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

X — „простое" мн-во

Example: Master's Admission

- $0.0 \leq$ GPA ≤ 4.0 (from F to A)
- $0 \leq$ Salary
- $1.0 \leq$ Perfomance ≤ 6.0 (final score of secess)
- Historical data:

GPA	Salary	Perfomance
3.52	100	3.92
3.66	109	4.34
3.76	113	4.80
3.74	100	4.67
3.93	100	5.52
3.88	115	5.44
3.77	115	5.04
3.66	107	4.73
3.87	106	5.03
3.84	107	5.06

Master's Admission: Linear model

Hypothesis:

$$\text{Perfomance} \approx w_0 + w_1 \cdot \text{GPA} + w_2 \cdot \text{Salary}$$

for weights w_0, w_1, w_2 to be learned.
Approach: Find w_0, w_1, w_2 by minimizing least squares error over the historical data.
Question: what we need to do with data before solving something?

Master's Admission: Linear model

Hypothesis:

$$\text{Perfomance} \approx w_0 + w_1 \cdot \text{GPA} + w_2 \cdot \text{Salary}$$

for weights w_0, w_1, w_2 to be learned.
Approach: Find w_0, w_1, w_2 by minimizing least squares error over the historical data.
Question: what we need to do with data before solving something?

- Relevant GPA scores span a range of 0.5 (take only top students).
- Relevant Salary scores span a range of 20 (from 100 to 120 - others go to jobs, not to master).
- $\Rightarrow$ normalize first so that w_1, w_2 can be compared

General setting

n inputs $\mathrm{x}_1, \ldots, \mathrm{x}_n$, $\mathrm{x}_i \in \mathbb{R}^d$ for all i

d input variables $1, 2, \ldots, d$

- 10 (GPA, Salary) pairs, two input variables

n outputs $y_1, \ldots, y_n \in \mathbb{R}$

- 10 Perfomance scores

(x_i, y_i): an observation

- $((3.93, 100), 5.52)$, observation (of a student doing very well)

With weights $w_0, \mathrm{w} = (w_1, \ldots, w_d) \in \mathbb{R}^d$, we plan to minimize the least squares objective

$$f(w_0, \mathrm{w}) = \boxed{\sum_{i=1}^{n} (\underline{w_0} + \underline{\mathrm{w}^T \mathrm{x}_i} - y_i)^2.}$$

General setting: centering

Want to assume that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0.$$

Can be achieved by

- subtracting the mean $\bar{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i$ from every input
- subtracting the mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ from every output.

Question: after centering what we can assume?

General setting: centering

Want to assume that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0.$$

Can be achieved by

- subtracting the mean $\bar{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i$ from every input
- subtracting the mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ from every output.

Question: after centering what we can assume?

After centering: $w_0^* = 0, \mathrm{w}^*$ is unaffected

$\Rightarrow$ From now on consider function

$$f(\mathrm{w}) = \sum_{i=1}^{n} (\mathrm{w}^T \mathrm{x}_i - y_i)^2.$$

General setting: normalization

Want to assume that for all j, the n input values $x_{1j}, \dots x_{nj}$ are on the same scale:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \dots, d.$$

Can be achieved by

- multiplying x_{ij} by $s(j) = \sqrt{n/\sum_{i=1}^{n} x_{ij}^2}$ for all i, j
- in w*, this just multiplies w_j^* by $1/s(j)$

Master's Admission: Centered and normalized data

x_{i1} (GPA)	x_{i2} (Salary)	y_i (Perfomance)
-2.04	-1.28	-0.94
-0.88	0.32	-0.52
-0.05	1.03	-0.05
-0.16	-1.28	-0.18
1.42	-1.28	0.67
1.02	1.39	0.59
0.06	1.39	0.19
-0.88	-0.04	-0.12
0.89	-0.21	0.17
0.62	-0.04	0.21

Least-squares objective:

$$\min \quad f(w_1, w_2) = \sum_{i=1}^{10} (w_1 x_{i1} + w_2 x_{i2} - y_i)^2.$$

Master's Admission: Results

Optimal solution: $\min_{(w_1, w_2) \in R^2}$ $w^* = (w_1^*, w_2^*) \approx (0.43, \boxed{0.097})$

Master's Admission: Results

Optimal solution: $\mathrm{w}^* = (w_1^*, w_2^*) \approx (0.43, 0.097)$

Under hypothesis (linear model), we expect $y_i \approx y_i^* = 0.43x_{i1} + 0.097x_{i2}$

x_{i1}	x_{i2}	y_i	y_i^*
-2.04	-1.28	-0.94	-1.00
-0.88	0.32	-0.52	-0.35
-0.05	1.03	-0.05	0.08
-0.16	-1.28	-0.18	-0.19
1.42	-1.28	0.67	0.49
1.02	1.39	0.59	0.57
0.06	1.39	0.19	0.16
-0.88	-0.04	-0.12	-0.38
0.62	-0.04	0.21	0.26

Questiob: what we can say about results? Salary has only very small influence ($w_2^* = 0.097$)

Predicting Perfomance in the future

Problems:

- least squares solution is optimized for the training data, not for the future (“overfitting”)
- “unimportant” variables should have weight 0, but they typically don’t

Predicting Perfomance in the future

Problems:

- least squares solution is optimized for the training data, not for the future (“overfitting”)
- “unimportant” variables should have weight 0, but they typically don’t

Subset selection heuristics: drop variables with seemingly “small” contribution

Predicting Perfomance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights. **Easy of not?**

n k C_n^k

Predicting Perfomance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights. **Easy of not?** Non-convex or NP-hard – various k might have to be tried.

Question: if we have 100 features, how many different subsets (of features) can we have?

Predicting Perfomance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights. **Easy of not?** Non-convex or NP-hard – various k might have to be tried.

Question: if we have 100 features, how many different subsets (of features) can we have? $2^{100} \approx \underline{1.26 \cdot 10}^{30}$.

LASSO: popular approach with some favorable statistical properties

The LASSO: a constrained optimization problem

$$
\begin{array}{ll}
\text{minimize} & \sum_{i=1}^{n} \|w^\top x_i - y_i\|^2 \\
\text{subject to} & \|w\|_1 \leq R,
\end{array}
\tag{1}
$$

where $R \in \mathbb{R}_+$ is some parameter.

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top \mathrm{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top \mathrm{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \le R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$:

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top \mathrm{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: Salary is gone!

The LASSO: a constrained optimization problem

$$
\begin{array}{ll}
\text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top \mathrm{x}_i - y_i\|^2 \\
\text{subject to} & \|\mathrm{w}\|_1 \leq R,
\end{array}
\tag{1}
$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: Salary is gone!
$R = 0.3 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.3, 0)$

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top \mathrm{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \le R, \end{array} \tag{1}$$

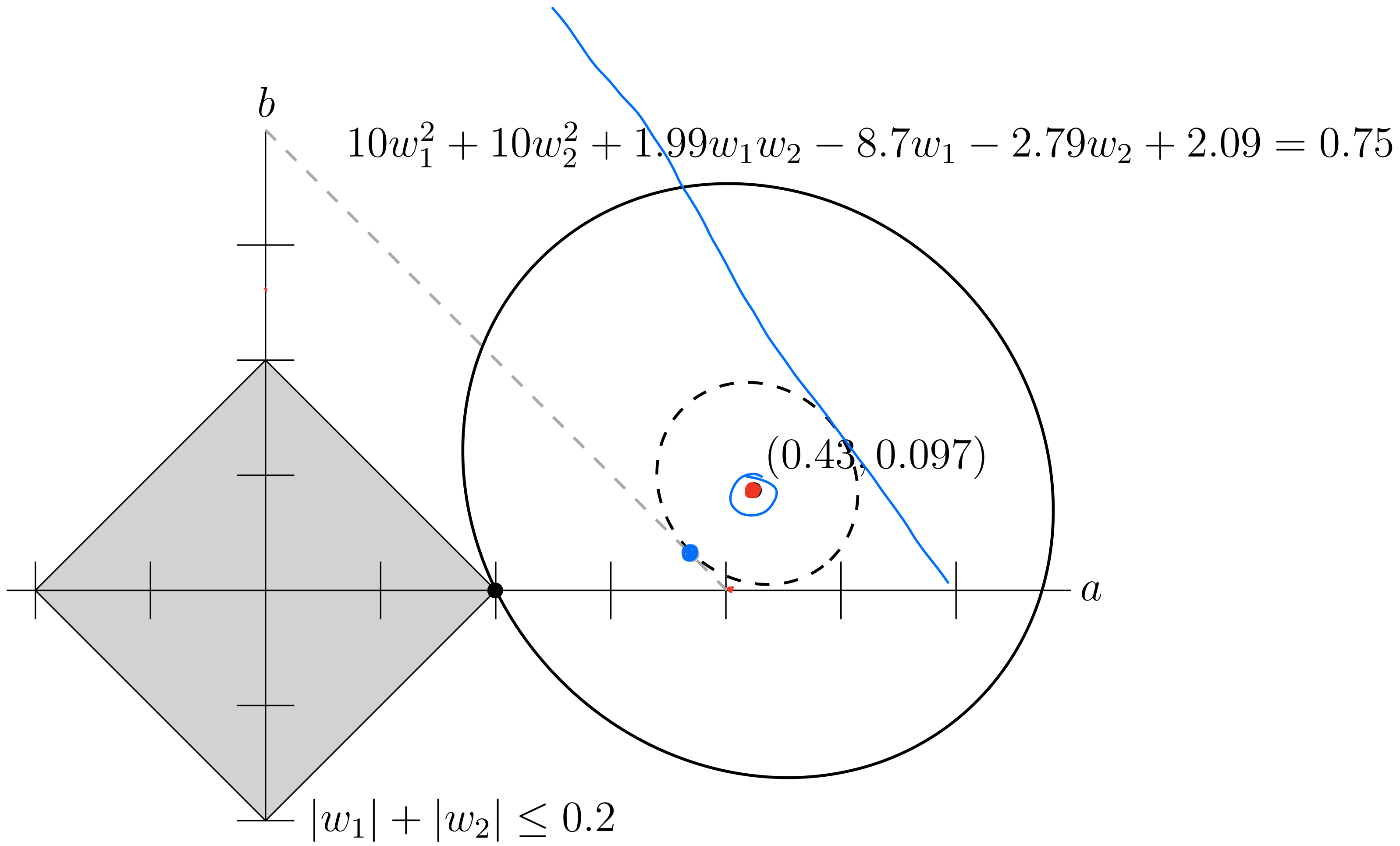
where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: Salary is gone!
$R = 0.3 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.3, 0)$
$R = 0.4 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.36, 0.036)$

The LASSO: a constrained optimization problem

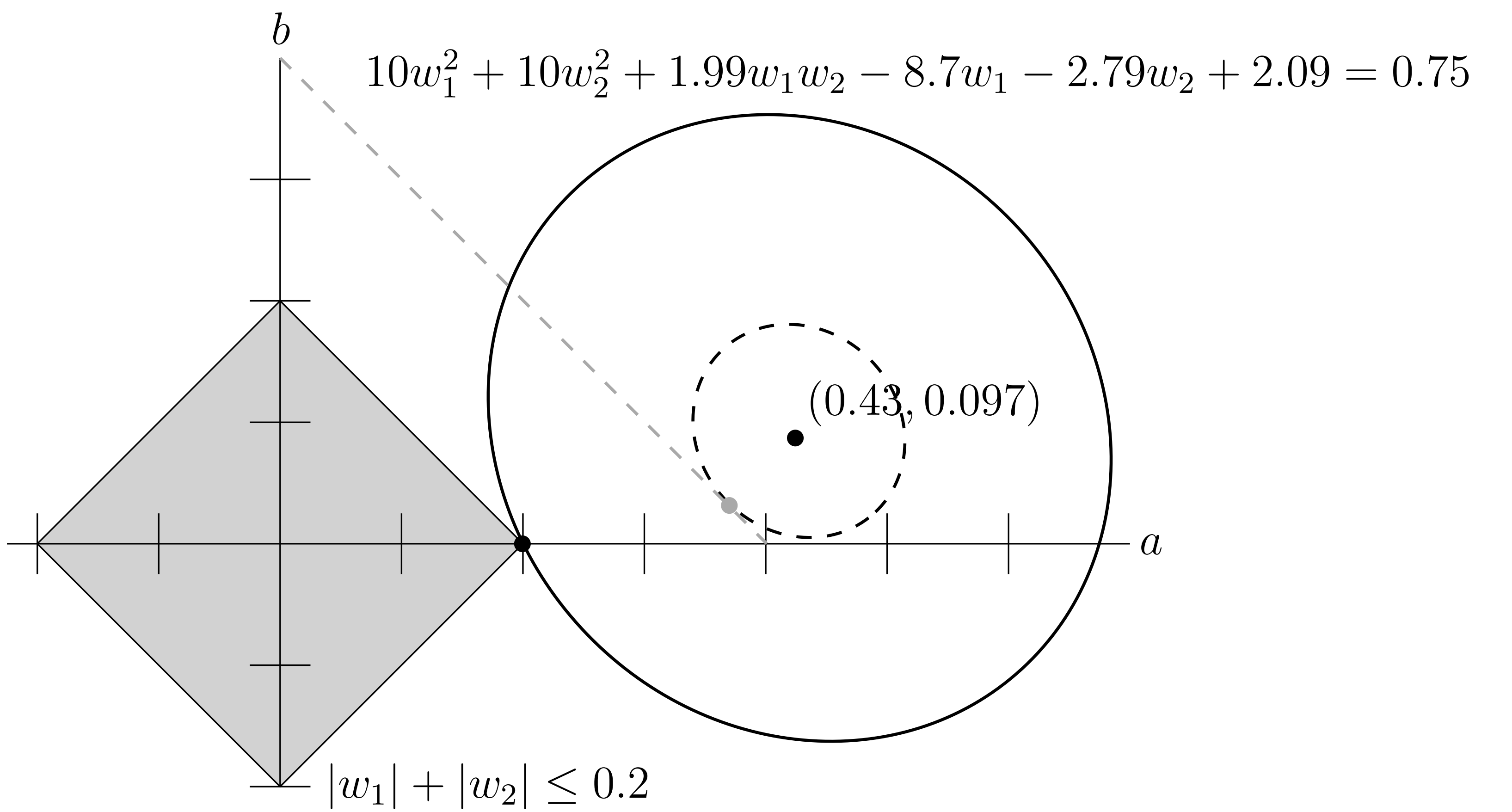
$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top \mathrm{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: Salary is gone!
$R = 0.3 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.3, 0)$
$R = 0.4 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.36, 0.036)$
$R \geq 0.6 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.43, 0.097)$

Geometry of the LASSO



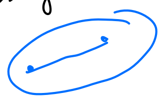
Geometry of the LASSO

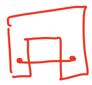


Question: Can we somehow modify gradient method to work with constraints?

Условие оптимальности

- f – непрер. дифф. на $\mathbb{R}^d$ $f: \mathbb{R}^d \to \mathbb{R}^d$
- f – выпуклая функция
- X – выпуклое

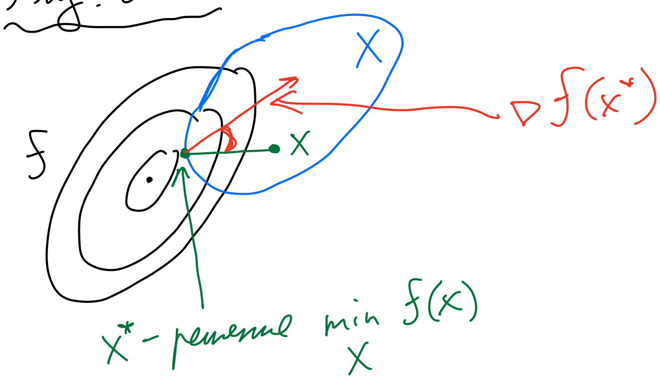




$x^* \in X$ – глобальный мин. $\min\limits_{x \in X} f(x) \iff$

$$\langle \nabla f(x^*); x - x^* \rangle \geq 0 \quad \forall x \in X$$

Физ. смысл:



угол между $\nabla f(x^*)$; $x - x^*$ – острый

Док-во:

- достаточно. $\Leftarrow$

$$\langle \nabla f(x^*); x - x^* \rangle \geq 0 \quad \forall x \in X$$

выпуклость f:

$$f(x) \geq f(x^*) + \underbrace{\langle \nabla f(x^*); x - x^* \rangle}_{\geq 0} \geq f(x^*) \quad \forall x \in X$$

x^* – глоб. минимум на X

- необходимость. $\Rightarrow$

x^* – глобальный минимум на X

От противного: $\exists x \in X$: $\langle \nabla f(x^*); x - x^* \rangle < 0$

$$x_\lambda = \lambda x + (1-\lambda) x^* \qquad \lambda \in [0; 1]$$

$$\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda) x^*)$$

$$\frac{d\varphi}{d\lambda} = \frac{d}{d\lambda}\Big(f\big(\lambda(x-x^*)+x^*\big)\Big) = \langle \nabla f(\lambda(x-x^*)+x^*); x-x^* \rangle$$

$$\left.\frac{d\varphi}{d\lambda}\right|_{\lambda=0} = \langle \nabla f(x^*); x-x^* \rangle \underbrace{< 0}_{\text{по усл.}}$$

φ убывает в окр. 0, а значит $\exists\, \lambda > 0$:

$$f(x^* + \lambda(x-x^*)) = \varphi(\lambda) < \varphi(0) = f(x^*)$$

противоречие x^* – глоб. минимум. ∎

<u>Метод град. спуска с проекцией</u>

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

$\in X$

не гарант. $x^{k+1} \in X$

$$\boxed{\begin{aligned} x^{k+1} &= \Pi_X\big(x^k - \gamma \nabla f(x^k)\big) \\ \Pi_X(y) &= \operatorname*{argmin}_{x \in X} \|x-y\|_2^2 \end{aligned}}$$

← проекция (евклидова)

<u>Св-ва проекции:</u>

- X – выпуклое, $x \in X$, $y \in \mathbb{R}^d$, тогда

$$\langle x - \Pi_X(y), y - \Pi_X(y) \rangle \le 0$$

<u>Док-во:</u> $\Pi_X(y) = \operatorname*{argmin}_{z \in \underbrace{X}_{\text{выпуклое}}} d(z) := \underbrace{\|z-y\|_2^2}_{\text{выпуклая}}$

Условие оптимальности для $d(z)$ и $\Pi_X(y)$:

$$\langle \nabla d(z^*); z - z^* \rangle \ge 0 \quad \forall z \in X$$

$$< \nabla d(\Pi_X(y)); \ x - \Pi_X(y) > \geq 0$$

$$\nabla d(z) = 2(z-y)$$

$$2 < \Pi_X(y) - y; \ x - \Pi_X(y) > \geq 0 \quad \blacksquare$$

- Нерастяжимость оператора проекции

X — выпуклое, $x_1, x_2 \in \mathbb{R}^d$, тогда

$$\|\Pi_X(x_1) - \Pi_X(x_2)\|_2 \leq \|x_1 - x_2\|_2$$

Док-во: пред. св-во $y = x_1$, $x = \Pi_X(x_2)$

$$<\underbrace{\Pi_X(x_2) - \Pi_X(x_1)}, \ x_1 - \Pi_X(x_1)> \leq 0$$

аналогично $y = x_2$, $x = \Pi_X(x_1)$

$$<\underbrace{\Pi_X(x_1) - \Pi_X(x_2)}, \ x_2 - \Pi_X(x_2)> \leq 0$$

складываем

$$<\Pi_X(x_2) - \Pi_X(x_1); \ x_1 - \Pi_X(x_1) - x_2 + \Pi_X(x_2)> \leq 0$$

$$\underbrace{<\Pi_X(x_2) - \Pi_X(x_1); \ \Pi_X(x_2) - \Pi_X(x_1)>}_{\|\Pi(x_2) - \Pi(x_1)\|_2^2} \leq <x_2 - x_1; \ \Pi(x_2) - \Pi(x_1)>$$

$$\|\Pi(x_2) - \Pi(x_1)\|_2^2 \leq \underbrace{<x_2 - x_1; \Pi(x_2) - \Pi(x_1)>}_{\text{КБШ}}$$

$$\|\Pi(x_2) - \Pi(x_1)\|_2^2 \leq \|\Pi(x_2) - \Pi(x_1)\|_2 \ \|x_2 - x_1\|_2$$

$$\|\Pi(x_2) - \Pi(x_1)\|_2 \leq \|x_1 - x_2\|_2 \quad \blacksquare$$

⊙ Стац. точка град. спуска с проекцией

$$x^* = \Pi_X(x^* - \gamma \nabla f(x^*))$$

Док-во:

$$\Pi_X(x^* - \gamma \nabla f(x^*)) = \arg\min_{x\in X}\left[\|x - x^* + \gamma \nabla f(x^*)\|_2^2\right]$$

$$= \arg\min_{x\in X}\Big[\underbrace{\underbrace{\|x - x^*\|_2^2}_{\geq 0} + 2\gamma \underbrace{\langle \nabla f(x^*); x - x^* \rangle}_{?\ \geq 0 \text{ по условию оптимальности}}}_{\geq 0\text{, но } 0 \text{ достигается на } x = x^*} + \cancel{\gamma^2 \|\nabla f(x^*)\|_2^2}\Big]$$

■

Док-во сходимости:

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

3е св-во $x^* = \ldots$

$$= \|\Pi_X(x^k - \gamma \nabla f(x^k)) - \Pi_X(x^* - \gamma \nabla f(x^*))\|_2^2$$

2е св-во $\|\Pi(x_1) - \Pi(x_2)\|_2 \leq \|x_1 - x_2\|_2$

$$\leq \|x^k - x^* - \gamma \nabla f(x^k) + \gamma \nabla f(x^*)\|_2^2$$

$$= \|x^k - x^*\|_2^2 - 2\gamma \langle \nabla f(x^k) - \nabla f(x^*); x^k - x^* \rangle + \gamma^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

μ-сильно выпуклость $\langle\ \rangle$ и L-гладкость для $\|\ \|_2^2$

$$\leq \|x^k - x^*\|_2^2 + 2\gamma \langle \nabla f(x^*); x^k - x^* \rangle - 2\gamma\left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right)$$

$$+ 2L\gamma^2 \left(f(x^k) - f(x^*) - \langle \nabla f(x^*); x^k - x^* \rangle \right)$$

$$= (1-\gamma\mu)\,\|x^k - x^*\|_2^2 + 2\gamma(\gamma L - 1)\underbrace{\left(f(x^k) - f(x^*) - \langle \nabla f(x^*); x^k - x^* \rangle \right)}_{\substack{\text{выпуклость } \geq 0 \\ \text{дивергенция Брэгмана, хар. } f}}$$

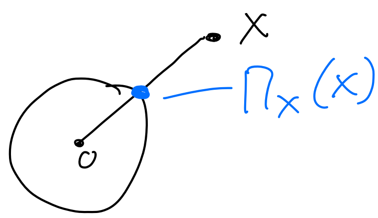
$\gamma \leq \frac{1}{L}$

$$\leq (1-\gamma\mu)\,\|x^k - x^*\|_2^2 \quad \blacksquare$$

Та же самая сходимость, что и у GD.

<u>Проекции:</u>

1) $X = \{ x \in \mathbb{R}^d \mid \|x\|_2^2 \leq 1 \}$ $\qquad \Pi_X(x) = \min\left\{1; \frac{1}{\|x\|_2}\right\} x$



2) $X = \{ x \in \mathbb{R}^d \mid a_i \leq x_i \leq b_i \}$ $\qquad [\Pi_X(x)]_i = \begin{cases} a_i & x_i \leq a_i \\ x_i & a_i < x_i < b_i \\ b_i & x_i \geq b_i \end{cases}$

3) $X = \{ x \in \mathbb{R}^d \mid Ax = b \}$ $\qquad \Pi_X(x) = x - A^T (AA^T)^{-1}(Ax - b)$

<u>Линейная задача</u> (как альт. проекции/кв. задачи):

$$\min_{s \in X} \langle s; g \rangle$$

ограничено

1) $X = \{ x \in \mathbb{R}^d \mid \|x\|_1 \le 1 \}$

$$s^* = -\operatorname{sign}(g_i) e_i \leftarrow \text{базисный вектор}$$

$$i = \arg\max_j |g_j|$$

2) $X = \{ x \in \mathbb{R}^d \mid \sum_{i=1}^{d} x_i = 1 ; \ x_i \ge 0 \}$

$$s^* = e_i \qquad i = \arg\min_j g_j$$

3) $X = \{ x \in \mathbb{R}^d \mid \|x\|_\infty \le 1 \}$

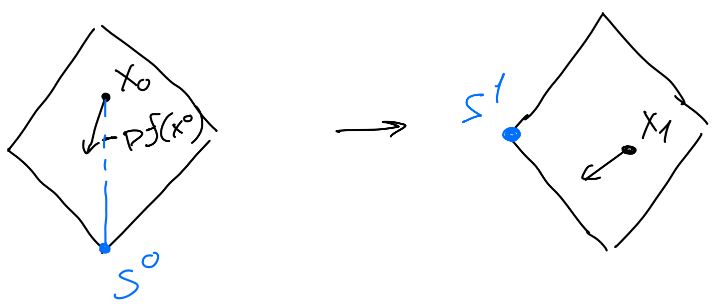
$$s^* = -\sum_{i=1}^{d} \operatorname{sign}(g_i) e_i$$

Метод Франк-Вульфа

$$s^k = \arg\min_{s \in X} \langle s ; \nabla f(x^k) \rangle$$

$$x^{k+1} = (1 - \gamma_k) x^k + \gamma_k s^k \qquad \gamma_k = \frac{2}{k+2}$$

Рисунок:



$s^0, s^1 \ldots$ – на границе (в „углах" мн-ва)

$$x^{k+1} = \underbrace{\left(1 - \frac{2}{k+2}\right)}_{1-\gamma_k} x^k + \underbrace{\frac{2}{k+2}}_{\gamma_k} s^k$$

- смотрим на границу согласно антиградиенту $\Leftarrow$
- усредняем точки на границах

$$x^{k+1} = \frac{k}{k+1} x^k + \frac{1}{k+1} s^k$$

подсчет среднего

Док-во сходимости:

$$f(x^{k+1}) = f(x^k + \gamma_k(s^k - x^k))$$

L – гладкость

$$\leq f(x^k) + \gamma_k \langle s^k - x^k; \nabla f(x^k)\rangle + \frac{\gamma_k^2 L}{2}\|s^k - x^k\|_2^2$$

X – ограничено, $D = \operatorname{diam} X$

$$\leq f(x^k) + \gamma_k \langle s^k - x^k; \nabla f(x^k)\rangle + \frac{LD^2\gamma_k^2}{2}$$

$-f(x^*)$

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \gamma_k \langle s^k - x^k; \nabla f(x^k)\rangle + \frac{LD^2\gamma_k^2}{2}$$

$$\langle s^k; \nabla f(x^k)\rangle = \min_{s\in X} \langle s; \nabla f(x^k)\rangle \leq \langle x^*; \nabla f(x^k)\rangle$$

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \gamma_k \langle x^* - x^k; \nabla f(x^k)\rangle + \frac{LD^2\gamma_k^2}{2}$$

Выпуклость

$$f(x^{k+1}) - f^* \leq (1-\gamma_k)(f(x^k) - f^*) + \frac{LD^2\gamma_k^2}{2}$$

зависит k

По индукции: если $\gamma_k = \frac{2}{k+2}$, то

$$f(x^k) - f^* \leq \frac{\max\{4C; f(x^0) - f^*\}}{k+2}$$

$$C = \frac{LD^2}{2}$$

$$f(x^{k+1}) - f^* \leq (1-\gamma_k)(f(x^k) - f^*) + C\gamma_k^2$$

$$\overset{\gamma_k}{=} \left(1 - \frac{2}{k+2}\right)(f(x^k) - f^*) + \frac{4}{(k+2)^2}C$$

ПИ

$$\leq \left(1 - \frac{2}{k+2}\right)\left(\frac{\max\{4C;\ f(x^0) - f^*\}}{k+2}\right) + \frac{\max\{4C; f(x^0) - f^*\}}{(k+2)^2}$$

$$= \frac{\max\{4C;\ f(x^0) - f^*\}}{(k+1)+2} \qquad \blacksquare$$

Итоги ФВ:

- сублинейная сход. для вып. задачи (как у GD)
- в случае сильной выпуклости все равно сублинейно
  тормоз = линейная миним.