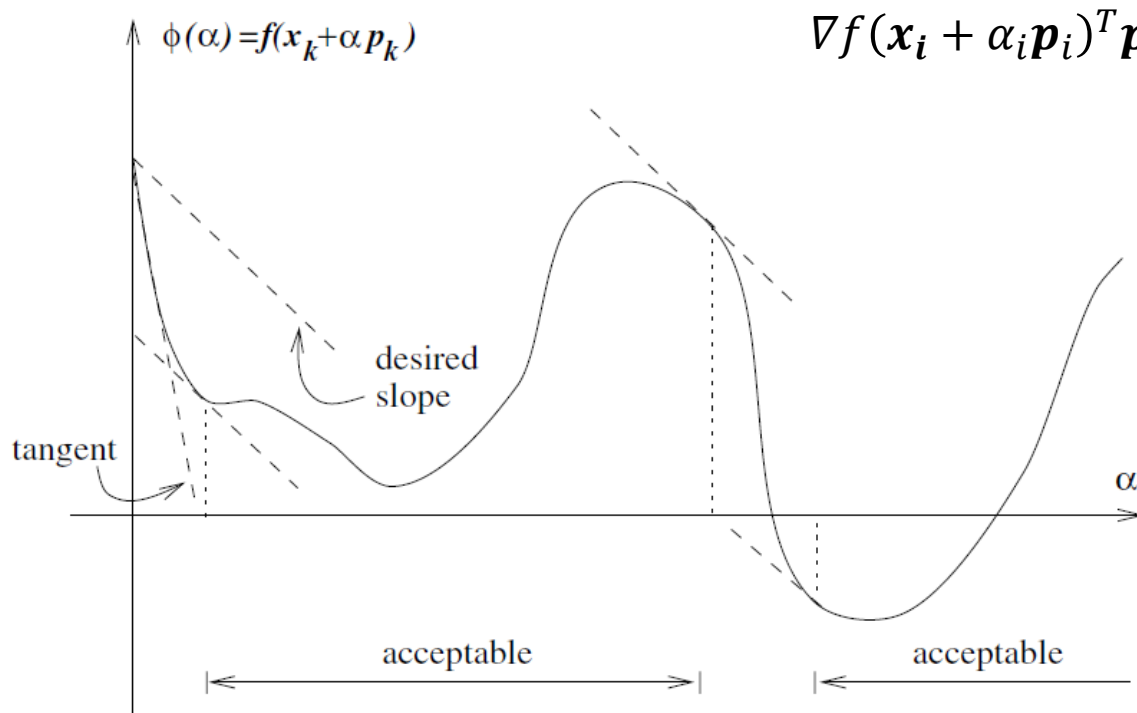
جهت حذف مقادير اقدام كوچك  
▪ شرط انحنا

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

$$c_2 \in (c_1, 1) \text{ ▪}$$

# شروط وولف

جهت حذف مقادیر اقدام کوچک  
شرط انحنا



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

$$c_2 \in (c_1, 1)$$

$$\phi'(\alpha_i) = \nabla f(\mathbf{x}_i + \alpha_i \mathbf{p}_i)^T \mathbf{p}_i$$

- شیب  $\phi$  در  $\alpha_i$  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

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

$$c_2 \in (c_1, 1) \quad \bullet$$

$$\phi'(\alpha_i) = \nabla f(\mathbf{x}_i + \alpha_i \mathbf{p}_i)^T \mathbf{p}_i \quad \bullet$$

▪ شیب  $\phi$  در  $\alpha_i$  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 معمولاً برابر ۰٫۹

▪ روش گرادیان مزدوج معمولاً برابر ۰٫۱

# شروط وولف

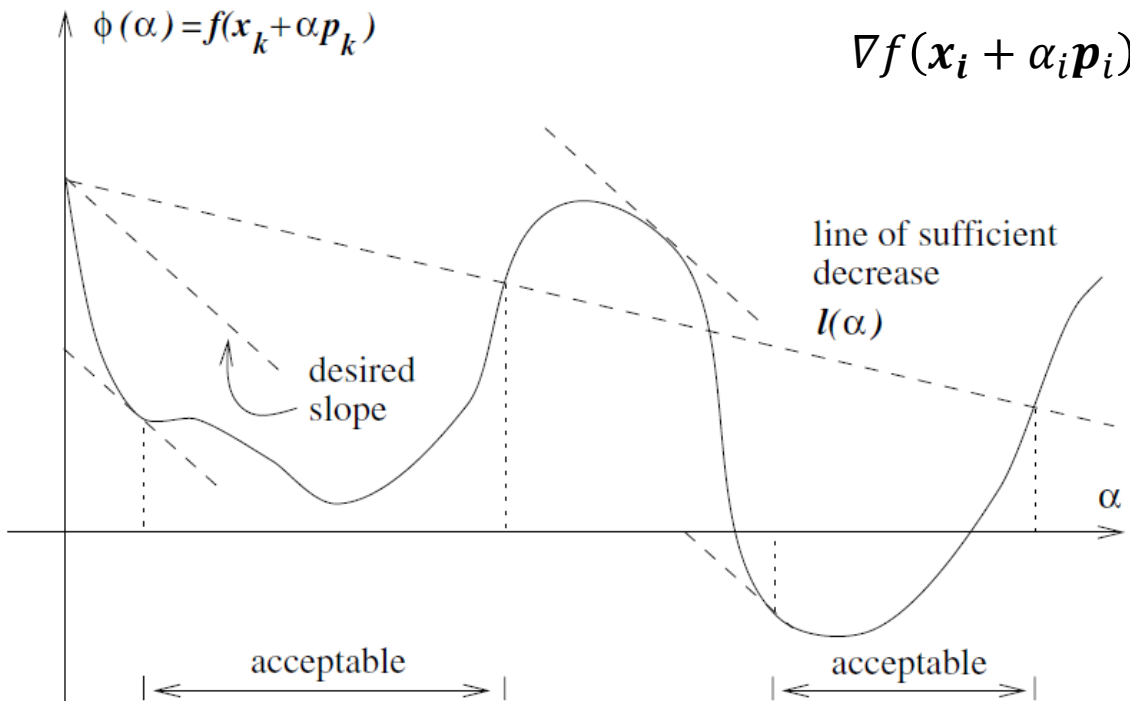
▪ شرط ارميخو

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

▪ شرط انحناء

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

$$0 < c_1 < c_2 < 1$$



# شروط قوی وولف

امکان وجود شرایط وولف بدون اینکه طول مناسبی باشد

شروط قوی وولف

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

$$|\nabla f(\mathbf{x}_i + \alpha_i \mathbf{p}_i)^T \mathbf{p}_i| \leq c_2 |\nabla f(\mathbf{x}_i)^T \mathbf{p}_i|$$

$$0 < c_1 < c_2 < 1 \quad \blacksquare$$

▪ جلوگیری از مقادیر خیلی مثبت

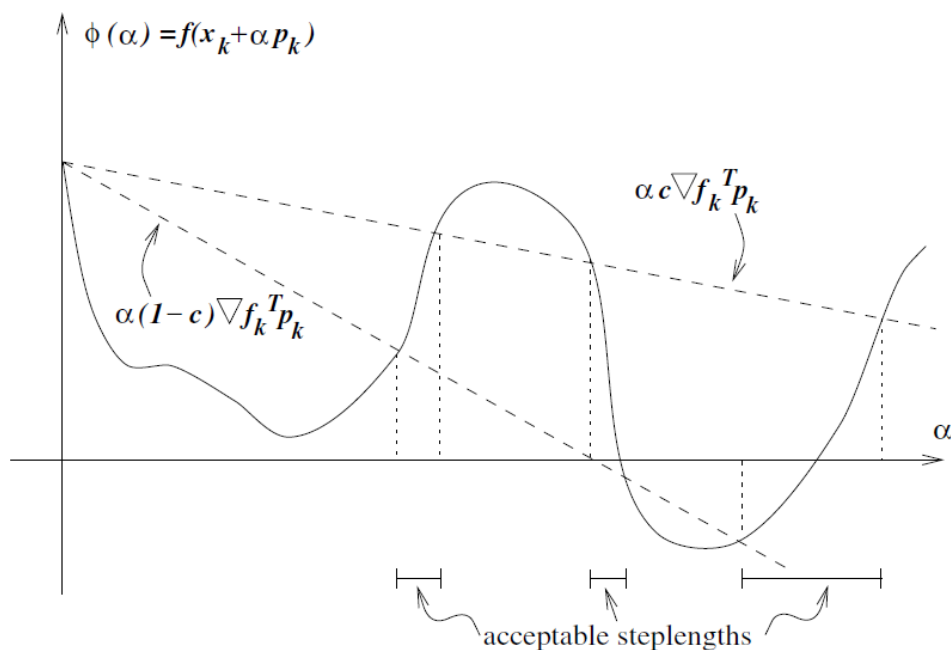
# شروط گلدشتین

اطمینان از دستیابی طول به کاهش کافی و جلوگیری از کوتاه قدم برداشتن

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

$$0 < c < 0.5$$

کنترل طول قدم از پائین



# کاهش کافی با پس روی

صرفاً استفاده از شرط کافی و نه شرط دوم

الگوریتم جستجو خط با پس روی

▪ انتخاب  $\bar{\alpha} > 0$  و  $\rho \in (0,1)$  و  $c \in (0,1)$  و  $\alpha = \bar{\alpha}$

▪ تا زمان  $f(\mathbf{x}_i + \alpha \mathbf{p}_i) \geq f(\mathbf{x}_i) + c\alpha \nabla f(\mathbf{x}_i)^T \mathbf{p}_i$

} ▪

$\alpha = \rho \alpha$  ▪

{ ▪

$\alpha_i = \alpha$  ▪

# کاهش کافی با پس روی

صرفاً استفاده از شرط کافی و نه شرط دوم

الگوریتم جستجو خط با پس روی

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

روش نیوتن و شبه نیوتن

$$\bar{\alpha} = 1$$

امکان تخصیص پویای  $\rho$  در هر مرحله

# طول قدم $\alpha$

طول قدم

اندازه هر پرش

▪ بزرگ

▪ همگرا نشدن

▪ کوچک

▪ کند

یکی از راه حل ها

▪ شروع با طول قدم بزرگ

▪ مثلاً ۱

▪ کاهش طول قدم هنگام اضافه رفتن  $\alpha = 0.8\alpha$

▪ کاهش دوباره مقدار در صورت کافی نبودن کاهش مقدار قبلی

▪ ادامه تا یافتن طول قدم مناسب

راه دیگر  $\alpha_i = \frac{c}{i}$

# طول قدم $\alpha$ - مومنتم

جهت بهبود سرعت یا طول قدم در مقاطعی که گرادیان کوچک دارند (سطوح پخت)

▪ در طول بهینه‌سازی

▪ پیگیری گرادیان‌های بدست آمده در گذشته

▪  $\beta = 0$  ؟

$$v_i = \mathbf{0}, \alpha, \beta$$

تازمان همگرا نشدن

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

$$v_{i+1} = \beta v_i - \alpha g$$

$$x_{i+1} = x_i + v_{i+1}$$

# طول قدم $\alpha$ - مومنتوم نستروف

مشکل مومنتوم. گذر از بهینه یا زیادی سریع بودن در نزدیکی بهینه  
▪ کاهش سرعت در ته دره

$$\mathbf{v}_i = \mathbf{0}, \alpha, \beta$$

تازمان همگرا نشدن

$$\mathbf{g} = \nabla f(\mathbf{x}_i + \beta \mathbf{v}_i)$$

$$\mathbf{v}_{i+1} = \beta \mathbf{v}_i - \alpha \mathbf{g}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{v}_{i+1}$$

# طول قدم $\alpha$ - آداگراد

مشکل مومنتوم و مومنتوم نستروف

▪ بروزرسانی کل مولفه‌های  $\mathbf{x}$  با اندازه یکسان

وابسته به شکل تابع

▪ در طول بهینه‌سازی

▪ پیگیری گرادیان‌های بدست آمده در گذشته

▪ تطبیق سرعت یادگیری به ازای هر مولفه  $\mathbf{x}$

▪ آداگراد

▪ مناسب برای گرادیان‌های تنک

▪ جهت گرادیان تصادفی و بی‌گیری

$$\begin{aligned} \mathbf{s} &= \mathbf{0} \\ \text{تازمان همگرا نشدن} \\ \mathbf{g} &= \nabla f(\mathbf{x}_i) \\ \text{ضرب جز به جز} \Rightarrow \mathbf{s} &= \mathbf{s} + \mathbf{g} \odot \mathbf{g} \\ \text{تقسیم جز به جز} \Rightarrow \mathbf{x}_{i+1} &= \mathbf{x}_i - \frac{\alpha}{\sqrt{\mathbf{s} + \epsilon}} \mathbf{g} \end{aligned}$$

# طول قدم $\alpha$

مشکل آداگراد

- جمع مربعات در مخرج
- مقادیری غیرکاهشی (افزایشی)
- بی‌اندازه کوچک شدن طول قدم

$$\begin{aligned} & \mathbf{s} = 0 \\ & \text{تازمان همگرا نشدن} \\ & \mathbf{g} = \nabla f(\mathbf{x}_i) \\ & \text{ضرب جز به جز} \Rightarrow \mathbf{s} = \mathbf{s} + \mathbf{g} \odot \mathbf{g} \\ & \text{تقسیم جز به جز} \Rightarrow \mathbf{x}_{i+1} = \mathbf{x}_i - \frac{\alpha}{\sqrt{\mathbf{s} + \epsilon}} \mathbf{g} \end{aligned}$$

# طول قدم $\alpha$ - ار-ام-اس پراپ

مشکل آداگراد

- جمع مربعات در مخرج

- مقادیری غیرکاهشی (افزایشی)

- بی‌اندازه کوچک شدن طول قدم

- ار-ام-اس پراپ RMSPROP

- جلوگیری از کوچک شدن بی‌اندازه مقدار طول قدم

- کاهش طول و استفاده از قبلی‌ها

- با استفاده از پنجره از چند مقدار قبلی

$$\begin{aligned} \mathbf{s} &= 0 \\ \text{تازمان همگرا نشدن} \\ \mathbf{g} &= \nabla f(\mathbf{x}_i) \\ \mathbf{s} &= \gamma \mathbf{s} + (1 - \gamma) \mathbf{g} \odot \mathbf{g} \\ \mathbf{x}_{i+1} &= \mathbf{x}_i - \alpha \frac{\mathbf{g}}{\epsilon + \sqrt{\mathbf{s}}} = \mathbf{x}_i - \alpha \frac{\mathbf{g}}{\epsilon + RMS(\mathbf{g})} \end{aligned}$$

# طول قدم $\alpha$ - آدالدا

مشکل ار-ام-اس-پراپ RMSPROP

▪ عدم توازن در یگاها

▪ آدالدا

$$\begin{aligned} \mathbf{s} &= 0 \\ \text{تازمان همگرا نشدن} \\ \mathbf{g} &= \nabla f(\mathbf{x}_i) \\ \mathbf{s} &= \gamma \mathbf{s} + (1 - \gamma) \mathbf{g} \odot \mathbf{g} \\ \mathbf{x}_{i+1} &= \mathbf{x}_i - \alpha \frac{\mathbf{g}}{\epsilon + \sqrt{\mathbf{s}}} = \mathbf{x}_i - \frac{RMS(\Delta \mathbf{x}_i)}{\epsilon + RMS(\mathbf{g})} \end{aligned}$$

# طول قدم $\alpha$ - آدام

- نگهداری مربع گرادیان کاهش یافته مانند آر-ام-اس پراپ
- نگهداری گرادیان کاهش یافته مانند مومنتوم

$$\begin{aligned} \mathbf{v} &= \mathbf{0}, \mathbf{s} = \mathbf{0} \\ &\text{تازمان همگرا نشدن} \\ \mathbf{g} &= \nabla f(\mathbf{x}_i) \\ \mathbf{v} &= \gamma_v \mathbf{v} + (1 - \gamma_v) \mathbf{g} \\ \mathbf{s} &= \gamma_s \mathbf{s} + (1 - \gamma_s) \mathbf{g} \odot \frac{\mathbf{v}}{\|\mathbf{v}\|} \\ \mathbf{v} &= \frac{\mathbf{v}}{(1 - \gamma_v^i)} \\ \mathbf{s} &= \frac{\mathbf{s}}{(1 - \gamma_s^i)} \\ \mathbf{x}_{i+1} &= \mathbf{x}_i - \alpha \frac{\mathbf{v}}{\epsilon + \sqrt{\mathbf{s}}} \end{aligned}$$

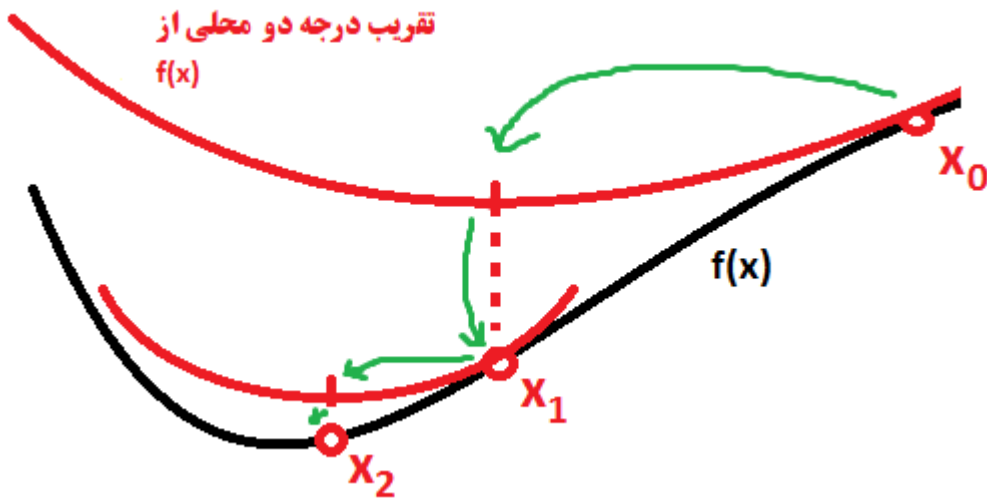
$$\begin{aligned} \alpha &= 0.001 \\ \gamma_v &= 0.9 \\ \gamma_s &= 0.999 \\ \epsilon &= 1 \times 10^{-8} \end{aligned}$$

# روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

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

روشی مبتنی بر مرتبه دوم  
▪ ماتریس هسی



# روش نیوتن

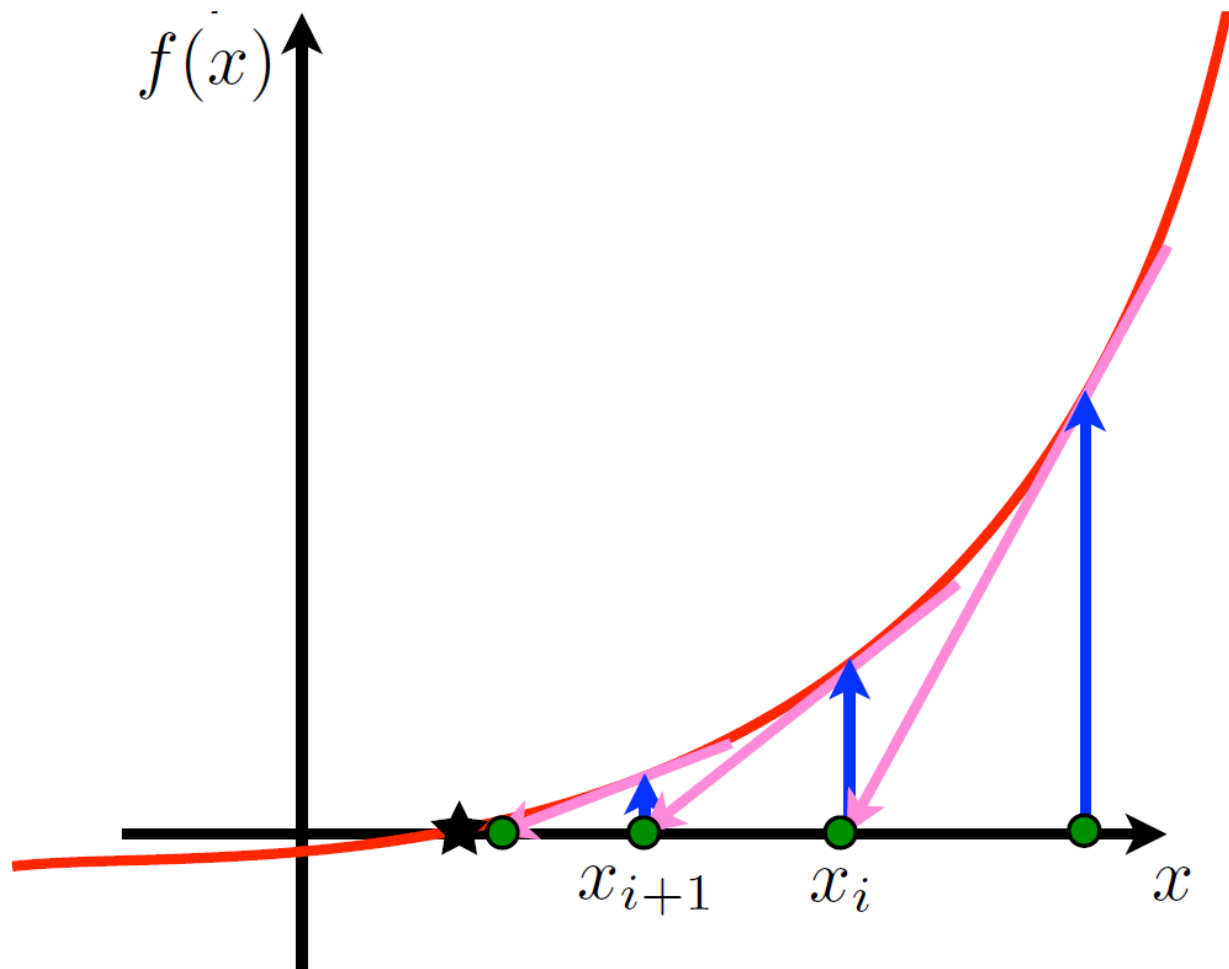
گرادیان نزولی مبتنی بر مشتق مرتبه اول

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

روشی مبتنی بر مرتبه دوم (استفاده از  $g_k = \nabla f(x_k)$  و  $H_k = \nabla^2 f(x_k)$ )  
▪ ماتریس هسی

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

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



# روش نیوتن

گرادیان نزولی مبتنی بر مشتق مرتبه اول

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

روشی مبتنی بر مرتبه دوم  
▪ ماتریس هسی

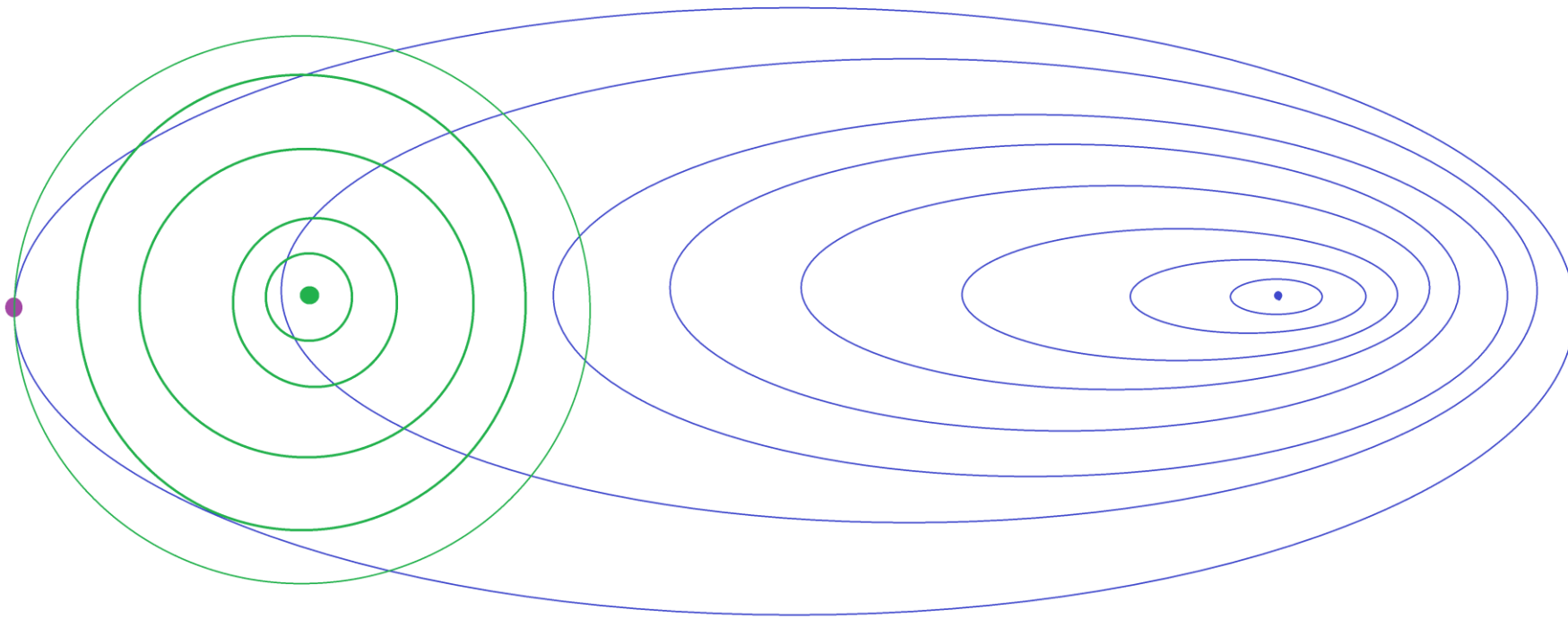
یادآوری جهت تقریب ذهن- برای یافتن صفر (یک بعدی)

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

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

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

▪ به دنبال یافتن کمینه

قضیه تیلور جهت  $\mathbf{x} \in \mathbb{R}^n$  نزدیک  $\mathbf{a}$

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

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

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

دیگر روش نوشتن

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

## روش نیوتن - ادامه

$$f(\mathbf{x}_i + \mathbf{p}) \approx f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \nabla^2 f(\mathbf{x}_i) \mathbf{p}$$
$$f(\mathbf{x}_i + \mathbf{p}) \approx \underbrace{f(\mathbf{x}_i) + \mathbf{g}_i^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{H}_i \mathbf{p}}_{m(\mathbf{p})}$$

روش نیوتن:  $\mathbf{p}$  کاهنده تقریب درجه دو  $m(\mathbf{p})$

$$\nabla m(\mathbf{p}) = \mathbf{g} + \mathbf{H}\mathbf{p} = \mathbf{0}$$

معادله نیوتن

$$\mathbf{H}\mathbf{p} = -\mathbf{g}$$
$$\mathbf{p}_i = -\mathbf{H}_i^{-1} \mathbf{g}_k$$

$\mathbf{p}_i$  جهت نیوتن

# روش نیوتن - ادامه

$$\mathbf{p}_i = -\mathbf{H}_i^{-1} \mathbf{g}_i$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{p}_i \text{ با جستجو خط}$$

# روش نیوتن - ادامه

الگوریتم روش نیوتن

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

تکرار  $\mathbf{x}_{i+1} = \mathbf{x}_i - H^{-1} g$

$$g = \nabla f(\mathbf{x}_i) \cdot$$

$$H = \nabla^2 f(\mathbf{x}_i) \cdot$$

# روش نیوتن - حل معادله غیرخطی

$$\min_{x \in \mathbb{R}^n} f(x)$$
$$\nabla f(x) = \mathbf{g}(x) = \mathbf{0}$$

نیاز به حل دستگاه n-معادله غیرخطی  $\mathbf{g}_i(x) = 0$

# روش نیوتن - حل معادله غیرخطی

$$\min_{x \in \mathbb{R}^n} f(x)$$
$$\nabla f(x) = \mathbf{g}(x) = \mathbf{0}$$

نیاز به حل دستگاه n-معادله غیرخطی  $\mathbf{g}_i(x) = 0$

$$g(x_i + \mathbf{p}) \approx g(x_i) + H(x_i)\mathbf{p} = \mathbf{0}$$

$$\mathbf{p} = H^{-1}(x_i)g(x_i)$$

# روش نیوتن - ادامه

$$\mathbf{p}_i = H^{-1}(\mathbf{x}_i)g(\mathbf{x}_i)$$
$$f'_p(\mathbf{x}_i) = g_i^T \mathbf{p}_i = -g_i^T H_i^{-1} g_i < 0, g_i \neq 0$$

برآورده‌پذیر در صورت مثبت معین بودن  $H$

در صورت مثبت معین نبودن  $H$ : افزودن قطر مثبت  $\Delta_i$  به طوری که

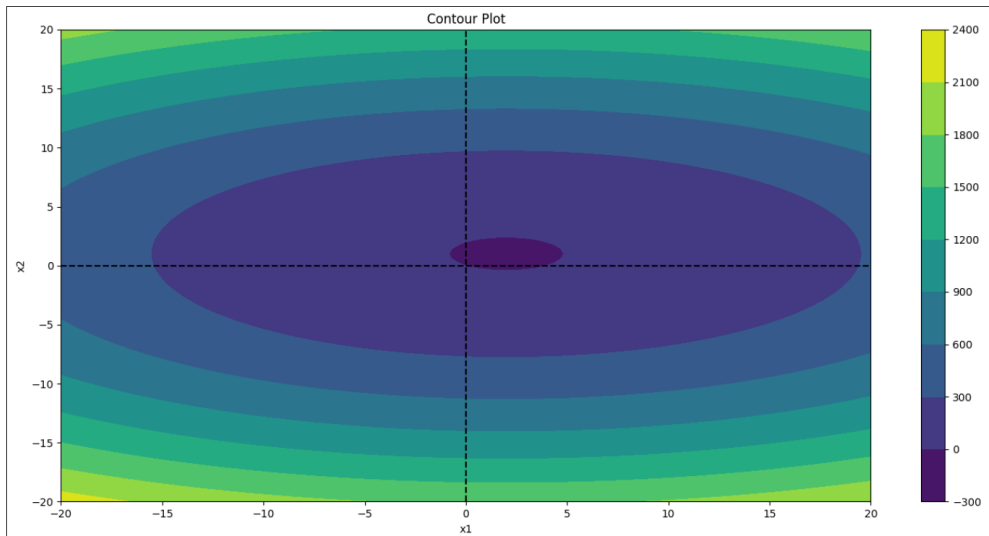
$$H_i + \Delta_i > \epsilon I \Leftrightarrow \lambda_i(H_i) > \epsilon$$

روش نیوتن تغییر یافته

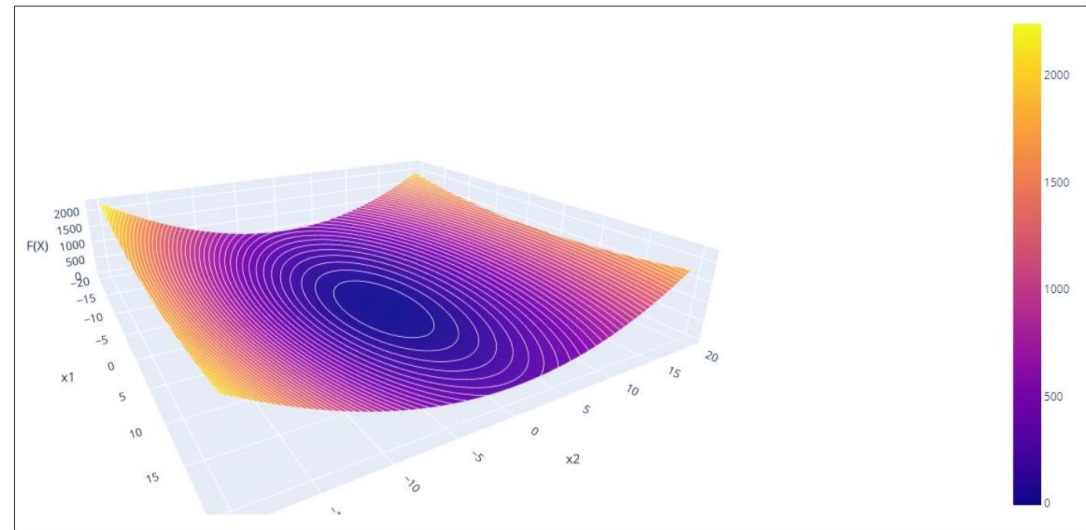
$$\mathbf{p}_i = (H_i + \Delta_i)^{-1} g_i$$

# روش تندترین نزول در مقابل روش نیوتن

$$f(x) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2$$



شکل-۲. رسم سطح ترازهای تابع  $f$ .

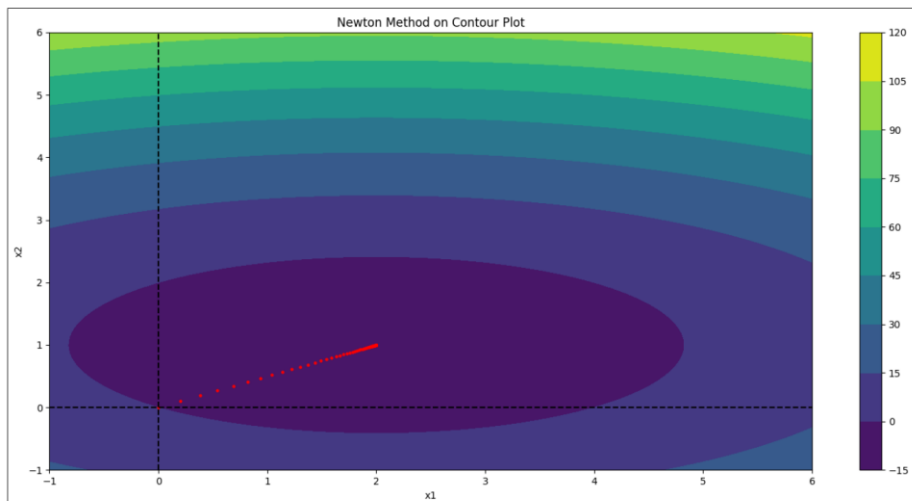


شکل-۱. نمایش ۳-بعدی تابع  $f$  به همراه سطح ترازهای آن.

# روش تندترین نزول در مقابل روش نیوتن

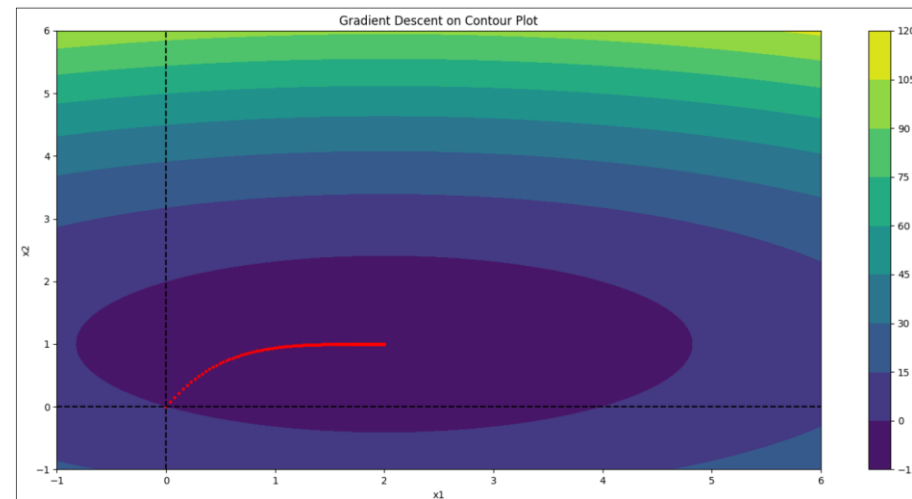
$$f(x) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2$$

با شروع از  $x_0 = [0,0]^T$  و سرعت یادگیری (طول قدم)  $\alpha_i = 0.1$



شکل-۴. نتیجه 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

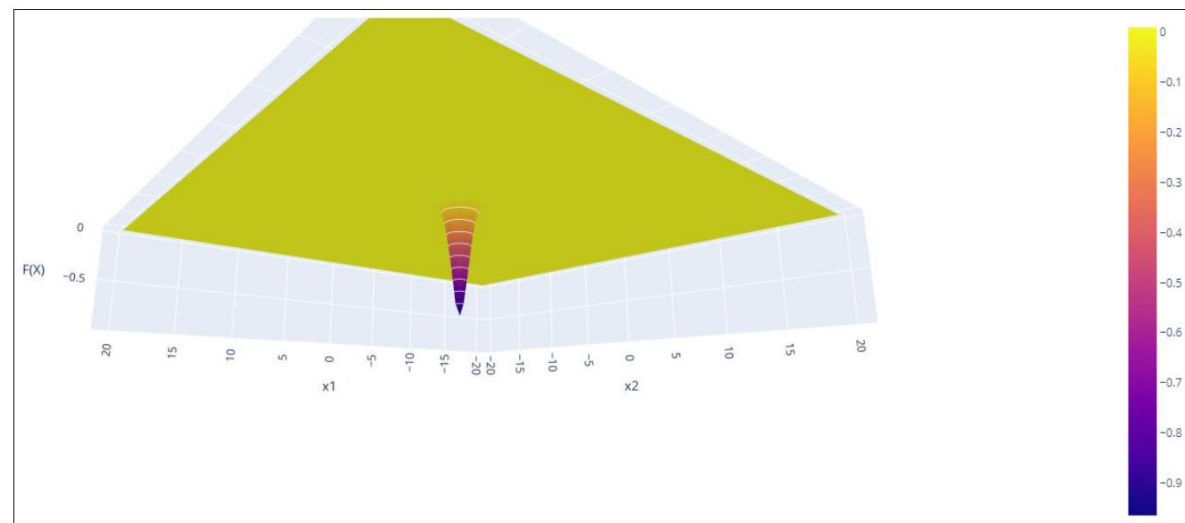


شکل-۳. نتیجه اجرای الگوریتم تندترین نزول روی تابع  $f$ . نقاط قرمز رنگ کوچکی که در تصویر قابل مشاهده است توسط الگوریتم تندترین بدست آمده است.

تندترین نزول با ۱۶۳۶ قدم

# روش تندترین نزول در مقابل روش نیوتن

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

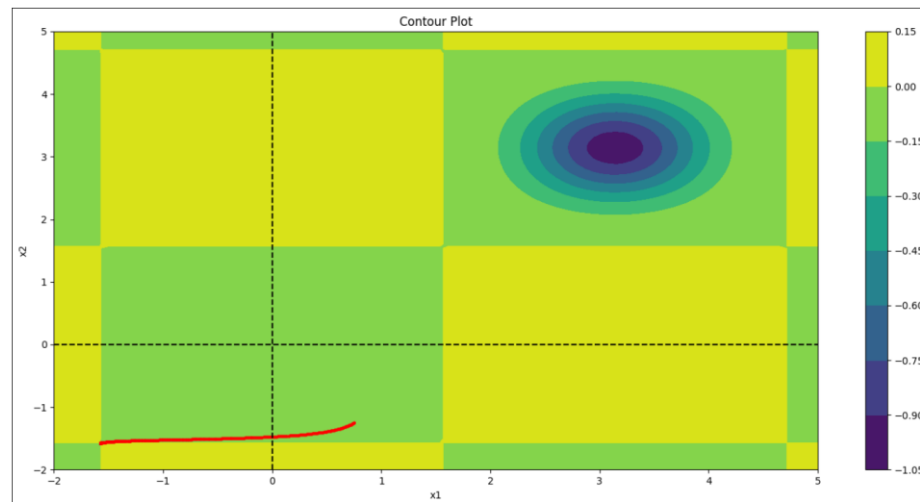


شکل-۵. نمایش ۳-بعدی تابع  $f$  به همراه سطح ترازهای آن.

# روش تندترین نزول در مقابل روش نیوتن

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

با شروع از  $\mathbf{x}_0 = [0.75, -1.25]^T$  و سرعت یادگیری (طول قدم)  $\alpha_i = 0.1$



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

همگرایی 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  $x_{i+1} = x_i - y$

- $x_{i+1} = x_i - \alpha y$

- بیشتر اوقات کار نمی کند

- 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