Расстояния между распределениями

f-дивергенции. Расстояние полной вариации. Расстояние Кульбака-Лейблера. Расстояние Йенсена-Шеннона. χ^2 расстояние. Расстояние Васерштейна.

ПМИ ФКН ВШЭ, 22 сентября 2018 г.

Денис Деркач[1]

[1] ФКН ВШЭ

Оглавление

f-дивергенции

Мотивация

Мы хотим по выборке экспериментальных точек $X_1, .., X_n \sim F_\theta$ оценить значение θ. Для этого необходимо построить некоторую оценку. Эта оценка должна быть достаточно качественной.

Напоминание

Для проверки качества оценки необходимо ввести функцию:

$$\begin{aligned} \ell : \mathcal{Y} \times \hat{\mathcal{Y}} &\to \mathcal{R}, \\ T \times \hat{T} &\mapsto \ell(T, \hat{T}). \end{aligned}$$

Какие функции можно предложить?

Определение

Определение

Пусть P и Q - распределение вероятностей на $\mathcal{X}$, причём P абсолютно непрерывна над Q. Тогда для выпуклой функции $f : (0, \infty) \to \mathbb{R}$, которая строго выпукла в 1 и $f(1) = 0$, f-дивергенцией P над Q называется:

$$D_f(P \parallel Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ.$$

Определение

Для обычных непрерывных распределений, определение можно переписать в виде:

$$D_f(P \parallel Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\, d\mu(x).$$

NB: для дискретного случая, определение:

$$D_f(P \parallel Q) \equiv \sum_{x\in\mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right).$$

Рассуждения

f-дивергенция не является расстоянием так как вообще говоря изменяется в случае замены $P \leftrightarrow Q$. Однако мы можем найти такие функции f, которые обладают нужным нам свойством.

Основные свойства

- Неотрицательность: $D_f(P \parallel Q)) \geq 0$ причём $D_f = 0$ тогда и только тогда, когда $P = Q$.
- Выпуклость: $(P, Q) \mapsto D_f(P \parallel Q))$ выпуклая функция.
- Возрастание на условных распределениях.

Часто встречающиеся расстояния

Напоминание: для определения нового расстояния достаточно задать функцию f, которая удовлетворяла бы условиям:

- f выпуклая;
- $f(1) = 0$;
- f строго выпуклая в точке $x = 1$, то есть для любых x, y, α, таких что $\alpha x + \alpha y = 1$, выполнено неравенство $f(1) < \alpha f(x) + \alpha f(y)$.

Часто встречающиеся расстояния

- расстояние полной вариации $f(t) = \frac{1}{2}|t - 1|$;
- расстояние Кульбака-Лейблера: $f(t) = t \log t$;
- расстояние Хеллингера $f(t) = (\sqrt{t} - 1)^2$;
- расстояние χ^2: $f(t) = t^2 - 1$;
- расстояние Йенсена-Шеннона*;
- расстояние Васерштейна*.

* - определения будут даны ниже.

Расстояние полной вариации

Определение расстояния

Определение

Расстояние полной вариации для случайных величин X и Y с плоностями f_X и g_Y определяется следующим образом:

$$D(f_X, g_Y) = \mathrm{sup_A} \left| \int_{\mathrm{A}} \mathrm{f_X(x)dx} - \int_{\mathrm{A}} \mathrm{g_Y(y)dy} \right|,$$

где sup вычисляется по всем измеримым множествам A.

Теорема Шеффе

Теорема

Если существуют плотности распределения $p_\xi(x)$, $p_\eta(x)$, $x \in \mathbb{R}^n$ случайных величин ξ и η, то:

$$D(p_\xi, p_\eta) = \frac{1}{2} \int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)| dx,$$

Теорема Шеффе (1/2)

Доказательство.

Пусть $A_o = \{x : p_\xi(x) \geq p_\eta(x)\}$ и $A_o^c = \{x : p_\xi(x) < p_\eta(x)\}$.
Тогда $D(p_\xi, p_\eta) \geq \int_{A_o} [p_\xi(x) - p_\eta(x)]dx$.
С другой стороны:

$$\int_{\mathbb{R}^p} |p_\xi(x) - p_\eta(x)|dx =$$
$$= \int_{A_o} [p_\xi(x) - p_\eta(x)]dx - \int_{A_o^c} [p_\xi(x) - p_\eta(x)]dx =$$
$$= 2 \int_{A_o} [p_\xi(x) - p_\eta(x)]dx,$$

то есть $D(p_\xi, p_\eta) \geq \frac{1}{2} \int_{\mathbb{R}^n} [p_\xi(x) - p_\eta(x)]dx$ $\square$

Теорема Шеффе (2/2)

Доказательство.

Заметим, что

$$\left|\int_A [p_\xi(x) - p_\eta(x)]dx\right| =$$

$$= \left|\int_{A\cap A_o} [p_\xi(x) - p_\eta(x)]dx + \int_{A\cap A_o^c} [p_\xi(x) - p_\eta(x)]dx\right| =$$

$$= \left|\int_{A\cap A_o} [p_\xi(x) - p_\eta(x)]dx - \int_{A\cap A_o^c} [p_\eta(x) - p_\xi(x)]dx\right| \leq$$

$$\leq \max\left(\int_{A\cap A_o} [p_\xi(x) - p_\eta(x)]dx, \int_{A\cap A_o^c} [p_\eta(x) - p_\xi(x)]dx\right) \leq$$

$$\leq \max\left(\int_{A_o} [p_\xi(x) - p_\eta(x)]dx, \int_{A_o^c} [p_\eta(x) - p_\xi(x)]dx\right) =$$

$$= \frac{1}{2}\int_{\mathbb{R}^n} [p_\xi(x) - p_\eta(x)]dx$$

Таким образом, верхяя оценка совпадает с нижней. □

Свойства

(+) Дистанция симметрична: $D(f_X, g_Y) = D(g_Y, f_X)$.

(+) Дистанция связана с ошибками при тестировании гипотез: $1 - D(f_X, g_Y)$ равна сумме false positive и false negative примеров.

(+) При увеличении числа испытаний, n, дистанция $D(f_{X^n}, g_{Y^n}) \to 1$. Больше того, если $D(f_X, g_Y) = \delta$, то для любого $k \in \mathbb{N}$ выполнено: $1 - 2e^{-k\frac{\delta^2}{2}} \leq D(f_{X^n}, g_{Y^n})$.

(−) Иногда дистанция может не реагировать на добавление семплов: $D(f_{X^2}, g_{Y^2}) = D(f_X, g_Y)$ (например, распределение Бернулли).

Расстояние Кульбака-Лейблера

Определение

Определение

Расстоянием Кульбака-Лейблера между плотностями вероятностей f_X и g_Y называют:

$$KL(f_X, g_Y) = \int_{\mathbb{R}^n} f_X(z) \log\left(\frac{f_X(z)}{g_Y(z)}\right) dz.$$

Несимметричность

Легко заметить, что:

$$KL(f_X, g_Y) \neq KL(g_Y, f_X).$$

Полезно вспомнить, что в реальности мы измеряем дистанцию плотности вероятностей для разных объектов: эмпирической плотности вероятностей и элемента параметрического семейства.

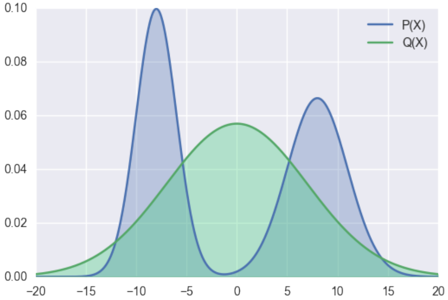
Обратная метрика

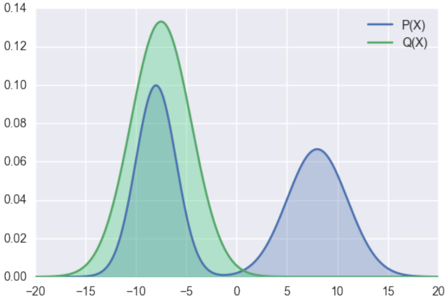
Можно также определить обратное расстояние Кульбака-Лейблера:

$$rKL(f_X, g_Y) = \int_{\mathbb{R}^n} g_X(z) \log\left(\frac{g_X(z)}{f_Y(z)}\right) dz.$$

Пример

$$KL = \int P(X) \log \frac{P(x)}{Q(x)} dx \qquad rKL = \int Q(x) \log \frac{Q(x)}{P(x)} dx$$

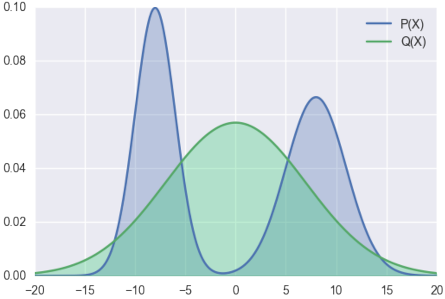




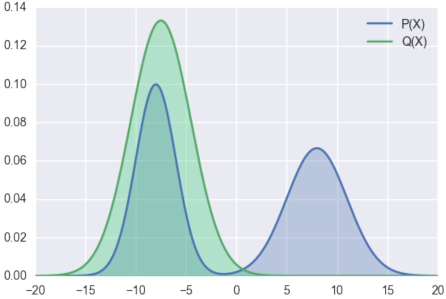
Picture credit:https://wiseodd.github.io/techblog/2016/12/21/forward-reverse-kl/

Пример

$$KL = \int P(X) \log \frac{P(x)}{Q(x)} dx$$



$$rKL = \int Q(x) \log \frac{Q(x)}{P(x)} dx$$



KL будет меньше, для левой картинки, так как $Q(X)$ там покрывает все места, где $P(x) \neq 0$. rKL будет меньше, для правой картинки, так как $Q(X)$ лучше приближает $P(X)$ в местах, где $Q(x)! = 0$.

Неравенство Йенсена

Теорема

Для вероятностной плотности распределения $p(x)$ и выпуклой функции f выполнено:

$$\int p(x) f[q(x)] dx \geq f\left[\int p(x) q(x) dx\right]$$

Неравенство Пинскера

Теорема

Если случайные величины X и Y имеют плотности $p_\xi(x)$ и $p_\eta(x)$, где $x \in \mathbb{R}^n$, то

$$KL(p_\xi, p_\eta) \geq 2 \left[D(p_\xi, p_\eta)\right]^2 .$$

Доказатальство неравенства Пинскера

Доказательство.

Пусть $A_o = \{x : p_\xi(x) \geq p_\eta(x)\}$ и $A_o^c = \{x : p_\xi(x) < p_\eta(x)\}$.
Тогда по теореме Шеффе:

$$D(p_\xi, p_\eta) = \frac{1}{2}\int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)| =$$
$$= \frac{1}{2}\int_{A_o} |p_\xi(x) - p_\eta(x)| + \frac{1}{2}\int_{A_o^c} |p_\xi(x) - p_\eta(x)| =$$
$$= P_\xi(A_o) - P_\eta(A_o),$$

где $P_\xi(A_o) = \int_{A_o} p_\xi(x)dx$, а $P_\eta(A_o) = \int_{A_o} p_\eta(x)dx$. $\square$

Доказательство неравенства Пинскера

Доказательство.

$$
\begin{aligned}
KL(p_\xi, p_\eta) = \int_{A_o} p_\xi(x)\log\frac{p_\xi(x)}{p_\eta(x)}dx\, + \\
+ \int_{A_o^c} p_\xi(x)\log\frac{p_\xi(x)}{p_\eta(x)}dx = \\
= -P_\xi(A_o)\int_{A_o}\frac{p_\xi(x)}{P_\xi(A_o)}\log\frac{p_\eta(x)}{p_\xi(x)}dx\, - \\
-P_\xi(A_o^c)\int_{A_o^c}\frac{p_\xi(x)}{P_\xi(A_o^c)}\log\frac{p_\eta(x)}{p_\xi(x)}dx \geq
\end{aligned}
$$

Учитывая неравенство Йенсена для $f(y) = \log(y)$ $\square$

Доказатальство неравенства Пинскера

Доказательство.

$$
\begin{aligned}
&\geq -P_\xi(A_o)\log\left(\frac{1}{P_\xi(A_o)}\int_{A_o} p_\eta(x)dx\right) \\
&-P_\xi(A_o^c)\log\left(\frac{1}{P_\xi(A_o^c)}\int_{A_o^c} p_\eta(x)dx\right) = \\
&= -P_\xi(A_o)\log\left(\frac{P_\eta(A_o)}{P_\xi(A_o)}\right) - P_\xi(A_o^c)\log\left(\frac{P_\eta(A_o^c)}{P_\xi(A_o^c)}\right) = \\
&= -P_\xi(A_o)\log\left(\frac{P_\eta(A_o)}{P_\xi(A_o)}\right) - (1 - P_\xi(A_o))\log\left(\frac{1-P_\eta(A_o)}{1-P_\xi(A_o)}\right)
\end{aligned}
$$

$\square$

Доказатальство неравенства Пинскера

Доказательство.

Возьмём $p = P_\xi(A_o)$, $q = P_\eta(A_o)$, тогда

$$D(p_\xi, p_\eta) = p - q,$$
$$KL(p_\xi, p_\eta) \geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

$\square$

Доказатальство неравенства Пинскера

Доказательство.

Рассмотрим:

$$f(q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - \lambda(p-q)^2,$$

тогда

$$\frac{\partial f}{\partial q} = (q-p)\left[\frac{1}{q(1-q)} - 2\lambda\right]$$

. При $\lambda < 2$ производная равна нулю только в точке $p = q$. При этом $f(0) = +\infty$ и $f(1) = +\infty$. То есть $f(q) \geq f(p) = 0$.

Таким образом, утверждение доказано. $\square$

КЛ-расстояние для нескольких случайных величин

Теорема

Пусть $X^n = X_1, ..X_n$ и $Y^n = Y_1, ..Y_n$ — случайные векторы, $X_1, ..X_n \sim p_X$, $Y_1, ..., Y_n \sim p_Y$. Тогда

$$KL(p_{X^n}, p_{Y^n}) = nKL(p_X, p_Y)$$

Свойства КЛ-расстояния

Доказательство.

Так как случайные величины iid:

$$KL(p_{X^n}, p_{Y^n}) =$$

$$= \int_{\mathbb{R}^n} p_{X^n}(y_1, \dots, y_n) \log \frac{p_{X^n}(y_1,\dots,y_n)}{p_{Y^n}(y_1,\dots,y_n)} dy_1 \dots dy_n =$$

$$= \int_{\mathbb{R}^n} p_{X^n}(y_1, \dots, y_n) \sum_i \log \frac{p_X(y_i)}{p_Y(y_i)} dy_1 \dots dy_n =$$

$$= \sum_i \int_{\mathbb{R}^n} p_{X^n}(y_1, \dots, y_n) \log \frac{p_X(y_i)}{p_Y(y_i)} dy_1 \dots dy_n =$$

$$= n \int_{\mathbb{R}} p_X(x) \frac{p_X(x)}{p_Y(x)} dx = nKL(p_X, p_Y).$$

$\square$

Свойства КЛ расстояний

КЛ расстояние:

- несимметрично;
- инвариантно относительно преобразований переменных;
- аддитивно для независимых переменных: если $P(X, Y)$ и $Q(X, Y)$ можно факторизовать, то выполняется: $KL(P\|Q) = KL(P_1\|Q_1) + KL(P_2\|Q_2)$.
- Для случайных векторов с iid компонентами выполнено $KL(p_{X^n}, p_{Y^n}) = nKL(p_X, p_Y)$.

Связь КЛ с ОМП

Максимизация $\ell_n(\theta)$ эквивалентна максимизации $M_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i;\theta)}{f(X_i;\theta_*)}$, поскольку $M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta_*))$ и $\ell_n(\theta_*)$ — константа.

Тогда

$$\mathbb{E}_{\theta_*}\left(\log\frac{f(x;\theta)}{f(x;\theta_*)}\right) = \int \log\left(\frac{f(x;\theta)}{f(x;\theta_*)}\right) f(x;\theta_*)dx =$$

$$= -\int \log\left(\frac{f(x;\theta_*)}{f(x;\theta)}\right) f(x;\theta_*)dx = -KL(f_{\theta_*}, f_\theta).$$

Связь КЛ с ОМП

Таким образом, $M_n(\theta) \approx -KL(f_{\theta_*}, f_\theta)$ принимает максимальное значение в точке θ_*, поскольку $-KL(f_{\theta_*}, f_{\theta_*}) = 0$ и $-KL(f_{\theta_*}, f\theta) < 0$ при $\theta \neq \theta_*$.

Теорема

Пусть θ_* — реальное значение параметра θ. Обозначим через

$$M_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i;\theta)}{f(X_i;\theta_*)}$$

и $M(\theta) = -KL(\theta_*, \theta)$.

Допустим, что $\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \to 0$ и для каждого $\epsilon > 0$ $\sup_{\theta:|\theta-\theta_*|\geq\epsilon} M(\theta) < M(\theta_*)$.

Пусть $\widehat{\theta}_n$ обозначает ОМП, тогда $\widehat{\theta}_n \to \theta_*$.

Доказательство.

Так как $\widehat{\theta}_n$ максимизирует $M_n(\theta)$, то $M_n(\widehat{\theta}_n) \geq M_n(\theta_*)$. Следовательно,

$$
\begin{aligned}
M(\theta_*) - M(\widehat{\theta}_n) = M_n(\theta_*) - M(\widehat{\theta}_n) + M(\theta_*) - M_n(\theta_*) \leq \\
\leq M(\widehat{\theta}_n) - M(\widehat{\theta}_n) + M(\theta_*) - M_n(\theta_*) \leq \\
\leq \sup_\theta |M_n(\theta) - M(\theta)| + M(\theta_*) - M_n(\theta_*) \to 0.
\end{aligned}
$$

Таким образом, для любого $\delta > 0$:

$$\Big(M(\widehat{\theta}_n) < M(\theta_*) - \delta\Big) \to 0.$$

Возьмем произвольное $\epsilon > 0$. Согласно условию теоремы найдется $\delta > 0$, для которого из неравенства $|\theta - \theta_*| \geq \epsilon$ следует, что $M(\theta) < M(\theta_*) - \delta$.

Значит, $\Big(|\widehat{\theta}_n - \theta_*| > \epsilon\Big) \leq \Big(M(\widehat{\theta}_n) < M(\theta_*) - \delta\Big) \to 0.$ $\square$

χ^2 расстояние

Определение

Определение

Расстоянием χ^2 между плотностями вероятностей f_X и g_Y называют

$$\chi^2(p_X, g_Y) = \int_{\mathbb{R}^n} \frac{(p_X(z) - g_Y(z))^2}{p_X(z)} dz.$$

Связь с КЛ

Предположим, что $p_X \approx p_Y$.

$$KL(p_X, p_Y) = -\int_{\mathbb{R}^n} p_X(z) \log\left(1 + \frac{p_X(z)}{p_Y(z)} - 1\right) dz \approx$$

$$\approx -\int_{\mathbb{R}^n} p_X(z)\left(\frac{p_X(z)}{p_Y(z)} - 1\right) dz + \int_{\mathbb{R}^n} p_X(z)\left(\frac{p_X(z)}{p_Y(z)} - 1\right)^2 dz =$$

$$= \int_{\mathbb{R}^n} \frac{(p_X(z) - {}_Y(z))^2}{p_X(z)} dz.$$

То есть χ^2 дистанция является первым членом разложения КЛ дистанции в ряд Тейлора.

Полезные неравенства

Можно заметить, что для распределений P и Q выполнены неравенства:

$$
\begin{aligned}
(D(P,Q))^2 &\le \chi^2(P,Q);\\
(D(P,Q))^2 &\le 2KL(P,Q);\\
KL(P,Q) &\le \chi^2(P,Q).
\end{aligned}
$$

NB: неравенства для KL выполняются в случае использования

натурального логарифма в определении.

Расстояние Йенсена-Шеннона

Определение

Определение

Расстоянием Йенсена-Шеннона между плотностями вероятностей f_X и g_y называют

$$JS(f_X, g_Y) = \frac{1}{2}\left(KL(f_X, \frac{1}{2}(f_X + g_Y)) + KL(g_Y, \frac{1}{2}(f_X + g_Y))\right),$$

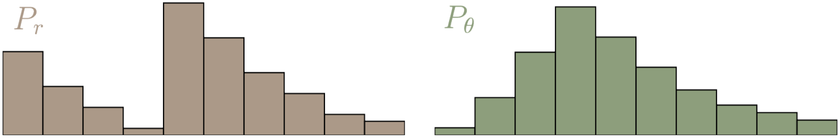
Свойства

- симметричность;
- ограниченность $0 \leq JS(P,Q) \leq \ln(2)$;
- $\sqrt{JS(.,.)}$ является метрикой.

Расстояние Васерштейна

Мотивация

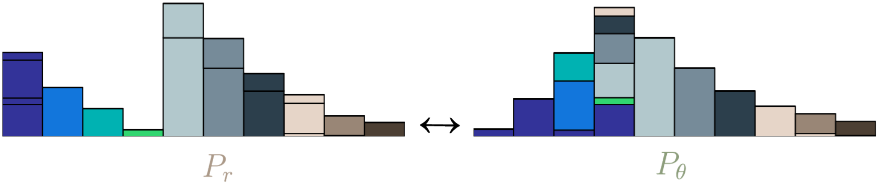
Предположим, что мы хотим переместить части распределения таким образом, чтобы из P_r получится P_θ



При этом мы хотим сэкономить усилия, то есть не перемещать большие куски на большие расстояния.

Picture credit:https://vincentherrmann.github.io/blog/wasserstein/

Мотивация: Earth Mover's Distance



$$EMD(P_r, P_\theta) = \inf_{\gamma \in \Pi} \sum_{x,y} ||x - y|| \gamma(x, y) = \inf_{\gamma \in \Pi} \mathbb{E}_{(x,y) \sim \gamma} ||x - y||,$$

где $\gamma(x, y)$ - усилия по перемещению из x в y, Π — набор всех перемещений между P_r и P_θ, $\gamma \in \Pi$. Заметим, что $\sum_x \gamma(x, y) = P_\theta(y)$, $\sum_y \gamma(x, y) = P_r(x)$. Фактически, мы ищем оптимальный переход между P_r и P_θ.

Picture credit:https://vincentherrmann.github.io/blog/wasserstein/

Определение

Перепишем определение в более общих терминах для метрики $d(x, y)$ и непрерывных распределений.

Определение

Для плотностей p_X и p_Y

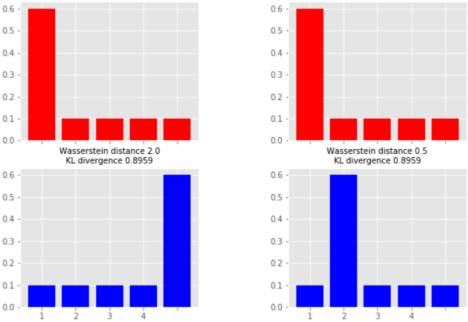
$$W(p_X, p_Y) = \inf_{\gamma \in \Pi(p_X, p_Y)} \int_{M \times M} \mathrm{dist}(\mathrm{x}, \mathrm{y}) \mathrm{d}\gamma(\mathrm{x}, \mathrm{y}),$$

где dist — заданная метрика на пространстве $\mathcal{X}$, γ - метрика в пространстве пар p.

NB: можно определить p-й момент Васерштейна.

Васерштейн и КЛ

Основным отличием расстояния Васерштейна является то, что он учитывает также расстояние, на котором находятся отличия в распределе



Picture credit:https://goo.gl/ncx3gt

Двойственность Канторовича—Рубинштейна

Метрику, определённую выше довольно трудно подсчитать на практике, потому обычно используют более применимую форму записи для первого момента Васерштейна:

Теорема

$$W(p_r, p_\theta) = \sup_{f \in \mathrm{Lip}_1(\mathrm{X})} \left(\mathbb{E}_{x \sim p_r}[f(x)] - \mathbb{E}_{x \sim p_\theta}[f(x)] \right),$$

где супремум берётся по всем 1-липшецевым функциям f.