- Историческая справка (17 в. Ньютон)

$$t^* : \varphi(t^*) = 0$$

$\varphi : \mathbb{R} \to \mathbb{R}$

Пусть $t^0 \in \mathbb{R}$ — стартовая точка

цель: найти Δt : $t^0 + \Delta t \approx t^*$

в ряд в окрестности t^0

$$\varphi(t^0 + \Delta t) = \varphi(t^0) + \varphi'(t^0)\Delta t + o(\Delta t)$$

Объединяем:

$$\varphi(t^0 + \Delta t) \approx \varphi(t^*) = 0 \Rightarrow \varphi(t^0) + \varphi'(t^0)\Delta t \approx 0$$

$$\Delta t \approx -\frac{\varphi(t^0)}{\varphi'(t^0)}$$

$$t^1 = t^0 + \Delta t = t^0 - \frac{\varphi(t^0)}{\varphi'(t^0)}$$

$$\boxed{t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)}}$$

исторический метод Ньютона

Пример работы:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

$$t^* = 0 \qquad \varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$$

м. Ньютона: $t^{k+1} = t^k - \dfrac{\varphi(t^k)}{\varphi'(t^k)} = t^k - \dfrac{\dfrac{t^k}{(1+(t^k)^2)^{1/2}}}{\dfrac{1}{(1+(t^k)^2)^{3/2}}}$

$$= t^k - t^k\left(1 + (t^k)^2\right) = -(t^k)^3$$

1) $|t^k| > 1$ расходится

2) $|t^k| = 1$ колебания $1 \rightleftarrows -1$

3) $|t^k| < 1$ сходимся

Сходимость только в окрестности

- Обратимся к оптимизации

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

Идея $\nabla f(x^*) = 0$

Метод Ньютона для задачи оптимизации

$$\boxed{x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)}$$

Другая интуиция (смотрим разложение)

$$\underbrace{f(x)}_{\text{вместо } f(x)} \approx \underbrace{f(x^k) + \langle \nabla f(x^k); x - x^k \rangle + \frac{1}{2} \langle x - x^k; \nabla^2 f(x^k)(x - x^k) \rangle}_{\text{минимизируем квадрат аппроксимацию}}$$

$$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$$

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

пришли к тому же

для квадратичной задачи сойдется за 1 итерацию, но дорого (обращение матрицы)

Сходимость метода Ньютона

- f — μ-сильно выпукла $\quad \nabla^2 f(x) \succeq \mu I$
- $\nabla^2 f$ — M-липшицев $\quad \|\nabla^2 f(x) - \nabla^2 f(y)\|_{\underset{\uparrow}{2}} \le M\|x-y\|_2$

  спектральная

Док-во сходимости:

$$x^{k+1} - x^* = x^k - \left(\nabla^2 f(x^k)\right)^{-1}\nabla f(x^k) - x^*$$

Формула Н-Л (криволинейн интеграл):

$$\nabla f(x^k) - \underbrace{\cancel{\nabla f(x^*)}}_{0} = \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*)\,d\tau$$

Подставляем

$$x^{k+1} - x^* = x^k - x^* - \left(\nabla^2 f(x^k)\right)^{-1}\int_0^1 \underbrace{\nabla^2 f(x^* + \tau(x^k - x^*))}_{\text{Нужна } \nabla^2 f(x) - \nabla^2 f(y)}(x^k - x^*)\,d\tau$$

Умножим „1"

$$x^{k+1} - x^* = \underbrace{\left(\nabla^2 f(x^k)\right)^{-1}\nabla^2 f(x^k)}_{1}\,\underline{(x^k - x^*)}$$

$$- \underbrace{\left(\nabla^2 f(x^k)\right)^{-1}}_{\text{за скобки}}\int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))\underbrace{(x^k - x^*)}_{\text{за интеграл и скобки}}\,d\tau$$

$$x^{k+1} - x^* = \left(\nabla^2 f(x^k)\right)^{-1}\left[\nabla^2 f(x^k) - \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))\,d\tau\right](x^k - x^*)$$

Гессиан по $\int$

$$x^{k+1} - x^* = \left(\nabla^2 f(x^k)\right)^{-1}\left[\int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right)d\tau\right](x^k - x^*)$$

Оцениваем расстояние

$$\|x^{k+1}-x^*\|_2 = \left\| \left(\nabla^2 f(x^k)\right)^{-1} \left[\int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right) d\tau\right](x^k-x^*)\right\|_2$$

$$\leq \left\| \left(\nabla^2 f(x^k)\right)^{-1} \left[\int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right) d\tau\right]\right\|_2 \|x^k-x^*\|_2$$

$$\leq \left\| \left(\nabla^2 f(x^k)\right)^{-1}\right\|_2 \left\| \int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right) d\tau\right\|_2 \|x^k-x^*\|_2$$

$\left(\nabla^2 f(x^k)\right)^{-1}$ спектр ? $\quad \nabla^2 f(x^k) \succeq \mu I \Rightarrow \left(\nabla^2 f(x^k)\right)^{-1} \preceq \frac{1}{\mu} I$

$$\leq \frac{\|x^k-x^*\|_2}{\mu} \left\| \int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right) d\tau\right\|_2 \qquad (1)$$

Работаем с $\| \int \|_2$

$$\left\| \int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right) d\tau\right\|_2 \leq \int_0^1 \left\| \nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*))\right\|_2 d\tau$$

M – липшицевость гессиана

$$\leq \int_0^1 M\tau \|x^k-x^*\|_2 \, d\tau$$

$$= M\|x^k-x^*\|_2 \int_0^1 \tau \, d\tau$$

$$= \frac{M}{2} \|x^k-x^*\|_2 \qquad (2)$$

Объединяем (1) и (2)

$$\|x^{k+1}-x^*\|_2 \leq \frac{\|x^k-x^*\|_2}{\mu} \cdot \frac{M}{2} \|x^k-x^*\|_2$$

$$= \frac{M}{2\mu} \|x^k-x^*\|_2^2$$

Сходимость метода Ньютона за 1 итерацию:

$$\boxed{\|x^{k+1}-x^*\|_2 \le \frac{M}{2\mu}\|x^k-x^*\|_2^2}$$

Анализ сходимости:

хотим $\|x^1-x^*\|_2 < \|x^0-x^*\|_2$

Если $\|x^0-x^*\|_2 < \frac{2\mu}{M}$, тогда получим, что хотим

Пример:

$M = 1 \quad \mu = \frac{1}{2}$

$$\|x^{k+1}-x^*\|_2 \le \|x^k-x^*\|_2^2$$

$$\|x^0-x^*\|_2 = \frac{1}{3} \Rightarrow \|x^1-x^*\| = \left(\frac{1}{3}\right)^2 \Rightarrow \left(\left(\frac{1}{3}\right)^2\right)^2 \Rightarrow \left(\left(\left(\frac{1}{3}\right)^2\right)^2\right)^2$$

квадратичная сходимость

Итого:

+ квадратичная сходимость
− локальная сходимость
− дорогая итерация

Борьба с локальной сходимостью

- Демпфирование

$$x^{k+1} = x^k - \gamma_k \underbrace{\left(\nabla^2 f(x^k)\right)^{-1}\nabla f(x^k)}_{-p^k}$$

добавим шаг

$$\gamma_k = \arg\min_{\gamma\in\mathbb{R}} f(x^k+\gamma p^k)$$

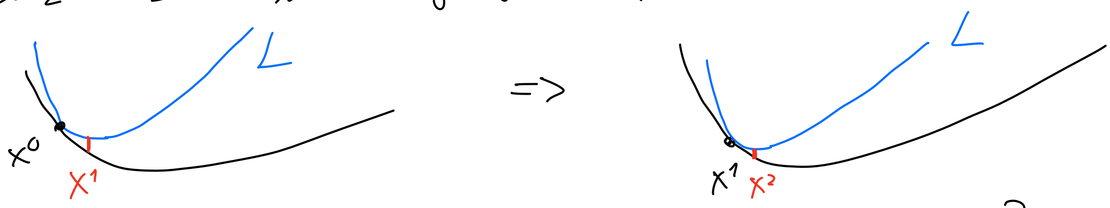
← выпукла по γ (если f выпукла)
← дихотомия / золотое сечение

• Вспомним GD, итерация GD с шагом $\frac{1}{L}$

$$x^{k+1} = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \Big(\underbrace{f(x^k) + \langle \nabla f(x^k); x - x^k\rangle + \frac{L}{2}\|x-x^k\|_2^2}\Big)$$

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$$

Шаг GD — минимизация аппроксимации сверху (L-гладкость)



Почему не $f(x) \approx \underbrace{f(x^k) + \langle \nabla f(x^k); x-x^k\rangle}_{\text{минимизировать}}$? минимум ∞

Аналогия для 2-го порядка

$$x^{k+1} = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \Big(f(x^k) + \langle \nabla f(x^k); x-x^k\rangle + \frac{1}{2}\langle x-x^k; \nabla^2 f(x^k)(x-x^k)\rangle + \frac{M}{6}\|x-x^k\|_2^3 \Big)$$

аппроксимация сверху   M-липшицев гессиан

Кубический метод Нестерова (Поляк, Нестеров)

Гессиан дорого, хотим удешевить

$$x^{k+1} = x^k - \underbrace{H_k}_{\text{матрица } d\times d} \nabla f(x^k)$$

В идеале $H_k = (\nabla^2 f(x^k))^{-1}$

Хотим считать H_k без $\nabla^2 f$, но со свойствами гессиана

⊛ разложение в ряд (в окр x^{k+1})

$$\nabla f(x^k) = \nabla f(x^{k+1}) + \nabla^2 f(x^{k+1})(x^k - x^{k+1}) + o(\|x^{k+1} - x^k\|_2)$$

$$\underbrace{x^{k+1}-x^k}_{s^k} \approx \underbrace{\left(\nabla^2 f(x^{k+1})\right)^{-1}}_{H^{k+1}} \underbrace{\left(\nabla f(x^{k+1})-\nabla f(x^k)\right)}_{y^k}$$

- Квазиньютоновское ур-е:

$$\boxed{s^k = H_{k+1} y_k}$$

- Симметричность: $H_{k+1}^T = H_{k+1}$

Сколько решений у квазиньютоновского уравнения?

Бесконечно много

Суть квазиньютоновских метод — решить уравнение „хорошо" с точки сходимости метода

Частные случаи:

- SR1 / Broyden

$$H_{k+1} = H_k + \underbrace{\underbrace{\mu_k}_{\in\mathbb{R}} \underbrace{(q^k)}_{\in\mathbb{R}^d} \underbrace{(q^k)^T}_{\in\mathbb{R}^d}}_{\text{дешевая одно-ранговая добавка}}$$

Найти q^k и μ^k

$$s^k = H_{k+1} y^k = H_k y^k + \mu_k q^k \underbrace{(q^k)^T y^k}_{\in\mathbb{R}}$$

$$= H_k y^k + \mu_k \left((q^k)^T y^k\right) q^k$$

$$s^k - H_k y^k = \underbrace{\mu_k \left((q_k)^T y^k\right)}_{\in\mathbb{R} = 1} q^k$$

$$\boxed{q^k = s^k - H_k y^k}$$

$$\boxed{\mu_k = \frac{1}{(q^k)^T y^k}}$$

- BFGS — ставим оптим. задачу

$$H_{k+1} = \arg\min_{H \in \mathbb{R}^{d\times d}} \|H - H_k\|^2_{\bigcirc}$$

любая норма (меняем норму получаем новый метод)

$$\text{s.t.} \quad Hy^k = s^k$$
$$H^T = H$$

В BFGS $\exists$ взвешенная Фроб. норма:

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F \qquad Wy^k = s^k$$

$$\boxed{H_{k+1} = \left(I - \rho_k s^k (y^k)^T\right) H_k \left(I - \rho_k y^k (s^k)^T\right) + \rho_k s^k (s^k)^T \qquad \rho_k = \frac{1}{(y^k)^T s^k}}$$

Как найти по-другому?

$$B_{k+1} = H_{k+1}^{-1} \quad (B \approx \nabla^2 f)$$

Квазиньютоновское ур-е для B:

$$B_{k+1} s^k = y^k$$

SR1 для B_{k+1}:

$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$$

$$y^k (y^k)^T \qquad \cancel{B_k s^k (y^k)^T} \qquad B_k s^k (B_k s^k)^T$$

мысль

Искать 2х-ранговое приближение:

$$B_{k+1} = B_k + \mu_{k,1}\, y^k (y^k)^T + \mu_{k,2}\, B_k y^k (B_k y^k)^T$$

используя квазиньют. ур-ние $\mu_{k,1}$, $\mu_{k,2}$:

$$B_{k+1} = B_k + \frac{y^k (y^k)^T}{(y^k)^T s^k} + \frac{B_k y^k (B_k y^k)^T}{(s^k)^T B_k s^k}$$

$$H_{k+1} = B_{k+1}^{-1} \quad \text{(ф. Шермана-Моррисона-Вудбери)}$$

можно H_{k+1} из BFGS

Апдейт BFGS:

$$H_{k+1} = \left(I - \rho_k s^k (y^k)^T\right) H_k \left(I - \rho_k y^k (s^k)^T\right) + \rho_k s^k (s^k)^T$$

Здесь есть матрица × матрицу? Если да $\Rightarrow O(d^3)$

обращение гессиана

$$H_{k+1} = H_k - \rho_k \underbrace{s^k \underbrace{(y^k)^T H_k}_{vec \times mat}}_{vec \times vec} - \rho_k \underbrace{\underbrace{H_k y^k}_{vec \times mat} (s^k)^T}_{vec \times vec} + \rho_k^2 \underbrace{s^k (y^k)^T H_k y^k (s^k)^T}_{\text{аналогично}}$$

не дороже $O(d^2)$

Итоги по квазиньютону:

+ глобальная суперлинейная сходимость
+ дешевизна вычислений (по сравнению с Ньютоном)
− дороговизна (по сравн. с GD)
− быстрая сходимость к лок. минимумам в практ. задачах