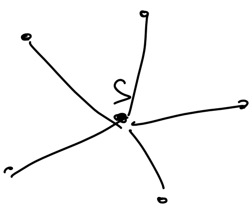
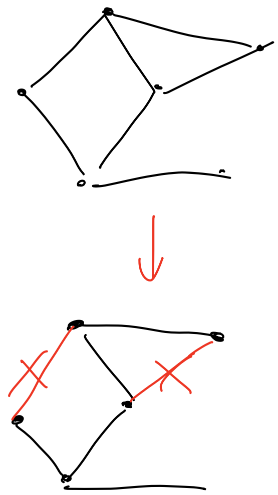
1) Децентрализованная оптимизация

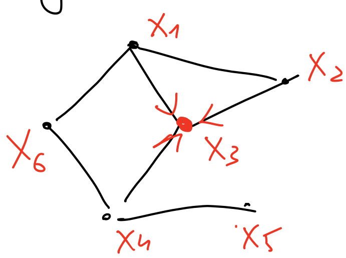
Ц



Д



gossip / диффузионные



$$\frac{1}{n}\sum x_i = \bar{x}$$

$$\tilde{x}_3 = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$$

$$\tilde{x}_3 = \alpha x_3 + \frac{1}{\gamma+\beta+c}(\gamma x_1 + \beta x_2 + c\, x_n)$$

$$\alpha+\gamma+\beta+c = 1$$

$$M_{ij} = \begin{cases} 0 & i \not\leftrightarrow j \\ \frac{1}{N(i)+1} & i \leftrightarrow j \end{cases}$$

$$x^{k+1} = M x^k \qquad x_1, x_2 \ldots x_n \in \mathbb{R} \qquad M \in \mathbb{R}^{n\times n}$$

$$x = (x_1 \ldots x_n)^T \in \mathbb{R}^n$$

$$x_i^{k+1} = \frac{1}{N(i)+1}\sum_{j\in N(i)} x_j^k$$

1) связность графа (можно ослабить, но несильно)

2) $M = M^T$ (ослабить / ор. графа)

3) стохастичность M $\quad M\underline{1} = \underline{1} \qquad \underline{1}^T \underset{M}{M^T} = \underline{1}^T$

ед. вектор

$$x_i^{k+1} = \frac{1}{N(i)+1} \sum_{j \in N(i)} x_j^k$$

$$\frac{1}{N(i)+1} \;\Big\|\; \frac{10^6}{N(i)+1}$$

$$x^{k+1} = Mx^k$$

$$Mx^k = \bar{x} \qquad Mx^{k+1} = \bar{x}$$

$$\bar{x}^{k+1} = \frac{1}{n}\left(\underline{1}^T x^{k+1}\right) = \bar{x}^k = \dots = \bar{x}^0$$

$$= \frac{1}{n}\left(\underline{1}^T M x^k\right) = \frac{1}{n}\left(\underline{1}^T x^k\right) = \bar{x}^k \; (!)$$

$$\|x^{k+1} - \bar{x}^0\| = \|Mx^k - \underbrace{M\bar{x}^0}_{\bar{x}^0}\|$$

$$\bar{x}^0 = \left(\frac{1}{n}\sum x_i^0; \dots\right) \qquad = \|M(x^k - \bar{x}^0)\| \overset{?}{<} \|x^k - \bar{x}^0\|$$

M — сжимающее отображение?

$\|M\|$?

с.з. M? $\quad \lambda_1(M) = 1$

$$\le \lambda_1(M)\,\|x^k - \bar{x}^0\|$$

$$= \|x^k - \bar{x}^0\|$$

с.в. $\underline{1}$ для с.з. 1

$$\langle x^0 - \bar{x}^0; \underline{1} \rangle = \langle (Ix^0 - \frac{1}{n}\underline{1}\,\underline{1}^T x^0; \underline{1} \rangle =$$

$x^0 \neq \bar{x}^0$

ед. матр.

$$= \left\langle \left(I - \tfrac{1}{n} 1 1^T\right) x^0 ; 1 \right\rangle$$

$$= \left\langle x^0 ; \underbrace{\left(I - \tfrac{1}{n} 1 1^T\right) 1}_{1 - \frac{1}{n} 1 (\underbrace{1^T 1}_{n})} \right\rangle$$

$$= 1 - 1 = 0$$

$$x^0 - \bar{x}^0 \in \left(\operatorname{span}(1)\right)^{\perp}$$

$$\langle x^{k+1} - \underset{\bar{x}^{k+1}}{\overset{}{\bar{x}^0}} ; 1 \rangle = \langle x^{k+1} - \bar{x}^k ; 1 \rangle$$

$$= \left\langle \left(M - \tfrac{1}{n} 1 1^T\right) x^k ; 1 \right\rangle$$

$$= \left\langle x ; \left(M - \tfrac{1}{n} 1 1^T\right) 1 \right\rangle$$

$$= \langle x ; \underbrace{M 1}_{1} - 1 \rangle = 0$$

$$x^k - \bar{x}^0 \in \left(\operatorname{span}(1)\right)^{\perp}$$

$$\|x^{k+1} - \bar{x}^0\| \le \lambda_{max}(M) \, \|x^k - \bar{x}^0\|$$

на $\left(\operatorname{span}(1)\right)^{\perp}$

если граф связен, то $\lambda_{max}(M) < 1$
на $\perp$

$x^k \to \bar{x}^0$ с экспоненц. скоростью

$$x^{k+1} = M(k)\, x^k$$

Как связать с оптимизацией?

Gossip GD

$$x_i^{k+1/2} = x_i^k - \gamma \nabla f_i(x_i^k)$$ ← локальный шаг

$$x^{k+1} = M(k)\, x^{k+1/2}$$ ← gossip усреднение

проблема → не сходится $\underbrace{\|\nabla f_i(x^*)\|^2}$

усредняем только x_i

идея: как-то аналогично делать с ∇f_i

Gradient Tracking

$\nabla f_i \to y_i$

$$x_i^{k+1/2} = x_i^k - \gamma y_i^k$$ ← локальный апдейт

$\approx \nabla f_i$, но хотим ∇f

$$x^{k+1} = M(k)\, x^{k+1/2}$$

$$y_i^{k+1/2} = y_i^k + \nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k)$$

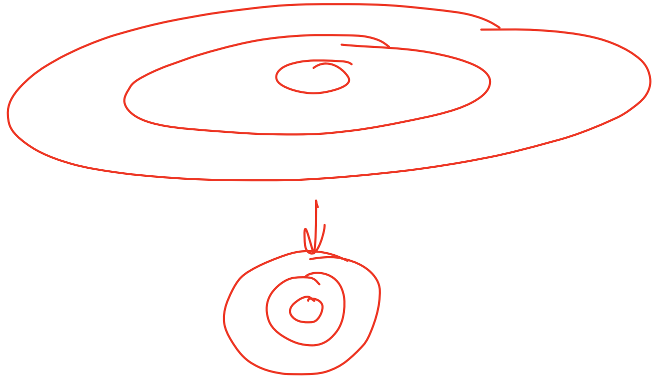
$$y^{k+1} = M(k)\, y^{k+1/2}$$

решили проблему сходимости

2) Адаптивные методы со шкалированием (адаптивность делать по компонентам)

$$x^{k+1} = x^k - \gamma (D^k)^{-1} \nabla f(x^k)$$

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



- D^k без $\nabla^2 f(x^k)$

$\frac{1}{2}(1000x^2 + y^2)$

$$\nabla^2 f(x,y) = \begin{pmatrix} 1000 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\nabla f(x,y) = \begin{pmatrix} 1000x \\ y \end{pmatrix} \qquad \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mathrm{diag}(\nabla f(x,y)) = \nabla^2 f(x,y)$$

$D_k = \mathrm{diag}(\nabla f_k)$

$\approx$

градиент гессиан интерпр.

О гессиане

- Ada Grad

$$D^k = \mathrm{diag}\left(\sum_{i=0}^{k} \nabla f(x^i) \odot \nabla f(x^i)\right)$$

$$= \begin{pmatrix} \sum_{i=0}^{k} \nabla f_{(1)}^2(x^i) & & 0 \\ & \ddots & \\ 0 & & \sum_{i=0}^{k} \nabla f_{(n)}^2(x^i) \end{pmatrix}$$

- RMS Prop

$$D^{k+1} = \beta D^k + (1-\beta)\,\mathrm{diag}\left(\nabla f(x^k) \odot \nabla f(x^k)\right)$$

$\beta \in (0;1)$ ← добавляем инерцию

- Adam: 1) Polyak HB

  2) biased correction $\beta(k)$

---

3) Безград.

$$\nabla f_{(i)}(x) \approx \frac{f(x+\tau e_i) - f(x-\tau e_i)}{2\tau} e_i$$

неточная

(для гладких функций:

$$\left\| \langle \nabla f(x); e_i \rangle e_i - \frac{f(x+\tau e_i) - f(x)}{\tau} e_i \right\|$$

$$= \frac{\|e_i\|}{\tau} \left| \underbrace{\langle \nabla f(x); \tau e_i \rangle - f(x+\tau e_i) + f(x)}_{\le \frac{L}{2}\|x+\tau e_i - x\|^2} \right|$$

$$= \frac{\|e_i\|}{\tau} \cdot \frac{L}{2} \tau^2 \|e_i\|^2 = \frac{L\tau}{2}$$