

$$x^{k+1} = x^k - \alpha_k \cdot g^k$$

# Intuition

## Recap

Suppose, we are to solve the following problem:

$$\min_{x \in S} f(x), \quad (\text{P})$$

$$g^k \in \partial f(x^k)$$

Нерн.  
вопр.  
 $\frac{1}{\sqrt{k}}$   
 Гл.  
вопр.  
 $\frac{1}{k}$   
 АЧН.

When  $S = \mathbb{R}^n$ , we have the unconstrained problem, which sometimes could be solved with (sub)gradient descent algorithm:

$$x_{k+1} = x_k - \alpha_k g_k, \quad \begin{matrix} \min f(x) \\ x \geq 0 \end{matrix} \quad (\text{SD})$$

For this method we have the following bounds:

- 1. Метод проекции (суб)градиента
- 2. Метод условного градиента (Франка-Вульфа)

## Bounds derivation

## Introduction

В этом разделе мы будем рассматривать работу в рамках какого-то выпуклого множества  $S \in \mathbb{R}^n$ , так, чтобы  $x_k \in S$ . Запишем для начала соотношение для итераций:

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|(x_{k+1} - x_k) + (x_k - x^*)\|^2 = \\ &= \|x_k - x_{k+1}\|^2 + \|x_k - x^*\|^2 - 2\langle x_k - x_{k+1}, x_k - x^* \rangle \\ 2\langle x_k - x_{k+1}, x_k - x^* \rangle &= \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \|x_k - x_{k+1}\|^2 \end{aligned}$$

Заметим, что при работе на ограниченном множестве у нас появилась небольшая проблема:  $x_{k+1}$  может не лежать в бюджетном множестве. Сейчас мы увидим, почему это является проблемой для выписывания оценок на число итераций: если мы имеем неравенство, записанное ниже, то процесс получения оценок будет абсолютно совпадать с описанными выше процедурами (потому что в случае субградиентного метода  $x_k - x_{k+1} = \alpha_k g_k$ ).

$$\langle \alpha_k g_k, x_k - x^* \rangle \leq \langle x_k - x_{k+1}, x_k - x^* \rangle \quad (\text{Target})$$

Однако, в нашем случае мы можем лишь получить (будет показано ниже) оценки следующего вида:

$$\langle \alpha_k g_k, x_{k+1} - x^* \rangle \leq \langle x_k - x_{k+1}, x_{k+1} - x^* \rangle \quad (\text{Forward Target})$$

Это связано с тем, что  $x_{k+1}$  нам легче контролировать при построении условного метода, а значит, легче записать на него оценку. К сожалению, привычной телескопической (сворачивающейся) суммы при таком неравенстве не получится. Однако, если неравенство (Forward Target) выполняется, то из него следует следующее неравенство:

$$\begin{aligned} \langle \alpha_k g_k, x_k - x^* \rangle &\leq \langle x_k - x_{k+1}, x_k - x^* \rangle - \\ &- \frac{1}{2} \|x_k - x_{k+1}\|^2 + \frac{1}{2} \alpha_k^2 g_k^2 \end{aligned} \quad (\text{Forward Target Fix})$$

Для того, чтобы доказать его, запишем (Forward Target Fix):

$$\begin{aligned} \langle \alpha_k g_k, x_k - x^* \rangle + \langle \alpha_k g_k, x_{k+1} - x_k \rangle &\leq \\ \langle x_k - x_{k+1}, x_k - x^* \rangle + \langle x_k - x_{k+1}, x_{k+1} - x_k \rangle \end{aligned}$$

Переписывая его еще раз, получаем:

$$\begin{aligned} \langle \alpha_k g_k, x_k - x^* \rangle &\leq \langle x_k - x_{k+1}, x_k - x^* \rangle - \|x_k - x_{k+1}\|^2 - \langle \alpha_k g_k, x_{k+1} - x_k \rangle = \\ &= \langle x_k - x_{k+1}, x_k - x^* \rangle - \frac{1}{2} \|x_k - x_{k+1}\|^2 - \frac{1}{2} (\|x_k - x_{k+1}\|^2 + 2\langle \alpha_k g_k, x_{k+1} - x_k \rangle) \leq \\ &\leq \langle x_k - x_{k+1}, x_k - x^* \rangle - \frac{1}{2} \|x_k - x_{k+1}\|^2 - \frac{1}{2} (-\alpha_k^2 g_k^2) = \\ &= \langle x_k - x_{k+1}, x_k - x^* \rangle - \frac{1}{2} \|x_k - x_{k+1}\|^2 + \frac{1}{2} \alpha_k^2 g_k^2 \blacksquare \end{aligned}$$

Итак, пускай мы имеем неравенство (Forward Target) - напомню, что мы его пока не доказали. Теперь покажем, как с его помощью получить оценки на сходимость метода. Для этого запишем неравенство (Forward Target Fix):

$$\begin{aligned} 2\langle \alpha_k g_k, x_k - x^* \rangle + \|x_k - x_{k+1}\|^2 - \alpha_k^2 g_k^2 &\leq \\ &\leq 2\langle x_k - x_{k+1}, x_k - x^* \rangle \\ &= \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \|x_k - x_{k+1}\|^2 \end{aligned}$$

$$2\langle \alpha_k g_k, x_k - x^* \rangle \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 g_k^2$$

Если внимательно посмотреть на полученный результат, то это в точности совпадает с исходной точкой доказательства для субградиентного метода в безусловном сеттинге.

Можем сразу получить оценки:

$$\begin{aligned} \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle &\leq GR\sqrt{T} \\ f(\bar{x}) - f^* &\leq GR \frac{1}{\sqrt{T}} \end{aligned}$$

Таким образом, мы показали, что для метода проекции субградиента справедлива точно такая же оценка на число итераций, если выполняется неравенство (Forward Target) :) Давайте разбираться с ним

Нам следует доказать, что:

$$\langle \alpha_k g_k, x_{k+1} - x^* \rangle \leq \langle x_k - x_{k+1}, x_{k+1} - x^* \rangle$$

В более общем случае  $\forall y \in S$ :

$$\begin{aligned} \langle \alpha_k g_k, x_{k+1} - y \rangle &\leq \langle x_k - x_{k+1}, x_{k+1} - y \rangle \\ \langle \alpha_k g_k, x_{k+1} - y \rangle - \langle x_k - x_{k+1}, x_{k+1} - y \rangle &\leq 0 \end{aligned}$$

Вспомним из неравенства для проекции (равно как и условия оптимальности первого порядка), что  $\forall y \in S$  для некоторой гладкой выпуклой минимизируемой функции  $g(x)$  в точке оптимума  $x \in S$ :

$$\langle \nabla g(x), x - y \rangle \leq 0$$

$$\begin{cases} y^{k+1} = x^k - \alpha^k g^k \\ x^{k+1} = \Pi_S(y^{k+1}) \end{cases} \quad (\text{PSD})$$

$g^k \in \partial f(x^k)$

$$x^{k+1} = \Pi_S(x^k - \alpha^k g^k)$$

$$\text{Одогнання } \ell(x) = f_{x^k}(x) + \frac{1}{2} \|x - x^k\|^2 =$$

$$= f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \|x - x^k\|^2$$

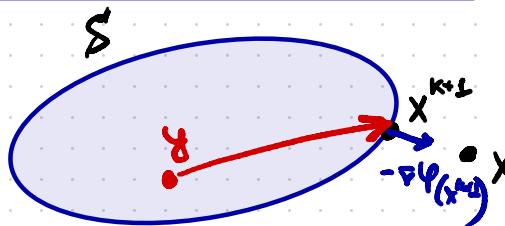
$$\underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \ell(x) = x^{k+1} \rightarrow \nabla_x \ell = 0$$

$$\alpha^k g^k + \frac{1}{2} \cdot 2 \cdot (x - x^k) = 0$$

$$x = x^k - \alpha^k g^k = x^{k+1}$$

$$\underset{x \in S}{\operatorname{argmin}} \ell(x) = x^{k+1}$$

$$\langle -\nabla \ell(x^{k+1}), x^{k+1} - y \rangle \geq 0$$



V

$$\langle \nabla \ell(x^{k+1}), x^{k+1} - y \rangle \leq 0$$

Yes

$$\langle \alpha^k g^k + x^{k+1} - x^k, x^{k+1} - y \rangle \leq 0$$

$$\langle \alpha^k g^k, x^{k+1} - y \rangle + \langle x^{k+1} - x^k, x^{k+1} - y \rangle \leq 0$$

$$\langle \alpha^k g^k, x^{k+1} - y \rangle - \langle x^k - x^{k+1}, x^{k+1} - y \rangle \leq 0$$

В противном бы случае, можно было бы сделать градиентный шаг в направлении  $y - x$  и уменьшить значение функции.

Рассмотрим теперь следующую функцию  $g(x)$ :

$$g(x) = \langle \alpha_k g_k, x \rangle + \frac{1}{2} \|x - x_k\|^2, \quad \nabla g(x) = \alpha_k g_k + x - x_k$$

И давайте теперь строить условный алгоритм как минимизацию этой функции:

$$x_{k+1} = \arg \min_{x \in S} \left( \langle \alpha_k g_k, x \rangle + \frac{1}{2} \|x - x_k\|^2 \right)$$

Тогда из условия оптимальности:

$$\begin{aligned} \langle \nabla g(x_{k+1}), x_{k+1} - y \rangle &\leq 0 \\ \langle \alpha_k g_k + x_{k+1} - x_k, x_{k+1} - y \rangle &\leq 0 \\ \langle \alpha_k g_k, x_{k+1} - y \rangle + \langle x_{k+1} - x_k, x_{k+1} - y \rangle &\leq 0 \\ \langle \alpha_k g_k, x_{k+1} - y \rangle - \langle x_k - x_{k+1}, x_{k+1} - y \rangle &\leq 0 \end{aligned}$$

Полученное неравенство в точности совпадает с неравенством (Forward Target), которое нам как раз таки и следовало доказать. Таким образом, мы получаем

$$\langle \alpha_k g_k, x \rangle + \frac{1}{2} \|x - x_k\|^2 \rightarrow \min_{x \in S}$$

## Algorithm

PSD

$$x_{k+1} = \arg \min_{x \in S} \left( \langle \alpha_k g_k, x \rangle + \frac{1}{2} \|x - x_k\|^2 \right)$$

Интересные фишки:

- Такая же скорость сходимости, как и для безусловного алгоритма. (Однако, стоимость каждой итерации может быть существенно больше из за необходимости решать задачу оптимизации на каждом шаге)
- В частном случае  $S = \mathbb{R}^n$  в точности совпадает с безусловным алгоритмом (убедитесь)

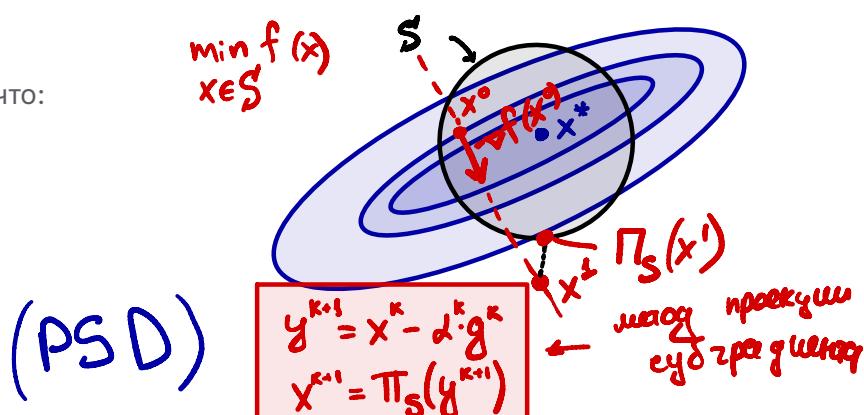
## Adaptive stepsize (without $T$ )

Разберем теперь одну из стратегий того, как избежать знания количества шагов  $T$  заранее для подбора длины шага  $\alpha_k$ . Для этого зададим "диаметр" нашего множества  $D$ :

$$D : \{ \max_{x, y \in S} \|x - y\| \leq D \}$$

Теперь зададим длину шага на  $k$ -ой итерации, как:  $\alpha_k = \tau \sqrt{\frac{1}{k+1}}$ . Константу  $\tau \geq 0$  подберем чуть позже.

Для начала легко заметить, что:



Lemma: Пусть  $S$  - выпуклое замкнутое множество  
 $\forall x, y \in \mathbb{R}^n$ :

$$\|\Pi_S(x) - \Pi_S(y)\| \leq \|x - y\|$$

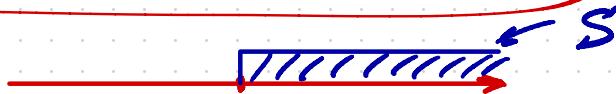
$$\text{т.е. } \Pi_S(t) = \arg \min_{\substack{x \\ \forall x \in S}} \|t - x\|$$

$$\|x^{k+1} - x^*\| \leq \dots$$

$$\|\Pi_S(x^{k+1}) - \Pi_S(x^*)\| \leq \|x^{k+1} - x^*\|$$

FUN FACT: характер сходимости PSD  
тот же, как и у безусловного метода (SD)

сложность в порядке порядка.

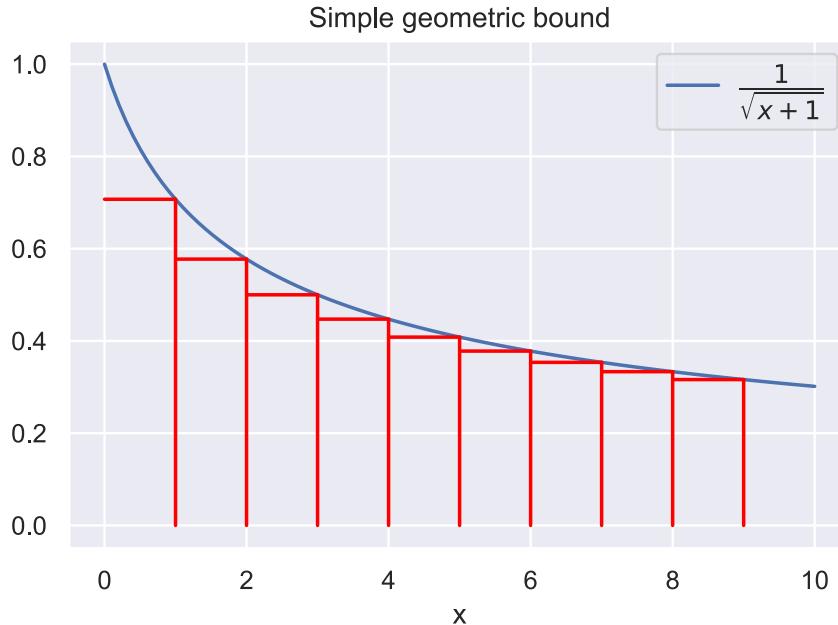


$$\Pi_S^\circ(t) = \begin{cases} t, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

Для (усложнённой) негладкой задачи е оракулом  
1 порядка лучше сделать нельзя.  $\frac{1}{\sqrt{k}}$

$$\begin{aligned}
\sum_{k=0}^{T-1} \alpha_k &= \tau \sum_{k=0}^{T-1} \frac{1}{\sqrt{k+1}} = \tau \left( 1 + \sum_{k=1}^{T-1} \frac{1}{\sqrt{k+1}} \right) \leq \\
&\leq \tau \left( 1 + \int_0^{T-1} \frac{1}{\sqrt{x+1}} dx \right) = \tau(2\sqrt{T} - 1)
\end{aligned}$$

См. геометрический смысл неравенства ниже:



Возьмем теперь равенство для классического субградиентного метода (БМ) (или неравенство в случае метода проекции субградиента (УМ)):

$$\begin{aligned}
2\langle \alpha_k g_k, x_k - x^* \rangle &= \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 g_k^2 \\
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle &= \sum_{k=0}^{T-1} \left( \frac{\|x_k - x^*\|^2}{2\alpha_k} - \frac{\|x_{k+1} - x^*\|^2}{2\alpha_k} + \frac{\alpha_k}{2} g_k^2 \right) \\
&\leq \frac{\|x_0 - x^*\|^2}{2\alpha_0} - \frac{\|x_T - x^*\|^2}{2\alpha_{T-1}} + \\
&\quad + \frac{1}{2} \sum_{k=0}^{T-1} \left( \frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}} \right) \|x_k - x^*\|^2 + \sum_{k=0}^{T-1} \frac{\alpha_k}{2} g_k^2 \leq \\
&\leq D^2 \left( \frac{1}{2\alpha_0} + \frac{1}{2} \sum_{k=0}^{T-1} \left( \frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}} \right) \right) + G^2 \sum_{k=0}^{T-1} \frac{\alpha_k}{2} \leq \\
&\leq \frac{D^2}{2\alpha_{T-1}} + G^2 \sum_{k=0}^{T-1} \frac{\alpha_k}{2} \leq \\
&\leq \frac{1}{2} \left( \frac{D^2}{\tau} \sqrt{T} + \tau G^2 (2\sqrt{T} - 1) \right) \leq \\
&\leq DG\sqrt{2T}
\end{aligned}$$

Где  $\tau = \frac{D}{G\sqrt{2}}$  - выбран путем минимизации данной оценки по  $\tau$ .

Таким образом, мы получили, что в случае, когда количество шагов  $T$  неизвестно заранее (весьма

важное свойство), оценка ухудшается в  $\sqrt{2}$  раз. Такие оценки называют anytime bounds.

## Online learning:

PSD - Projected Subgradient Descent

$$R_{T-1} = \sum_{k=0}^{T-1} f_k(x_k) - \min_{x \in S} \sum_{k=0}^{T-1} f_k(x) \leq DG\sqrt{2T} \quad (\text{anytime PSD})$$

$$R_{T-1} = \sum_{k=0}^{T-1} f_k(x_k) - \min_{x \in S} \sum_{k=0}^{T-1} f_k(x) \leq DG\sqrt{T} \quad (\text{PSD})$$

## Examples

### Least squares with $l_1$ regularization

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

### Nonnegativity

$$S = \{x \in \mathbb{R}^n \mid x \geq 0\}$$

### $l_2$ - ball

$$S = \{x \in \mathbb{R}^n \mid \|x - x_c\| \leq R\}$$

$$x_{k+1} = x_k - \alpha_k (A^\top (Ax_k - b) + \lambda \text{sign}(x_k))$$

### Linear equality constraints

$$S = \{x \in \mathbb{R}^n \mid Ax = b\}$$

## Bounds

Conditions	Convergence rate	Iteration complexity	Type of convergence
Convex Lipschitz-continuous function( $G$ )	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	Sublinear
Strongly convex Lipschitz-continuous function( $G$ )	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$	Sublinear

## References

- Comprehensive [presentation](#) on projected subgradient method.
- [Great cheatsheet](#) by Sebastian Pokutta
- [Lecture](#) on subgradient methods @ Berkley

Метод зеркального спуска является естественным обобщением метода проекции субградиента в случае обобщения  $l_2$  нормы на более общий случай какой-то функции расстояния.

## Dual norm:

Определение: Сопряженной нормой  $\|\cdot\|_*$  к данной  $\|\cdot\|$  называется:

$$\|y\|_* = \max\{\langle y, x \rangle : \|x\| = 1\}$$

Пример:  $(\|\cdot\|_p)_* = \|\cdot\|_q$ ,  $\frac{1}{p} + \frac{1}{q} = 1$

Доказательство:

Неравенство Гельдера:

$$\sum_{k=1}^n |x_k y_k| \leq \left( \sum_{k=1}^n |x_k|^p \right)^{\frac{1}{p}} \left( \sum_{k=1}^n |y_k|^q \right)^{\frac{1}{q}} \text{ for all } x, y \in \mathbb{C}^n$$

Свойства:

- Двойственная норма  $\|\cdot\|_*$  является нормой
- $l_2$  норма сопряжена сама себе
- Двойственная норма к двойственной норме - исходная норма
- $(\|\cdot\|_1)_* = \|\cdot\|_\infty$ ,  $(\|\cdot\|_\infty)_* = \|\cdot\|_1$
- Обобщенное неравенство Коши Шварца:  $\langle y, x \rangle \leq \|y\|_* \|x\|$ , следствие:  $\|x\|^2 \pm 2\langle y, x \rangle + \|y\|_*^2 \geq 0$

## Bregman divergence

Попробуем интуитивно ввести понятие обобщенного расстояния, именуемого расстоянием Брэгмана. Для каждой точки  $y$  она возвращает расстояние этой точки до  $x$  -  $V_x(y)$ . В самом простом случае можно взять

$V_x(y) = \frac{1}{2} \|x - y\|^2$ ,  $\nabla V_x(y) = y - x$ . Рассмотрим уже классическую запись:

$$\begin{aligned} \|x_{k+1} - y\|^2 &= \|x_{k+1} - x_k\|^2 + \|x_k - y\|^2 - 2\langle x_k - x_{k+1}, x_k - y \rangle \\ V_{x_{k+1}}(y) &= V_{x_{k+1}}(x_k) + V_{x_k}(y) - \langle \nabla V_{x_{k+1}}(x_k), x_k - y \rangle \end{aligned} \quad (\text{Req1})$$

Для вводимого обобщенного расстояния будем требовать выполнения (Req1), кроме того (как будет видно при получении оценок), приятным свойством было бы еще следующее требование:

$$V_x(y) \geq \frac{1}{2} \|x - y\|^2 \quad (\text{Req2})$$

**Определение:** Дивергенцией (расстоянием) Брэгмана называется функция следующая  $V_x(y)$ . Пусть  $S \subseteq \mathbb{R}^n$  - замкнутое выпуклое множество, тогда функция  $\phi : S \rightarrow \mathbb{R}$  называется прокс-функцией (distance generating function), если  $\phi$  является 1 - сильно выпуклой, т.е.:

$$\phi(y) \geq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{1}{2} \|y - x\|^2, \quad \forall x, y \in S$$

Тогда прокс-функцией индуцируется **расстояние Брэгмана**:

$$V_x(y) = \phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle$$

Заметим, что определение сильной выпуклости зависит от выбора прямой нормы  $\|\cdot\|$ . Это важное замечание, поскольку именно это свойство позволит в будущем подстраивать расстояние под геометрию пространства.

## Examples

- Выберем норму в прямом пространстве  $\|\cdot\| = \|\cdot\|_2$ , пусть  $\phi(x) = \frac{1}{2} \|x\|^2$ , тогда расстояние Брэгмана  $V_x(y) = \frac{1}{2} \|x - y\|^2$ . Такой выбор совпадает с тем, что мы видели ранее в методе проекции субградиента
- Выберем теперь другую норму  $\|\cdot\| = \|\cdot\|_1$ , пусть  $\phi(x) = \sum_{i \in [n]} x_i \log x_i$  - антиэнтропия. Тогда эта функция будет 1 сильно выпукла на выпуклом множестве  $S : \left\{ x \in S : x \geq 0, \sum_{i \in [n]} x_i = 1 \right\}$  (вероятностном симплексе), а соответствующая ей дивергенция Брэгмана:  $V_x(y) = \sum_{i \in [n]} y_i \log \frac{y_i}{x_i} = D(y\|x)$  - расстояние Кульбака - Ляйблера.
- Еще немного примеров [отсюда](#):

TABLE 2.1  
Common seed functions and the corresponding divergences.

Function name	$\phi(x)$	$\text{dom } \phi(x)$	$V_x(y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty, +\infty)$	$\frac{1}{2}(x - y)^2$
Shannon entropy	$x \log x - x$	$[0, +\infty)$	$x \log \frac{x}{y} - x + y$
Bit entropy	$x \log x + (1 - x) \log(1 - x)$	$[0, 1]$	$x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0, +\infty)$	$\frac{x}{y} - \log \frac{x}{y} - 1$
Hellinger	$-\sqrt{1 - x^2}$	$[-1, 1]$	$(1 - xy)(1 - y^2)^{-1/2} - (1 - x^2)^{1/2}$
$\ell_p$ quasi-norm	$-x^p$ ( $0 < p < 1$ )	$[0, +\infty)$	$-x^p + pxy^{p-1} - (p - 1)y^p$
$\ell_p$ norm	$ x ^p$ ( $1 < p < \infty$ )	$(-\infty, +\infty)$	$ x ^p - px \operatorname{sgn} y  y ^{p-1} + (p - 1) y ^p$
Exponential	$\exp x$	$(-\infty, +\infty)$	$\exp x - (x - y + 1)\exp y$
Inverse	$1/x$	$(0, +\infty)$	$1/x + x/y^2 - 2/y$

TABLE 2.2  
Common exponential families and the corresponding divergences.

Exponential family	$\psi(\theta)$	$\text{dom } \psi$	$\mu(\theta)$	$\phi(x)$	Divergence
Gaussian ( $\sigma^2$ fixed)	$\frac{1}{2}\sigma^2\theta^2$	$(-\infty, +\infty)$	$\sigma^2\theta$	$\frac{1}{2\sigma^2}x^2$	Euclidean
Poisson	$\exp \theta$	$(-\infty, +\infty)$	$\exp \theta$	$x \log x - x$	Relative entropy
Bernoulli	$\log(1 + \exp \theta)$	$(-\infty, +\infty)$	$\frac{\exp \theta}{1 + \exp \theta}$	$x \log x + (1 - x) \log(1 - x)$	Logistic loss
Gamma ( $\alpha$ fixed)	$-\alpha \log(-\theta)$	$(-\infty, 0)$	$-\alpha/\theta$	$-\alpha \log x + \alpha \log \alpha - \alpha$	Itakura–Saito

## Свойства

- Аксиома тождества  $V_x(x) = 0$
- Совместимость с Евклидовой нормой:  $V_x(y) \geq \frac{1}{2}\|x - y\|^2 \geq 0$
- (Не)равенство треугольника:  $\langle -\nabla V_x(y), y - z \rangle = V_x(z) - V_y(z) - V_x(y)$

Первые два свойства очевидны из определения. Докажем третье:

$$\begin{aligned}
 \langle -\nabla V_x(y), y - z \rangle &= \langle \nabla \phi(x) - \nabla \phi(y), y - z \rangle = \\
 &= (\phi(z) - \phi(x) - \langle \nabla \phi(x), z - x \rangle) \\
 &\quad - (\phi(z) - \phi(y) - \langle \nabla \phi(y), z - y \rangle) \\
 &\quad - (\phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle) \\
 &= V_x(z) - V_y(z) - V_x(y)
 \end{aligned}$$

## Возвращение к истокам

Пусть задано выпуклое замкнутое множество  $S \in \mathbb{R}^n$ , кроме того, есть алгоритм оптимизации, возвращающий последовательность точек  $x_1, \dots, x_k, \dots$ . Тогда запишем (не)равенство треугольника для расстояния Брэгмана, полагая  $y = x_{k+1}$ ,  $x = x_k$  и произвольный  $z \in S$  (который мы в дальнейшем для целостности изложения будем обозначать  $y$ )

$$\begin{aligned}
 \langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - z \rangle &= V_{x_k}(z) - V_{x_{k+1}}(z) - V_{x_k}(x_{k+1}) \\
 \langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle &= V_{x_k}(y) - V_{x_{k+1}}(y) - V_{x_k}(x_{k+1}) \quad (\text{baseMD})
 \end{aligned}$$

Просуммируем полученные равенства:

$$\sum_{k=0}^{T-1} \langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle = V_{x_0}(y) - V_{x_T}(y) - \sum_{k=0}^{T-1} V_{x_k}(x_{k+1})$$

Имея ввиду полученное уравнение, давайте, наконец, попробуем сформулировать метод **зеркального спуска**:

$$x_{k+1} = \arg \min_{x \in S} (\langle \alpha_k g_k, x \rangle + V_{x_k}(x))$$

Посмотрим внимательнее на условие проекции для точки  $x_{k+1}$ :

$$\langle \alpha_k g_k, x_{k+1} - y \rangle + \langle \nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle \leq 0$$

$$\langle \alpha_k g_k, x_{k+1} - y \rangle \leq -\langle \nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle$$

Попробуем теперь получить наше базовое неравенство, используя (baseMD):

$$\begin{aligned} \langle \alpha_k g_k, x_k - y \rangle &\leq -\langle \nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle - \langle \alpha_k g_k, x_{k+1} - x_k \rangle = \\ &= V_{x_k}(y) - V_{x_{k+1}}(y) - V_{x_k}(x_{k+1}) - \langle \alpha_k g_k, x_{k+1} - x_k \rangle \leq \\ &\leq V_{x_k}(y) - V_{x_{k+1}}(y) - \frac{1}{2} \|x_k - x_{k+1}\|^2 - \langle \alpha_k g_k, x_{k+1} - x_k \rangle \leq \\ &\leq V_{x_k}(y) - V_{x_{k+1}}(y) + \left( \langle \alpha_k g_k, x_k - x_{k+1} \rangle - \frac{1}{2} \|x_k - x_{k+1}\|^2 \right) \leq \\ &\leq V_{x_k}(y) - V_{x_{k+1}}(y) + \frac{\alpha_k^2}{2} \|g_k\|_*^2 \end{aligned}$$

ТЕЛЕСКОПИРУЕМ

$$\begin{aligned} \sum_{k=0}^{T-1} \langle \alpha_k g_k, x_k - y \rangle &\leq V_{x_0}(y) - V_{x_T}(y) + \sum_{k=0}^{T-1} \frac{\alpha_k^2}{2} \|g_k\|_*^2 \\ &\leq V_{x_0}(y) + \sum_{k=0}^{T-1} \frac{\alpha_k^2}{2} \|g_k\|_*^2 \\ &\leq M + \frac{\alpha^2 G^2 T}{2} \end{aligned}$$

Здесь мы подразумеваем  $\|g_k\|_* \leq G$  равномерно по  $k$ , а  $V_{x_0}(y) \leq M$ . В итоге:

$$\begin{aligned}
f(\bar{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T} \left( \sum_{k=0}^{T-1} f(x_k) - f^* \right) \\
&\leq \frac{1}{T} \left( \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\
&\leq \frac{M}{\alpha T} + \frac{\alpha G^2}{2} \leq \sqrt{\frac{2MG^2}{T}}
\end{aligned}$$

Выбирая шаг  $\alpha_k = \alpha = \sqrt{\frac{2M}{G^2 T}}$

## Алгоритм зеркального спуска (mirror descent):

$$x_{k+1} = \arg \min_{x \in S} (\langle \alpha_k g_k, x \rangle + V_{x_k}(x))$$

Интересные фишки:

- Такая же скорость сходимости, как и для метода проекции субградиента.
- Работает в существенно более широком классе практических задач

## Онлайн версия

Совершенно ясно, что в наших оценках на каждом шаге может быть новая функция  $f_k(x)$  на заданном классе. Поэтому, аналогичные оценки получаются и для онлайн постановки:

$$R_{T-1} = \sum_{k=0}^{T-1} f_k(x_k) - \min_x \sum_{k=0}^{T-1} f_k(x) \leq \sqrt{2MG^2 T}$$

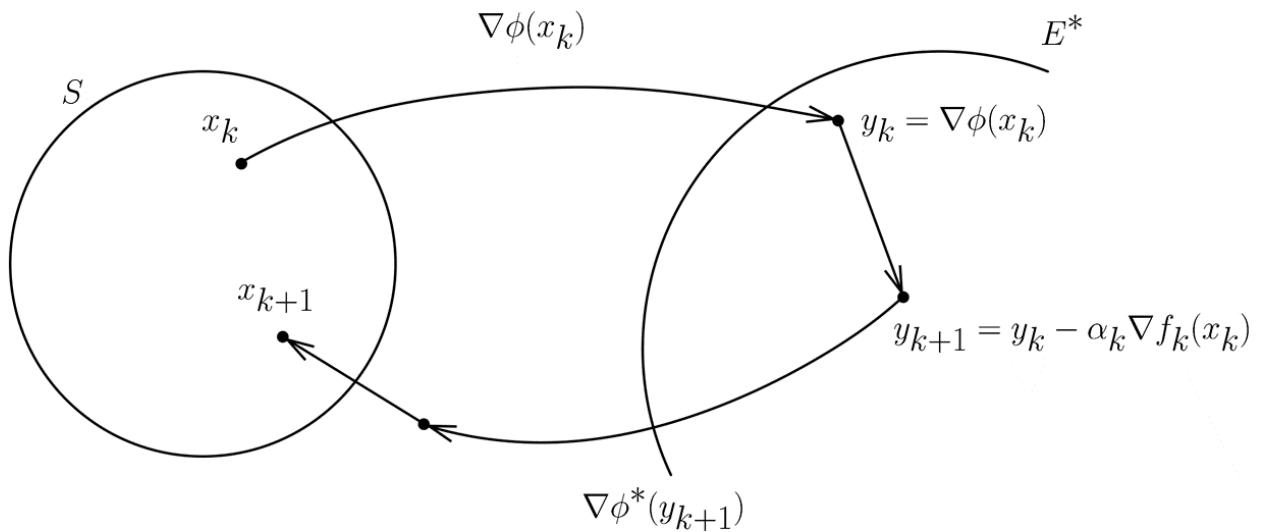
$$\overline{R_{T-1}} = \frac{1}{T} R_{T-1} \leq \sqrt{\frac{2MG^2}{T}}$$

## Еще одна интерпретация

Давайте покажем, что полученный алгоритм имеет еще одну очень интуитивную интерпретацию:

- 1  $y_k = \nabla \phi(x_k)$  Отображение в сопряженное пространство с помощью функции  $\nabla \phi(x)$

- 2  $y_{k+1} = y_k - \alpha_k \nabla f_k(x_k)$  Градиентный шаг в сопряженном пространстве
- 3  $x_{k+1} = \arg \min_{x \in S} V_{\nabla \phi^*(y_{k+1})}(x)$  Обратное отображение с помощью функции  $\nabla \phi^*(x)$  и проекция на бюджетное множество



Для доказательства эквивалентности таких записей, следует сначала доказать факт того, что:

$$(\nabla \phi(x))^{-1} = \nabla \phi^*(y)$$

Для этого пусть  $y = \nabla \phi(x)$ . Заметим, что для сопряженной функции справедливо неравенство Фенхеля - Юнга:  $\phi^*(y) + \phi(x) \geq xy$ , в случае, если  $\phi(x)$  - дифференцируема, такое преобразование называется преобразованием Лежандра и выполняется равенство:  $\phi^*(y) + \phi(x) = xy$ . Дифференцируя равенство по  $y$ , получаем  $\nabla \phi^*(y) = x$ . Таким образом,

$$\nabla \phi^*(y) = \nabla \phi^*(\nabla \phi(x)) = x, \quad \nabla \phi(x) = \nabla \phi(\nabla \phi^*(y)) = y$$

Доказательство:

$$\begin{aligned}
x_{k+1} &= \arg \min_{x \in S} \{V_{\nabla \phi^*(y_{k+1})}(x)\} = \\
&= \arg \min_{x \in S} \{\phi(x) - \phi(\nabla \phi^*(y_{k+1})) - \langle \nabla \phi(\nabla \phi^*(y_{k+1})), x - \nabla \phi^*(y_{k+1}) \rangle\} = \\
&= \arg \min_{x \in S} \{\phi(x) - \langle y_{k+1}, x \rangle\} = \\
&= \arg \min_{x \in S} \{\phi(x) - \langle \nabla \phi(x_k) - \alpha_k g_k, x \rangle\} = \\
&= \arg \min_{x \in S} \{\phi(x) - \phi(x_k) - \langle \nabla \phi(x_k), x \rangle + \langle \alpha_k g_k, x \rangle\} = \\
&= \arg \min_{x \in S} \{V_{x_k}(x) + \langle \alpha_k g_k, x \rangle\}
\end{aligned}$$

В последней строчке мы пришли к той формулировке, которую писали раньше.

Заметим так же, еще одну интересную концепцию:

$$\begin{aligned}
x_{k+1} &= \arg \min_{x \in S} (\langle \alpha_k g_k, x \rangle + V_{x_k}(x)) \\
&= \arg \min_{x \in S} \left( \langle g_k, x \rangle + \frac{1}{\alpha_k} V_{x_k}(x) \right) \\
&= \arg \min_{x \in S} \left( f(x_k) + \langle g_k, x \rangle + \frac{1}{\alpha_k} V_{x_k}(x) \right)
\end{aligned}$$

Здесь левая часть минимизируемого выражения представляет собой аппроксимацию первого порядка, а правая часть представляет собой проекционный член.

## НАФИГА?

Резонный вопрос, ведь в случае, если мы выбрали  $\|\cdot\| = \|\cdot\|_2$  Евклидову норму и Евклидово расстояние, то этот метод в точности совпадает с тем, что мы уже называем метод проекции субградиента.

Значит, надо предоставить сценарий, когда МЗС работает лучше, давайте рассмотрим  $S = \Delta_n$  - вероятностный симплекс, а так же следующее расстояние Брэгмана  $V_x(y) = \sum_{i \in [n]} y_i \log \frac{y_i}{x_i} = D(y \| x)$ . Норма в прямом пространстве при этом  $\|\cdot\|_1$ , а в сопряженном -  $\|\cdot\|_\infty$ . Кроме того, для заданной дивергенции Брэгмана справедливо:

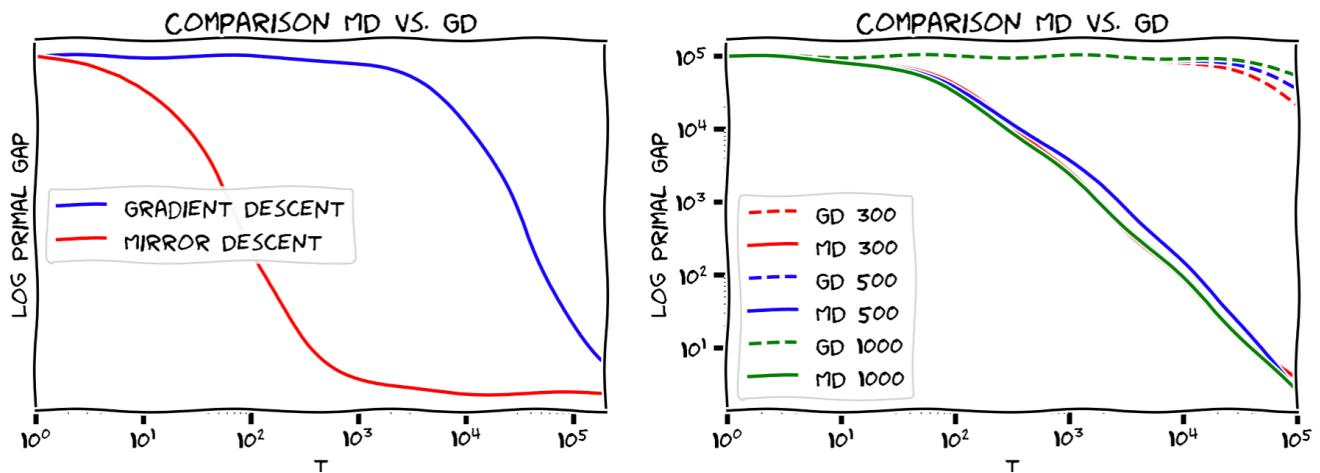
$$x_0 = (1/n, \dots, 1/n) \rightarrow V_{x_0}(x) \leq \log n \quad \forall x \in \Delta_n$$

Тогда алгоритм зеркального спуска запишется в виде:

$$\begin{aligned}
x_{k+1} &= \arg \min_{x \in S} (\langle \alpha_k g_k, x \rangle + V_{x_k}(x)) \\
&= \arg \min_{x \in S} \left( \langle \alpha_k g_k, x \rangle + \sum_{i \in [n]} x_i \log \frac{x_i}{x_{k_i}} \right) \\
&= x_k \cdot \frac{e^{-\alpha_k g_k}}{\|x_k \cdot e^{-\alpha_k g_k}\|_1}
\end{aligned}$$

А оценки с учетом того, что  $M = \log n$ ,  $\|g_k\|_\infty \leq G$  запишутся, как:

$$f(\bar{x}) - f^* \leq \sqrt{\frac{2 \log(n) G^2}{T}}$$



```

import numpy as np
from matplotlib import pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set()

Ts = np.logspace(0, 4, 10)

m = 10
n = 1000
A = np.random.randn(m, n)
x_true = np.random.randn(n)
x_true[x_true < 0] = 0
x_true = x_true / (np.linalg.norm(x_true, 1))

```

```

b = A@x_true

x0 = np.ones(n)/n

def f(x):
    return np.linalg.norm(A@x - b, 1)

def grad(x):
    return np.sum(A.T * np.sign(A@x - b), axis=1)

def mirror_descent(x0, grad, T):
    n = len(x0)
    M = np.log(n)
    # G = np.linalg.norm(A,np.inf)*1+np.linalg.norm(b,np.inf)
    # alpha = np.sqrt(2*M/(G**2*T))
    alpha = 0.0001
    xk = x0
    sequence = []
    # print('MD %.3f'%alpha)
    for i in range(int(T)):
        sequence.append(xk)
        g = grad(xk)
        xk = xk * np.exp(-alpha * g) / np.sum(xk * np.exp(-alpha * g))
    return sequence

def projection_subgradient_descent(x0, grad, T):
    n = len(x0)
    M = 0.5
    # G = np.linalg.norm(A,2)*1+np.linalg.norm(b,2)
    # alpha = np.sqrt(2*M/(G**2*T))
    alpha = 0.0001
    # print('GD %.3f'%alpha)
    xk = x0
    sequence = []
    for i in range(int(T)):
        sequence.append(xk)
        g = grad(xk)
        xk = xk - alpha*g

```

```

xk[xk<0] = 0
xk = xk/(np.linalg.norm(xk, 1))

return sequence

result_md = []
result_gd = []

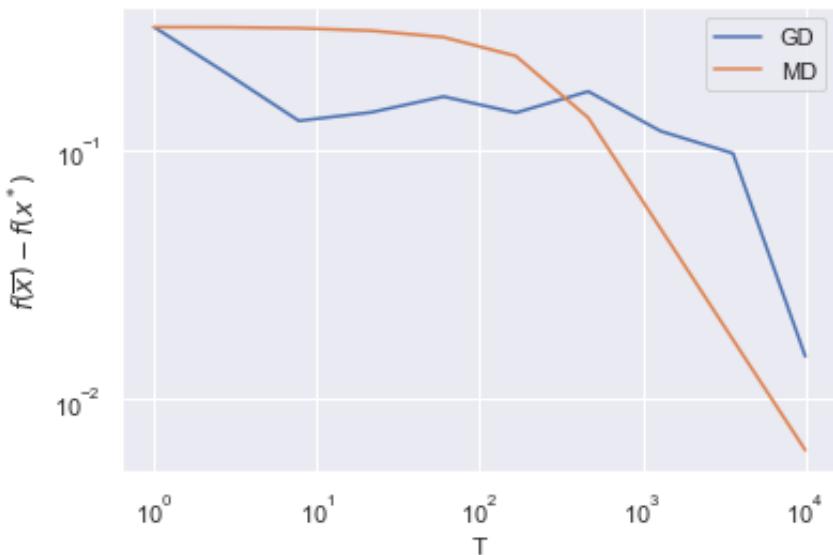
for T in Ts:
    print(T)
    md_T = mirror_descent(x0, grad, T)
    gd_T = projection_subgradient_descent(x0, grad, T)

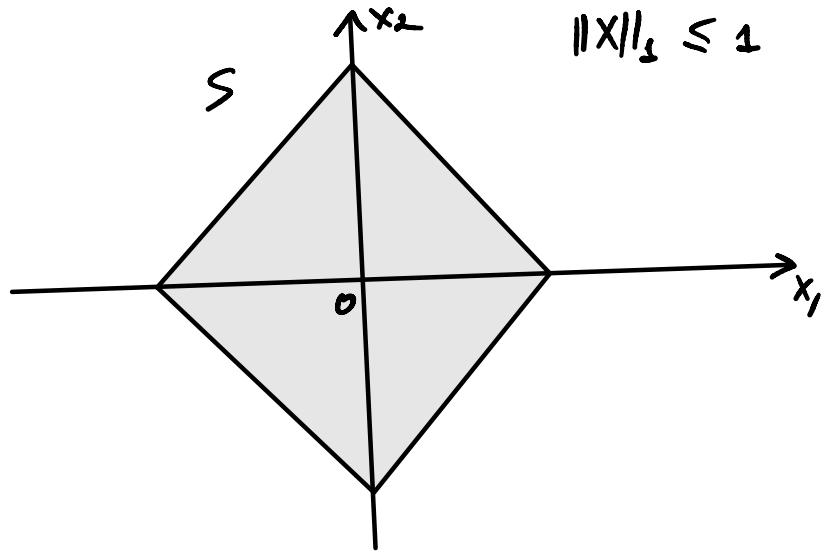
    x_md = np.mean(md_T, axis = 0)
    x_gd = np.mean(gd_T, axis = 0)

    result_md.append(f(x_md) - f(x_true))
    result_gd.append(f(x_gd) - f(x_true))

plt.loglog(Ts, result_gd, label = 'GD')
plt.loglog(Ts, result_md, label = 'MD')
plt.xlabel('T')
plt.ylabel(r'$f(\overline{x}) - f(x^*)$')
plt.legend()

```





# Projection

## Distance between point and set

The distance  $d$  from point  $\mathbf{y} \in \mathbb{R}^n$  to closed set  $S \subset \mathbb{R}^n$ :

$$d(\mathbf{y}, S, \|\cdot\|) = \inf\{\|\mathbf{x} - \mathbf{y}\| \mid \mathbf{x} \in S\}$$

## Projection of a point on set

Projection of a point  $\mathbf{y} \in \mathbb{R}^n$  on set  $S \subseteq \mathbb{R}^n$  is a point  $\pi_S(\mathbf{y}) \in S$ :

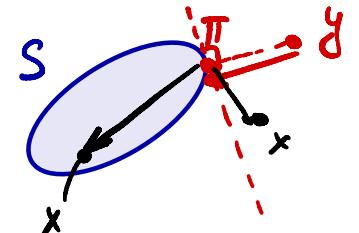
$$\|\pi_S(\mathbf{y}) - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x} \in S$$

- if set is open, and a point is beyond this set, then its projection on this set does not exist.
- if a point is in set, then its projection is the point itself

$$\pi_S(\mathbf{y}) = \underset{\mathbf{x} \in S}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{y}\|$$

- Let  $S \subseteq \mathbb{R}^n$  - convex closed set. Let the point  $\mathbf{y} \in \mathbb{R}^n$  и  $\pi \in S$ . Then if for all  $\mathbf{x} \in S$  the inequality holds:

$$\langle \pi - \mathbf{y}, \mathbf{x} - \pi \rangle \geq 0,$$



then  $\pi$  is the projection of the point  $\mathbf{y}$  on  $S$ , so  $\pi_S(\mathbf{y}) = \pi$ .

- Let  $S \subseteq \mathbb{R}^n$  - affine set. Let we have points  $\mathbf{y} \in \mathbb{R}^n$  and  $\pi \in S$ . Then  $\pi$  is a projection of point  $\mathbf{y}$  on  $S$ , so  $\pi_S(\mathbf{y}) = \pi$  if and only if for all  $\mathbf{x} \in S$  the inequality holds:

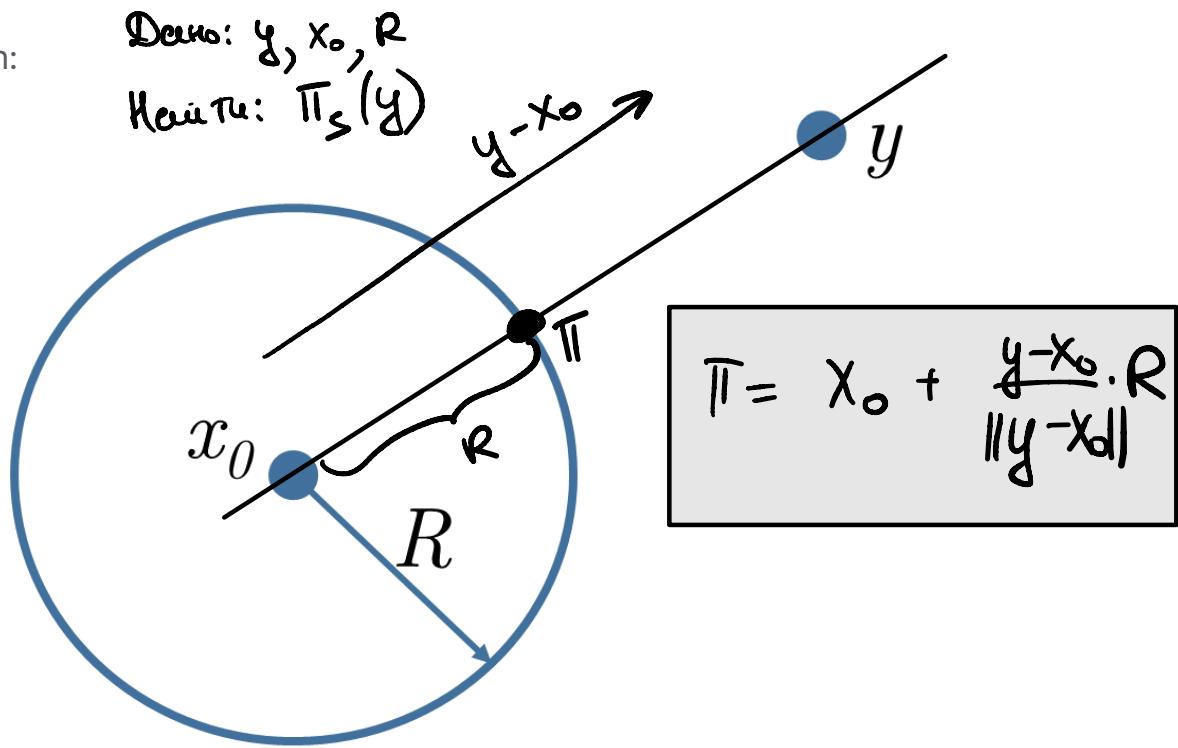
$$\langle \pi - \mathbf{y}, \mathbf{x} - \pi \rangle = 0$$

- Sufficient conditions of existence of a projection.** If  $S \subseteq \mathbb{R}^n$  - closed set, then for all points exists projection on set  $S$ .
- Sufficient conditions of uniqueness of a projection.** Если  $S \subseteq \mathbb{R}^n$  - convex set, then projection for all point on set  $S$  is unique (if exists).

### EXAMPLE 1

Find  $\pi_S(y) = \pi$ , if  $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$ ,  $y \notin S$

Solution:



- Build a hypothesis from the figure:  $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$
- Check the inequality for a convex closed set:  $(\pi - y)^T(x - \pi) \geq 0$

$$\begin{aligned}
 & \left( x_0 - y + R \frac{y - x_0}{\|y - x_0\|} \right)^T \left( x - x_0 - R \frac{y - x_0}{\|y - x_0\|} \right) = \\
 & \left( \frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|} \right)^T \left( \frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|} \right) = \\
 & \frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T ((x - x_0)\|y - x_0\| - R(y - x_0)) = \\
 & \frac{R - \|y - x_0\|}{\|y - x_0\|} \left( (y - x_0)^T (x - x_0) - R\|y - x_0\| \right) = \\
 & (R - \|y - x_0\|) \left( \frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R \right)
 \end{aligned}$$

- The first factor is negative for point selection  $y$ . The second factor is also negative, which follows from the Cauchy-Bunyakovsky inequality:

$$(y - x_0)^T (x - x_0) \leq \|y - x_0\| \|x - x_0\|$$

$$\frac{(y - x_0)^T(x - x_0)}{\|y - x_0\|} - R \leq \frac{\|y - x_0\|\|x - x_0\|}{\|y - x_0\|} - R = \|x - x_0\| - R \leq 0$$

EXAMPLE 2

Find  $\pi_S(y) = \pi$ , if  $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$ ,  $y \notin S$ .

Demo: Y, C

Solution:

$$c^T \pi = b$$

$$\pi = y - ct$$

$$y - t = \frac{c^T y - b}{c^T c}$$

$$\pi = y - \frac{c^T y - b}{c^T c} c$$

$$\frac{c^T \pi}{b} = \underbrace{c^T y}_{0} - \underbrace{(c^T y - b)}_{b}$$

$$\pi$$

$$c$$

$$1) \quad \pi = y - c \cdot t$$

непр.

2) Норма робок

$$c^T \pi = b$$

$$c^T (y - ct) = b$$

$$\hookrightarrow t = \dots$$

- Build a hypothesis from the figure:  $\pi = y + \alpha c$ . Coefficient  $\alpha$  is chosen so that  $\pi \in S: c^T \pi = b$ , so:

$$c^T (y + \alpha c) = b$$

$$c^T y + \alpha c^T c = b$$

$$c^T y = b - \alpha c^T c$$

- Check the inequality for a convex closed set:  $(\pi - y)^T(x - \pi) \geq 0$

$$(y + \alpha c - y)^T(x - y - \alpha c) =$$

$$\alpha c^T (x - y - \alpha c) =$$

$$\alpha(c^T x) - \alpha(c^T y) - \alpha^2(c^T c) =$$

$$\alpha b - \alpha(b - \alpha c^T c) - \alpha^2 c^T c =$$

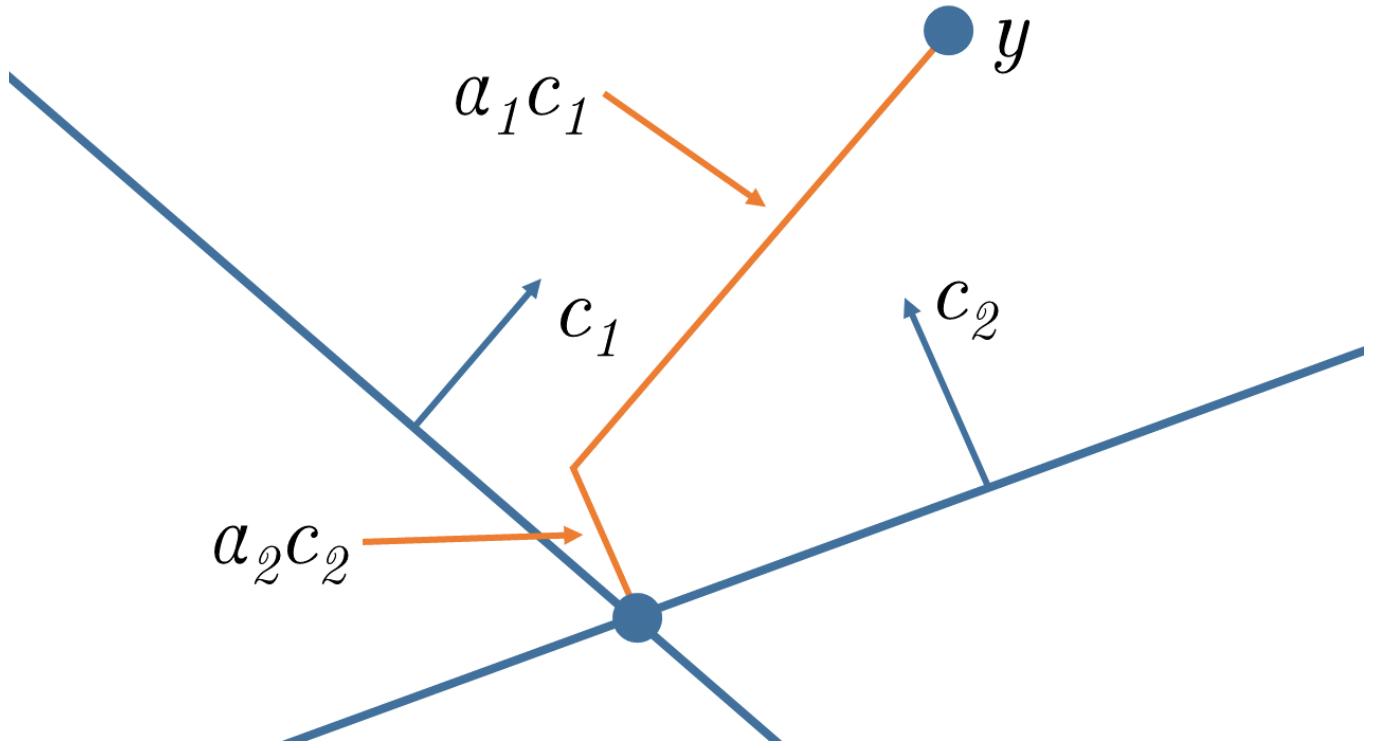
$$\alpha b - \alpha b + \alpha^2 c^T c - \alpha^2 c^T c = 0 \geq 0$$

EXAMPLE 3

Find  $\pi_S(y) = \pi$ , if  $S = \{x \in \mathbb{R}^n \mid Ax = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m\}$ ,  $y \notin S$ .

Solution:

$$\text{rg } A = n$$



- Build a hypothesis from the figure:  $\pi = y + \sum_{i=1}^m \alpha_i A_i = y + A^T \alpha$ . Coefficient  $\alpha$  is chosen so that  $\pi \in S$ :  $A\pi = b$  so:

$$A(y + A^T \alpha) = b$$

$$Ay = b - AA^T \alpha$$

- Check the inequality for a convex closed set:  $(\pi - y)^T(x - \pi) \geq 0$

$$(y + A^T \alpha - y)^T(x - y - A^T \alpha) =$$

$$\alpha^T A(x - y - A^T \alpha) =$$

$$\alpha^T(Ax) - \alpha^T(Ay) - \alpha^T(AA^T \alpha) =$$

$$\alpha^T b - \alpha^T(b - AA^T \alpha) - \alpha^T AA^T \alpha =$$

$$\alpha^T b - \alpha^T b + \alpha^T AA^T \alpha - \alpha^T AA^T \alpha = 0 \geq 0$$

# Метод условного градиентного (Франка - Вульфа)

$$\min_{x \in S} f(x)$$

ищем в т.  $x^k$

$$f_{x^k}^I(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$$

на каждом шаге ищем

$$\min_{x \in S} f_{x^k}^I(x) \Leftrightarrow$$

$$\Leftrightarrow \min_{x \in S} \left( f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \right)$$

$$\Leftrightarrow \boxed{\min_{x \in S} \langle \nabla f(x^k), x \rangle}$$

Алгоритм Франка - Вульфа

$$1. S^k = \operatorname{argmin}_{x \in S} \langle \nabla f(x^k), x \rangle$$

$$2. x^{k+1} = x^k + \lambda^k (S^k - x^k)$$

