Методы оптимизации в машинном обучении

Метод сопряженных градиентов

Метод сопряженных градиентов (CG). Общая схема

Решаемая задача: $Ax = b$, где $A \in \mathbb{S}^n_{++}$, $b \in \mathbb{R}^n$.

Эквивалентно: $f(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \to \min_{x\in\mathbb{R}^n}$.

- Метод спуска: $x_{k+1} = x_k + \alpha_k d_k$.
- Для квадратичной функции можно аналитически найти

$$\alpha_k = \underset{\alpha\geq 0}{\operatorname{argmin}} f(x_k + \alpha_k d_k) = \frac{-\langle g_k, d_k\rangle}{\langle Ad_k, d_k\rangle}, \qquad g_k := Ax_k - b$$

Сопряженные направления: $\langle Ad_i, d_j\rangle = 0$ для $i \neq j$.

Возможные варианты:

- Ортогональные собственные векторы $q_1, \ldots, q_n$ матрицы A.
- Любые л.н.з. векторы + модифицированный процесс Грамма-Шмидта.

Это все неэффективно для больших матриц!

Метод сопряженных градиентов (CG). Общая схема – 2

Основная идея: строить d_k онлайн (в итерациях алгоритма):

- Пусть уже есть $d_0, \ldots, d_k$, т. ч. $\langle Ad_i, d_j \rangle = 0$ для $i \neq j$.
- Ищем d_{k+1} как линейную комбинацию g_{k+1} и d_k:
$$d_{k+1} = -g_{k+1} + \beta_k d_k.$$
- Коэффициент β_k найдем из условия сопряженности:
$$0 = \langle Ad_k, d_{k+1} \rangle = -\langle d_k, Ag_{k+1} \rangle + \beta_k \langle Ad_k, d_k \rangle.$$
Отсюда
$$\beta_k = \frac{\langle Ad_k, g_{k+1} \rangle}{\langle Ad_k, d_k \rangle}.$$
- Таким образом мы обеспечили лишь $\langle Ad_k, d_{k+1} \rangle = 0$. Оказывается, что если выбрать $d_0 = -g_0$ (это важно!), то автоматически будет $\langle Ad_i, d_{k+1} \rangle = 0$ и для $i < k$.

Метод сопряженных градиентов (CG). Общая схема – 3

Метод сопряженных градиентов (неэффективная версия):

1. $g_k \leftarrow Ax_k - b$
2. $\alpha_k \leftarrow \frac{-\langle g_k, d_k \rangle}{\langle Ad_k, d_k \rangle}$
3. $x_{k+1} \leftarrow x_k + \alpha_k d_k$
4. $g_{k+1} \leftarrow Ax_{k+1} - b$
5. $\beta_k \leftarrow \frac{\langle g_{k+1}, Ad_k \rangle}{\langle Ad_k, d_k \rangle}$
6. $d_{k+1} \leftarrow -g_{k+1} + \beta_k d_k$

- Недостаток: три матрично-векторных произведения.

Устраним этот недостаток:

1. Важное свойство CG: $\langle g_i, g_{k+1} \rangle = 0$ и $\langle d_i, g_{k+1} \rangle = 0$ для $i \leq k$.
2. Заметим, что $g_{k+1} = A(x_k + \alpha_k d_k) - b = g_k + \alpha_k Ad_k$.
3. Отсюда $Ad_k = \alpha_k^{-1}(g_{k+1} - g_k)$. Значит,

$$\beta_k = \frac{\langle g_{k+1} - g_k, g_{k+1} \rangle}{\langle g_{k+1} - g_k, d_k \rangle} = \frac{\|g_{k+1}\|^2}{-\langle g_k, d_k \rangle}$$
$$= \frac{\|g_{k+1}\|^2}{-\langle g_k, -g_k + \beta_{k-1} d_{k-1} \rangle} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$$

4. Аналогично $\alpha_k = \frac{\|g_k\|^2}{\langle Ad_k, d_k \rangle}$.

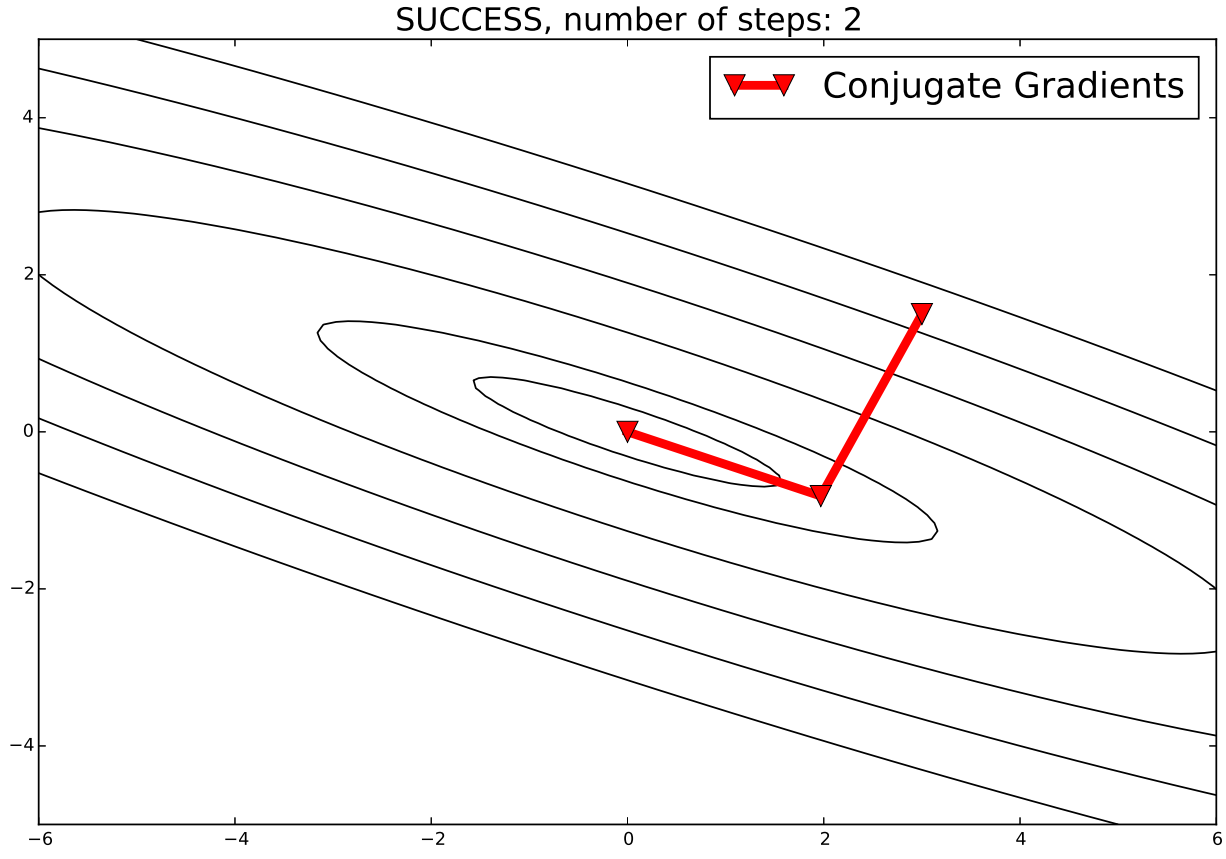
Метод сопряженных градиентов (CG). Общая схема – 4

Метод сопряженных градиентов (эффективная версия):

$g_0 \leftarrow Ax_0 - b$
$d_0 \leftarrow -g_0$
$k \leftarrow 0$
while $\|g_k\| > \varepsilon\|b\|$ **do**
  $\alpha_k \leftarrow \frac{\|g_k\|^2}{\langle Ad_k, d_k\rangle}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $g_{k+1} \leftarrow g_k + \alpha_k Ad_k$
  $\beta_k \leftarrow \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$
  $d_{k+1} \leftarrow -g_{k+1} + \beta_k d_k$
  $k \leftarrow k + 1$
end while

- Одно матрично-векторное произведение за итерацию!

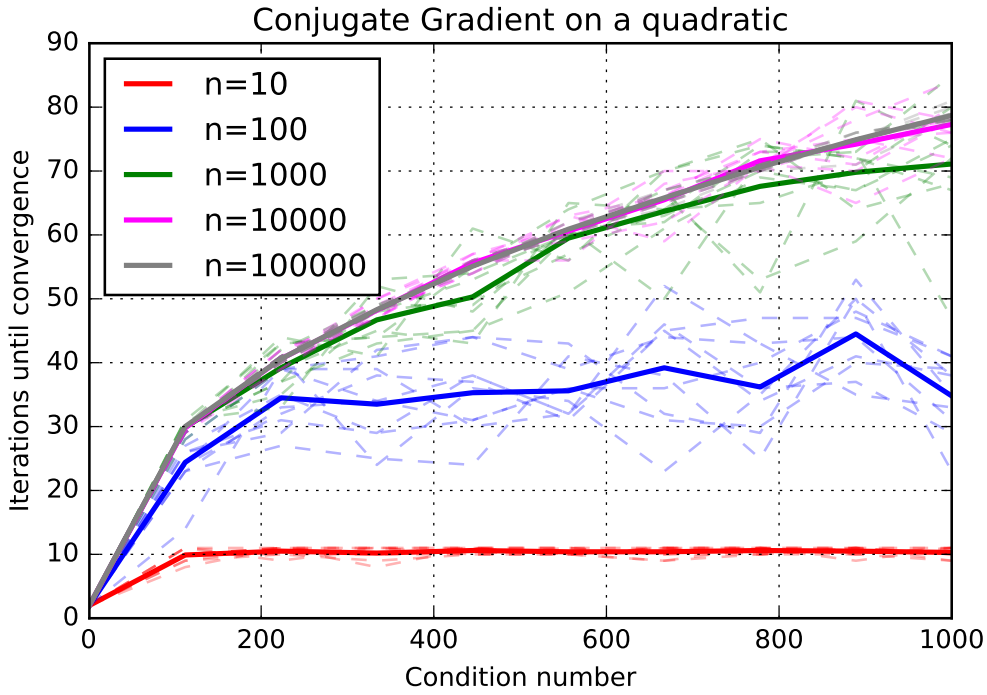
Метод сопряженных градиентов: траектория



- Всегда ≤ 2 итерации в двумерном случае.

Зависимость от обусловленности и размерности задачи

$$A = \mathrm{Diag}(a), \qquad b \equiv \mathcal{N}(0, I_n), \qquad a_i = \begin{cases} 1, & i = 1 \\ \sim \mathrm{Unif}(1, \kappa), & 2 \le i \le n-1 \\ \kappa, & i = n \end{cases}$$

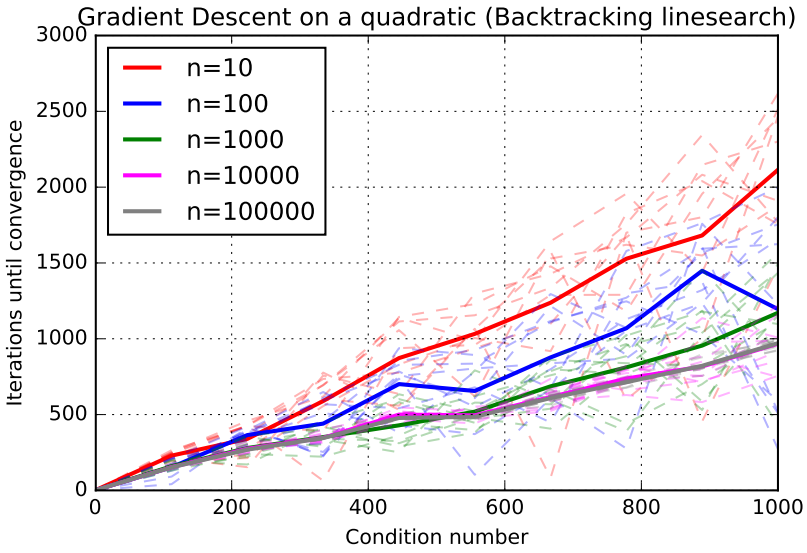


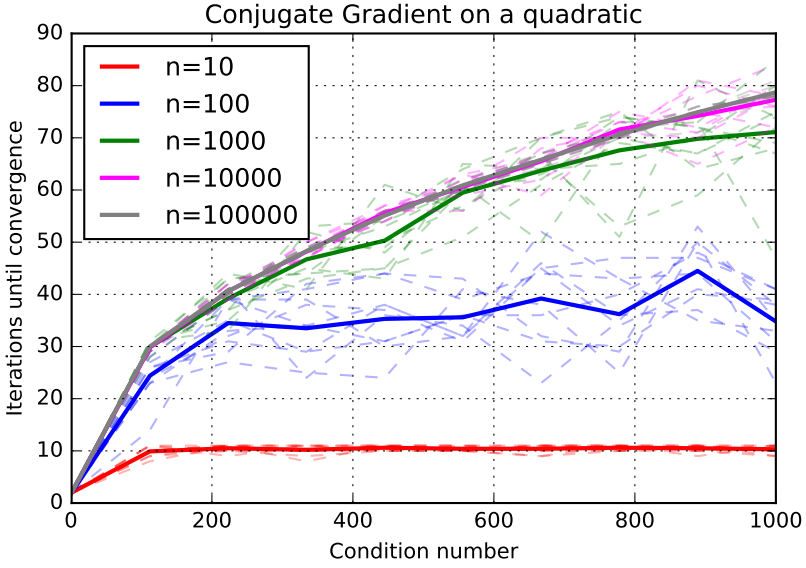
- Всегда $\le n$ итераций (м.б. чуть больше из-за погрешностей).
- В худшем случае: $O(\sqrt{\kappa})$ (κ — число обусловленности).
- С ростом размерности число итераций не увеличивается!

Сравнение CG с градиентным спуском

Квадратичная функция: $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$.

$$A = \mathrm{Diag}(a), \qquad b \equiv \mathcal{N}(0, I_n), \qquad a_i = \begin{cases} 1, & i = 1 \\ \sim \mathrm{Unif}(1, \kappa), & 2 \le i \le n-1 \\ \kappa, & i = n \end{cases}$$





- $O(\kappa)$ против $O(\sqrt{\kappa})$ (κ — число обусловленности).
- Огромная разница уже даже при небольших κ!

Предобуславливание в методе сопряженных градиентов

Оригинальная задача: $Ax = b$, где $A \in \mathbb{S}^n_{++}$, $b \in \mathbb{R}^n$.

- Выполним эквивалентное преобразование системы:

$$Ax = b \quad \Leftrightarrow \quad (S^{-T}AS^{-1})(Sx) = S^{-T}b,$$

где $S \in \mathbb{R}^{n \times n}$ — невырожденная матрица.

Новая задача: $\tilde{A}\tilde{x} = \tilde{b}$, где $\tilde{A} := S^{-T}AS^{-1}$, $\tilde{b} := S^{-T}b$.

- Решение исходной системы: $x = S^{-1}\tilde{x}$.

Предобуславливатель: $M := S^T S$.

- Идеальный предобуславливатель: $M \approx A$

$$\tilde{A} \approx S^{-T}(S^T S)S^{-1} = I_n$$

В этом случае сходимость будет примерно за одну итерацию.

Схема предобусловленного CG

Обычный CG:

$g_0 \leftarrow Ax_0 - b$
$d_0 \leftarrow -g_0$
$k \leftarrow 0$
while $\|g_k\| > \varepsilon\|b\|$ **do**
  $\alpha_k \leftarrow \frac{\|g_k\|^2}{\langle Ad_k, d_k\rangle}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $g_{k+1} \leftarrow g_k + \alpha_k Ad_k$
  $\beta_k \leftarrow \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$
  $d_{k+1} \leftarrow -g_{k+1} + \beta_k d_k$
  $k \leftarrow k + 1$
end while

Предобусловленный CG:

$g_0 \leftarrow Ax_0 - b$
$d_0 \leftarrow -M^{-1}g_0$
$k \leftarrow 0$
while $\|g_k\| > \varepsilon\|b\|$ **do**
  $\alpha_k \leftarrow \frac{\langle M^{-1}g_k, g_k\rangle}{\langle Ad_k, d_k\rangle}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $g_{k+1} \leftarrow g_k + \alpha_k Ad_k$
  $\beta_k \leftarrow \frac{\langle M^{-1}g_{k+1}, g_{k+1}\rangle}{\langle M^{-1}g_k, g_k\rangle}$
  $d_{k+1} \leftarrow -M^{-1}g_{k+1} + \beta_k d_k$
  $k \leftarrow k + 1$
end while

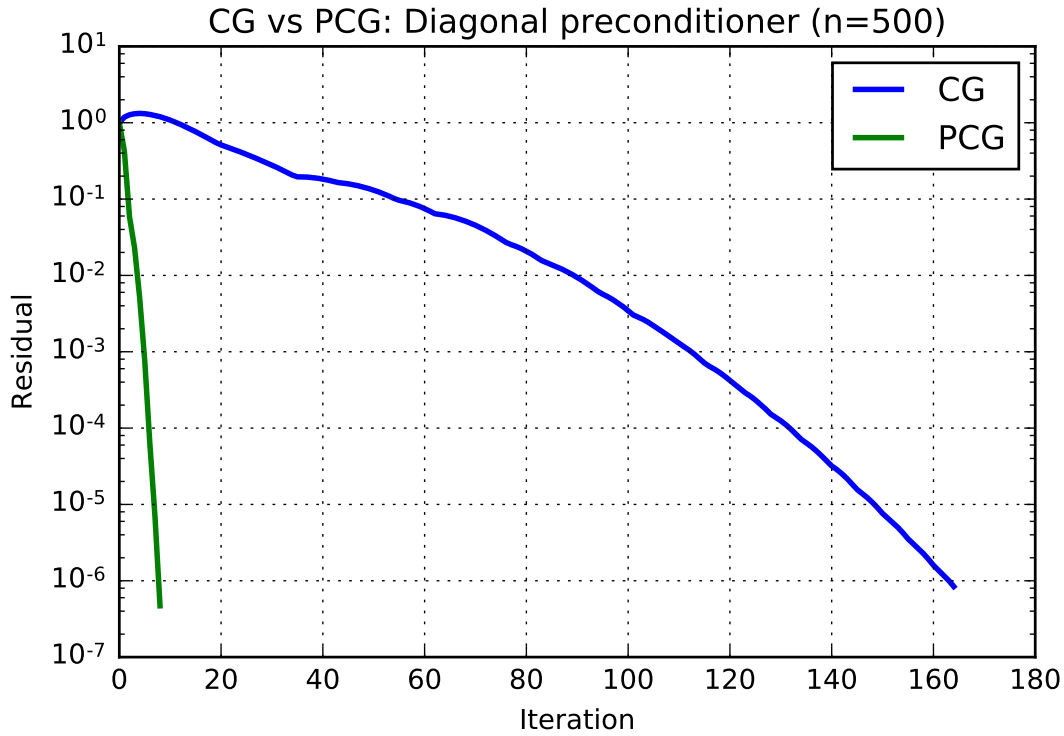
- Дополнительно нужна процедура вычисления $M^{-1}g_k$.
- Обычно применяют для хорошо структурированных M.
  - Примеры: диагональная, ленточная, разреженная и т. д.

Предобуславливание: пример

- Система $Ax = b$ размера $n = 500$, где $b := (1, \ldots, 1)$ и

$$a_{ij} := \begin{cases} 1 + i^{1.2} & \text{если } i = j \\ 1 & \text{если } |i - j| = 1 \text{ или } |i - j| = 100 \\ 0 & \text{иначе} \end{cases}$$

- Диагональный предобуславливатель: $M := \mathrm{Diag}(A)$.



- Хороший предобуславливатель существенно ускоряет метод.

Нелинейный метод CG. Версия Флетчера–Ривса

Обычный CG:

$g_0 \leftarrow Ax_0 - b$
$d_0 \leftarrow -g_0$
$k \leftarrow 0$
while $\|g_k\| > \varepsilon\|b\|$ **do**
  $\alpha_k \leftarrow \frac{\|g_k\|^2}{\langle Ad_k, d_k\rangle}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $g_{k+1} \leftarrow g_k + \alpha_k Ad_k$
  $\beta_k \leftarrow \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$
  $d_{k+1} \leftarrow -g_{k+1} + \beta_k d_k$
  $k \leftarrow k + 1$
end while

Метод Флетчера–Ривса:

$g_0 \leftarrow \nabla f(x_0)$
$d_0 \leftarrow -g_0$
$k \leftarrow 0$
while $\|g_k\| > \varepsilon\|\nabla f(x_0)\|$ **do**
  $\alpha_k \leftarrow \{\text{линейный поиск}\}$
  $x_{k+1} \leftarrow x_k + \alpha_k d_k$
  $g_{k+1} \leftarrow \nabla f(x_{k+1})$
  $\beta_k \leftarrow \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$
  $d_{k+1} \leftarrow -g_{k+1} + \beta_k d_k$
  $k \leftarrow k + 1$
end while

Другие схемы нелинейного метода CG

Существует много версий нелинейного CG. Все они отличаются лишь выбором коэффициента β_k:

Полак–Рибье: $$\beta_k^{\mathrm{PR}} := \frac{\langle g_{k+1}, y_k \rangle}{\|g_k\|^2}$$

Хестинс–Штифель: $$\beta_k^{\mathrm{HS}} := \frac{\langle g_{k+1}, y_k \rangle}{\langle d_k, y_k \rangle}$$

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

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

Дай–Юань: $$\beta_k^{\mathrm{DY}} := \frac{\|g_{k+1}\|^2}{\langle d_k, y_k \rangle}$$

Агер–Джан: $$\beta_k^{\mathrm{HZ}} := \frac{\langle g_{k+1}, y_k - \frac{2\|y_k\|^2}{\langle d_k, y_k \rangle} d_k \rangle}{\langle d_k, y_k \rangle}$$

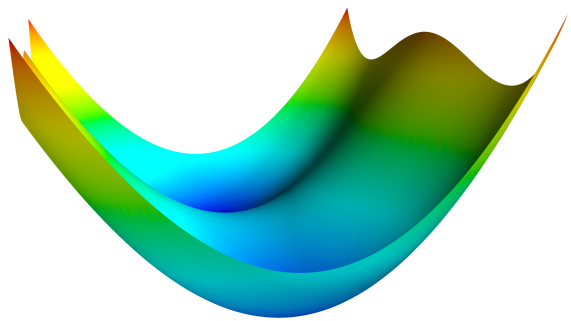
Во всех формулах $y_k := g_{k+1} - g_k$.

- Все эти схемы совпадают на строго выпуклой квадратичной функции.
- На неквадратичной функции их поведение различается.

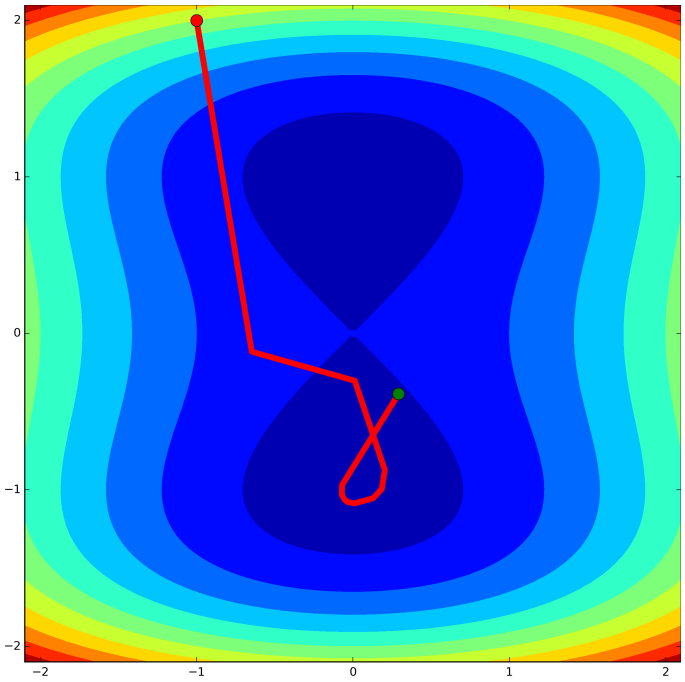
Флетчер–Ривс и направление спуска

- В общем случае метод Флетчера–Ривса не гарантирует того, что направление d_k является направлением спуска.
- Однако можно доказать, что если в линейном поиске использовать сильные условия Вульфа с константой $c_2 < 0.5$, то d_k будет направлением спуска (т. е. поиск должен быть точнее, чем обычно).
- Пример: рассмотрим двумерную функцию

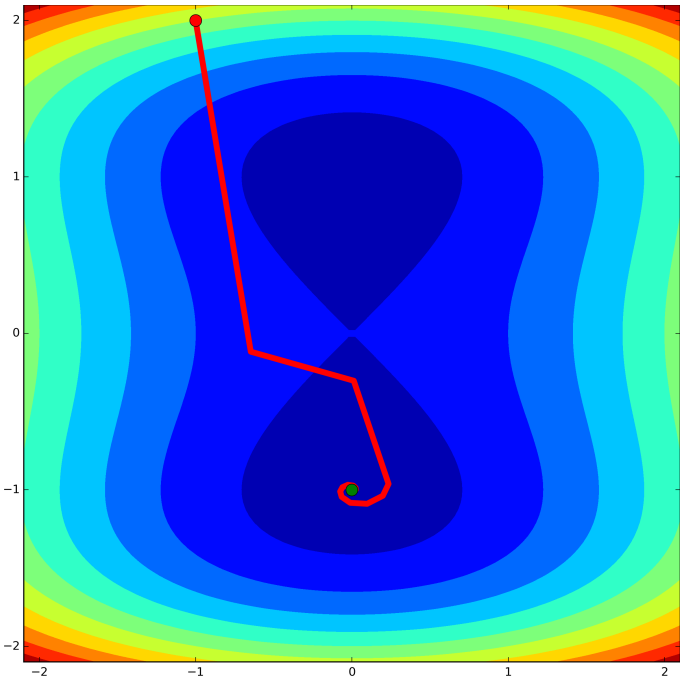
$$f(x) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2,$$



Флетчер–Ривс и направление спуска 2



$c_2 = 0.9$



$c_2 = 0.2$

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

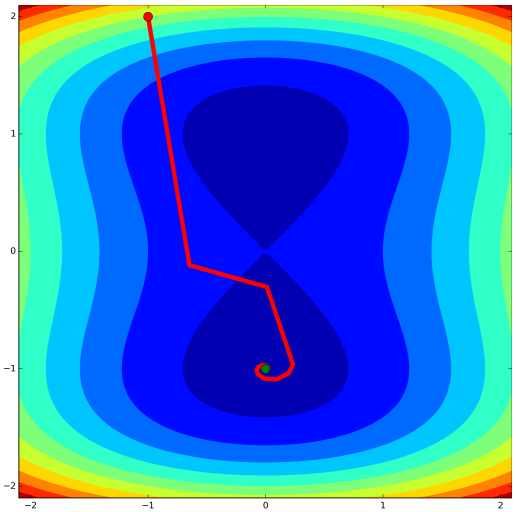
Нелинейный метод CG и рестарт

- На квадратичной функции CG сходится $\leq$ за n итераций. Для функций общего вида это необязательно так.
- Многие функции вблизи оптимума близки к квадратичной. Как гарантировать быструю сходимость в окрестности оптимума?
- Для линейного CG очень важно, что $d_0 = -g_0$.
- Этого можно добиться, если делать рестарт каждые n итераций.
- Такое условие неэффективное, т.к. на практике метод обычно запускается на менее чем n итераций (например, $n \sim 10^6$). Более эффективным условием является **условие Пауэлла**:

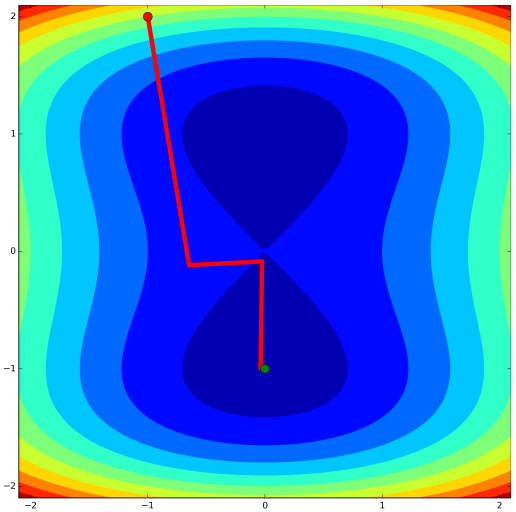
$$\text{рестарт} \Leftrightarrow |\langle g_k, g_{k+1} \rangle| \geq \nu \|g_{k+1}\|^2.$$

Обычно берут $\nu = 0.1$. Основано на том, что в CG соседние градиенты ортогональны.

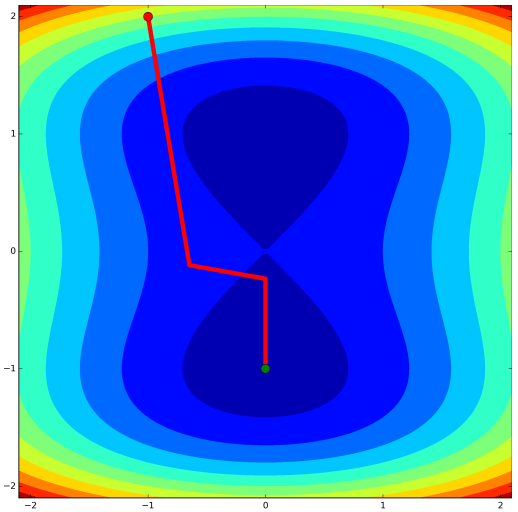
Сравнение версий FR, PR и FR+restart



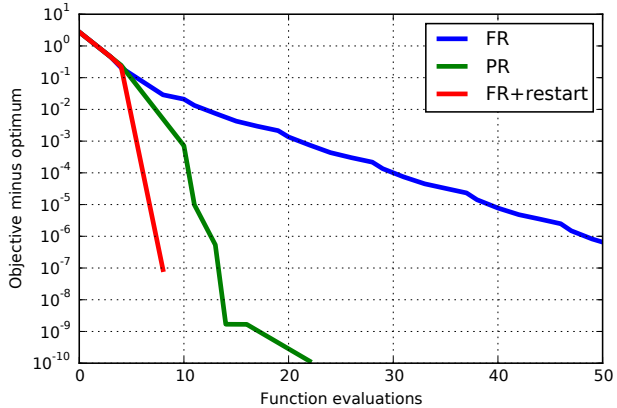
FR



PR



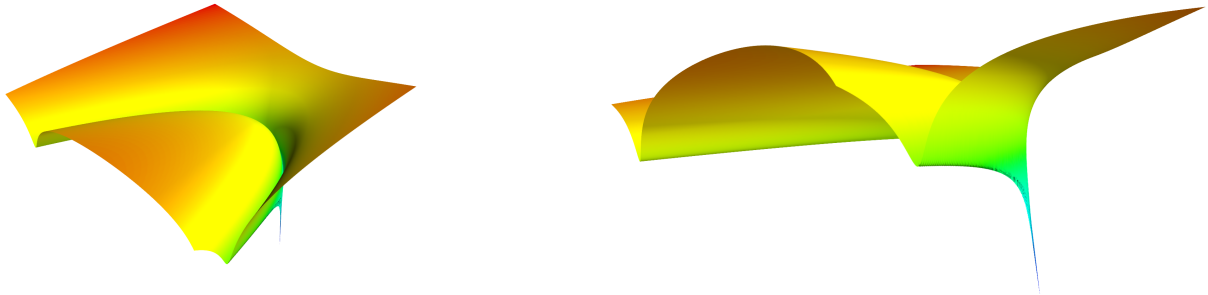
FR+restart

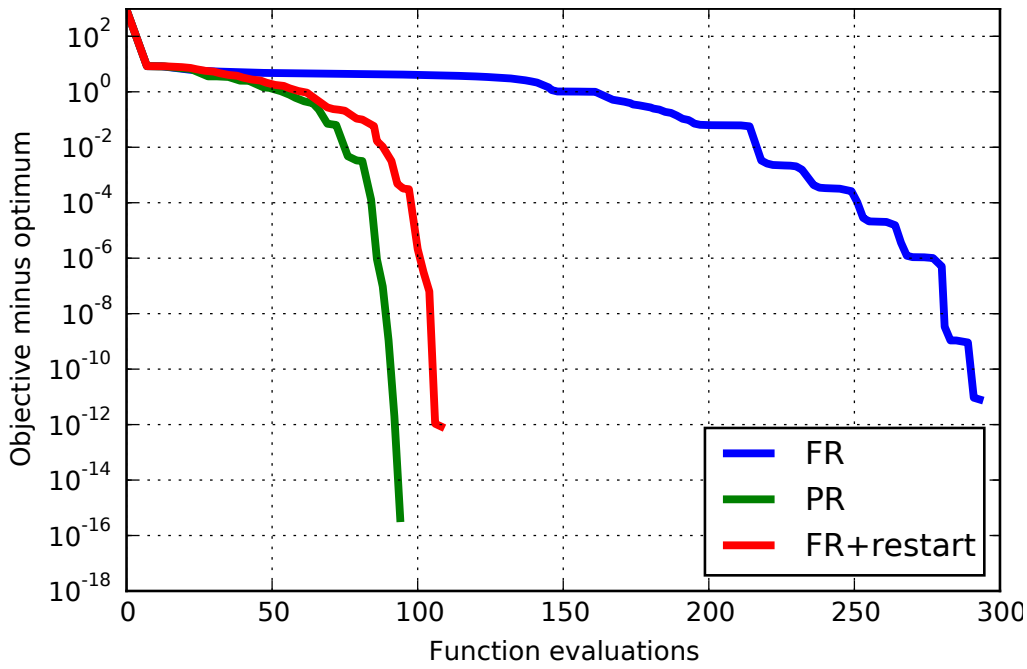


- Методы PR и FR+restart сходятся сильно быстрее, чем просто FR.
- Метод PR делает «автоматический» рестарт.

Пример 2. Функция Розенброка

Функция Розенброка: $f(x) := (1 - x_1)^2 + 100(x_2 - x_1^2)^2$.

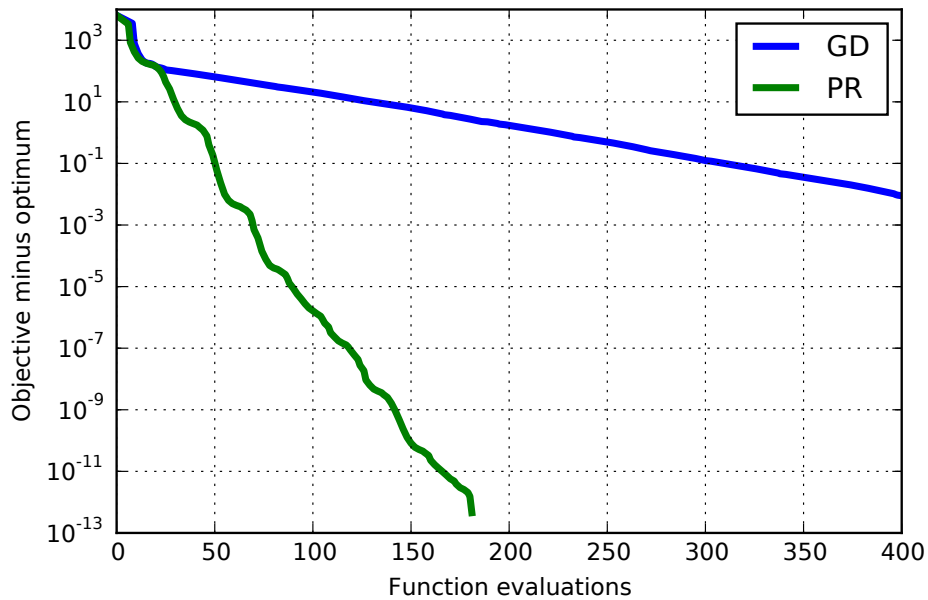




Сравнение нелинейного CG и градиентного спуска

Логистическая регрессия с l^2-регуляризатором:

$$f(x) := \sum_{i=1}^{m} \ln(1 + e^{\langle a_i, x \rangle}) + \frac{\tau}{2}\|x\|^2 \quad \to \quad \min_{x \in \mathbb{R}^n}.$$



- На практике нелинейный метод сопряженных градиентов работает существенно быстрее градиентного спуска.

Нелинейный метод CG: заключение

- Нелинейный метод CG является методом первого порядка.
- Существуют много разных вариантов, отличающихся лишь выбором коэффициента β_k.
- В линейном поиске обычно используются сильные условия Вульфа с константой $c_2 < 0.5$. Также обычно применяют рестарт.
- Метод Флетчера–Ривса без рестартов является наименее эффективным. Наиболее популярным является метод Полака–Рибье.