Лекция 5. Оптимизация в пространствах большой размерности:

нелинейный метод сопряжённых градиентов и неточный метод Ньютона

Курс «Методы оптимизации в машинном обучении»

19 октября 2015 г.

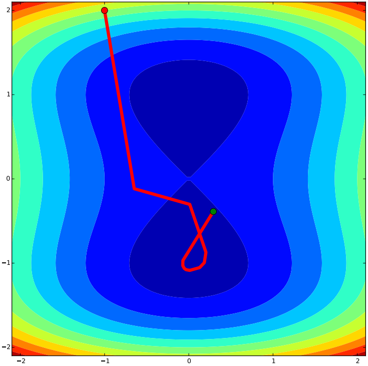
$$\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b} \to \min_{\boldsymbol{x}}$$

$\boldsymbol{g}_0 = A\boldsymbol{x}_0 - \boldsymbol{b}$
$\boldsymbol{d}_0 = -\boldsymbol{g}_0$, $\boldsymbol{u}_0 = A\boldsymbol{d}_0$;
для $k = 0, 1, \ldots, \#iter$
 $\alpha_k = \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{d}_k^T\boldsymbol{u}_k}$;
 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$;
 $\boldsymbol{g}_{k+1} = \boldsymbol{g}_k + \alpha_k\boldsymbol{u}_k$;
 если $\|\boldsymbol{g}_{k+1}\| < \varepsilon$ **то**
  $\boldsymbol{x}_* = \boldsymbol{x}_{k+1}$;
  Выход из цикла;
 $\beta_k = \frac{\boldsymbol{g}_{k+1}^T\boldsymbol{g}_{k+1}}{\boldsymbol{g}_k^T\boldsymbol{g}_k}$;
 $\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_k\boldsymbol{d}_k$;
 $\boldsymbol{u}_{k+1} = A\boldsymbol{d}_{k+1}$;

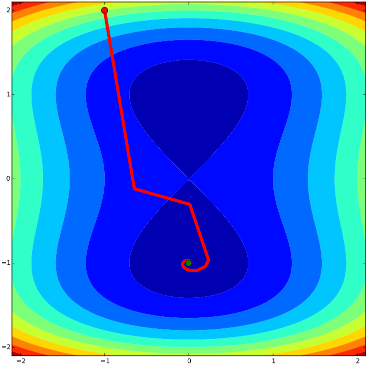
$$f(\boldsymbol{x}) \to \min_{\boldsymbol{x}}$$

$\boldsymbol{g}_0 = \nabla f(\boldsymbol{x}_0)$;
$\boldsymbol{d}_0 = -\boldsymbol{g}_0$;
для $k = 0, 1, \ldots, \#iter$
 $\alpha_k = \arg\min_{\alpha\geq 0} f(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$;
 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$;
 $\boldsymbol{g}_{k+1} = \nabla f(\boldsymbol{x}_{k+1})$;
 если $\|\boldsymbol{g}_{k+1}\| < \varepsilon$ **то**
  $\boldsymbol{x}_* = \boldsymbol{x}_{k+1}$;
  Выход из цикла;
 $\beta_k = \frac{\boldsymbol{g}_{k+1}^T\boldsymbol{g}_{k+1}}{\boldsymbol{g}_k^T\boldsymbol{g}_k}$;
 $\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_k\boldsymbol{d}_k$;

$$f(x,y) := \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2 \to \min_{x,y},$$



$c_2 = 0.9$

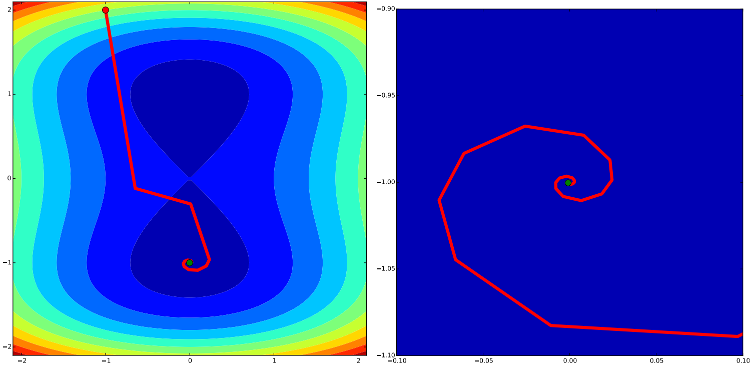


$c_2 = 0.2$

- При $c_2 > 0.5$ метод может не сходиться.
- При $c_2 < 0.5$ метод сходится.

Плохое поведение метода Флетчера–Ривса

- Часто метод Флетчера–Ривса может делать очень маленькие шаги.
- Обычно это происходит когда $\boldsymbol{d}_k$ почти ортогонально $\boldsymbol{g}_k$.
- Пример (движение по спирали):



Другие схемы метода сопряжённых градиентов

$$\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_k \boldsymbol{d}_k$$

Другие способы выбора β_k:

Полак–Рибье: $\beta_k^{\text{PR}} := \dfrac{\boldsymbol{g}_{k+1}^\top \boldsymbol{y}_k}{\|\boldsymbol{g}_k\|_2^2}$

Хестинс–Штифель: $\beta_k^{\text{HS}} := \dfrac{\boldsymbol{g}_{k+1}^\top \boldsymbol{y}_k}{\boldsymbol{d}_k^\top \boldsymbol{y}_k}$

Полак–Рибье+: $\beta_k^{\text{PR+}} := \max\{0, \beta_k^{\text{PRP}}\}$

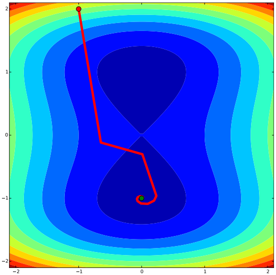
Гильберт-Ноусидаль: $\beta_k^{\text{GN}} := \max\{-\beta_k^{\text{FR}}, \min\{\beta_k^{\text{PR}}, \beta_k^{\text{FR}}\}\}$

Дай–Юань: $\beta_k^{\text{DY}} := \dfrac{\|\boldsymbol{g}_{k+1}\|_2^2}{\boldsymbol{d}_k^\top \boldsymbol{y}_k}$

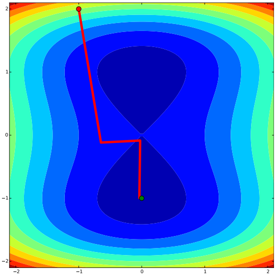
Агер–Джан: $\beta_k^{\text{HZ}} := \dfrac{\boldsymbol{g}_{k+1}^\top}{\boldsymbol{d}_k^\top \boldsymbol{y}_k}\left(\boldsymbol{y}_k - \dfrac{2\,\|\boldsymbol{y}_k\|_2^2}{\boldsymbol{d}_k^\top \boldsymbol{y}_k}\boldsymbol{d}_k\right)$

Во всех формулах $\boldsymbol{y}_k := \boldsymbol{g}_{k+1} - \boldsymbol{g}_k$.

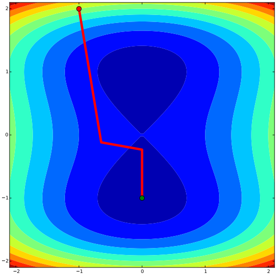
Пример: FR vs PR vs FR+restart



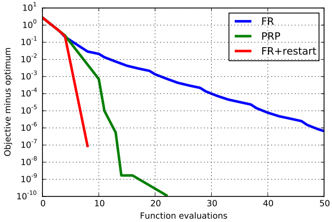
FR



PR



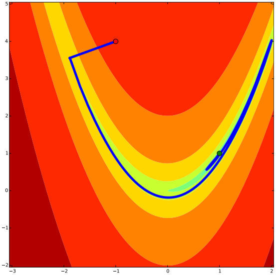
FR+restart



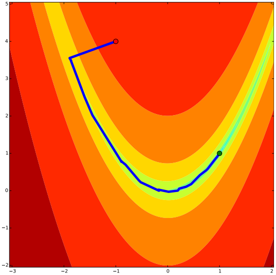
- Методы PR и FR+restart сходятся гораздо быстрее, чем просто FR.

Пример #2: функция Розенброка

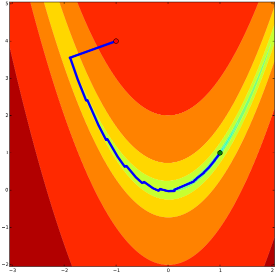
Функция Розенброка: $f(x, y) := (1 - x)^2 + 100(y - x^2)^2 \to \min\limits_{x,y}$



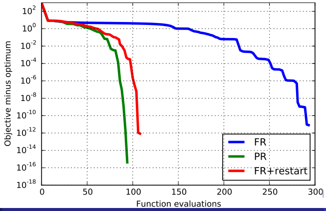
FR



PR



FR+restart



Вычислить $\nabla f(\boldsymbol{x}_0)$;
для $k = 0, 1, \ldots, \#iter$
$\varepsilon_k = \min(1/2, \sqrt{\|\nabla f_k\|})\|\nabla f_k\|$;
$\boldsymbol{z}_0 = \boldsymbol{0}$, $\boldsymbol{g}_0 = H_k \boldsymbol{z}_0 + \nabla f_k$, $\boldsymbol{d}_0 = -\boldsymbol{g}_0$;
для $j = 0, 1, \ldots,$
$\gamma_j = \dfrac{\boldsymbol{g}_j^T \boldsymbol{g}_j}{\boldsymbol{d}_j^T H_k \boldsymbol{d}_j}$;
$\boldsymbol{z}_{j+1} = \boldsymbol{z}_j + \gamma_j \boldsymbol{d}_k$, $\boldsymbol{g}_{j+1} = \boldsymbol{g}_j + \gamma_j H_k \boldsymbol{d}_j$;
если $\|\boldsymbol{z}_{j+1}\| < \varepsilon_k$ **то**
$\boldsymbol{p}_k = \boldsymbol{z}_{j+1}$;
Выход из цикла;
$\beta_j = \dfrac{\boldsymbol{g}_{j+1}^T \boldsymbol{g}_{j+1}}{\boldsymbol{g}_j^T \boldsymbol{g}_j}$;
$\boldsymbol{d}_{j+1} = -\boldsymbol{g}_{j+1} + \beta_j \boldsymbol{d}_j$;
$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k)$, $\alpha_{start} = 1$;
$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$;
если $\|\nabla f_{k+1}\| < \varepsilon$ **то**
$\boldsymbol{x}_* = \boldsymbol{x}_{k+1}$, Выход.