Метод синтеза устойчивой оценки функции плотности распределения вероятностей

А.Н. Назаров, Д.А. Карандеев

1 О методе стохастической регуляризации

Исследуются вопросы построения оценки функции плотности распределения вероятностей с помощью метода стохастической регуляризации. Этот метод был описан в [1]-[3]. Он широко применяется для решения стохастических некорректно поставленных задач [4]. Для оценивания плотности распределения вероятностей предлагается следовать методу стохастической регуляризации с использованием специального непрерывного полигона. Отдельный раздел статьи посвящен сравнению новой оценки с несколькими наиболее часто употребляемыми на практике парзеновскими оценками. Там же производится сравнительный анализ двух вариантов использования метода стохастической регуляризации (с использованием различных полигонов) для оценивания плотности распределения вероятностей по эмпирическим данным.

Известно, что плотность распределения вероятностей $p(x)$ - функция, удовлетворяющая следующему уравнению:

$$\int_{-\infty}^{z} p(x)dx = F(z), \tag{1}$$

где $F(z)$ - функция распределения. Это уравнение можно переписать с помощью функции Хевисайда $\theta(t)$:

$$\int_{-\infty}^{+\infty} \theta(z-x)p(x)dx = F(z), \tag{2}$$

где $\theta(t) = \begin{cases} 0, & t < 0, \\ 1, & t \geq 0. \end{cases}$ Таким образом, функция $p(x)$ является решением интегрального уравнения Фредгольма первого рода. Задача решения такого уравнения является некорректно поставленной, так как его решение не является устойчивым по отношению к малым изменениям правой части

[5]. Для решения таких задач широко применяется метод регуляризации по Тихонову.

Предположим, что у нас имеется случайная выборка $x_1, x_2, \ldots, x_n$ из неизвестного распределения. Будем искать оценку плотности распределения в виде решения операторного уравнения:

$$Af = F, \tag{3}$$

где A оператор, осуществляющий взаимно-однозначное отображение элементов $f(x)$ множества Φ_1 метрического пространства E_1 в элементы $F(x)$ множества Φ_2 метрического пространства E_2 в ситуации, когда вместо правой части $F(x)$ задана последовательность случайных функций $F_n(x),\ n = 1, 2, \ldots,$ такая, что:

$$\rho_{E_2}(F, F_n) \xrightarrow{\mathrm{P}} 0, \text{ при } n \to \infty.$$

Будем решать уравнение (3) методом регуляризации по Тихонову. Суть метода состоит в том, что по последовательности $F_n(x)$ строится последовательность функций $f_n(x)$, минимизирующая функционал

$$R(f, F_n) = \rho^2_{E_2}(Af, F_n) + \alpha_n \Omega(f), \tag{4}$$

где $\Omega(f)$ - стабилизирующий функционал, а константы регуляризации $\alpha_n \to 0$ при $n \to 0$.

Для стабилизирующего функционала $\Omega(f)$, удовлетворяющего следующим трем условиям:

1) точное решение f_0 уравнения (3) принадлежит $D(\Omega(f))$ - области определения стабилизирующего функционала $\Omega(f)$,
2) функционал $\Omega(f)$ принимает на $D(\Omega(f))$ только вещественные неотрицательные значения,
3) все множества $M_C = \{f : \Omega(f) \le C\},\ C \ge 0$ являются компактами в метрике $\rho_{E_2}(f_1, f_2)$,

доказаны следующие теоремы [6].

Теорема 1 *Если для каждого n выбирается положительное α_n, такое, что $\alpha_n \to 0$ при $n \to 0$, то для любых положительных μ и ν найдется такой номер $N = N(\mu, \nu)$, что при всех $n > N$ элементы f_n, минимизирующие функционал (4), удовлетворяют неравенству:*

$$\mathrm{P}\{\rho_{E_1}(f_n, f_0) > \nu\} \le \mathrm{P}\{\rho^2_{E_2}(F, F_n) > \mu\alpha_n\},$$

где $f_0(x)$ - точное решение операторного уравнения (3) с правой частью F.

Теорема 2 *Пусть E_1 - гильбертово пространство, $\Omega(f) = ||f||^2$ и выполнены остальные условия теоремы 1. Тогда для любого $\varepsilon > 0$ найдется такой номер $N = N(\varepsilon)$, что при $n > N(\varepsilon)$:*

$$\mathrm{P}\left\{\|f_n - f_0\|^2 > \varepsilon\right\} < 2\,\mathrm{P}\left\{\rho^2_{E_2}(F_n, F) > \frac{\varepsilon}{2}\alpha_n\right\}.$$

Итак, эти две теоремы при правильном выборе параметров настройки позволяют решить задачу (2) методом регуляризации по Тихонову [3]. Проблема выбора параметров настройки алгоритма - это отдельная задача, которая рассматривается в [7].

Подобная задача рассматривалась для оценки плотности распределения вероятностей из класса $L_2(-\pi, \pi)$, из класса функций, k-я производная которых интегрируема с квадратом на (a, b), и других классов. В случае, когда плотность распределения вероятностей $f(x) \in L_2(-\infty, \infty)$, была получена оценка Розенблатта-Парзена, которая строилась как решение уравнения (2) с эмпирической функцией распределения вероятностей в правой части [4]. Мы рассматриваем задачу построения оценки $f(x) \in L_2(-\infty, \infty)$, когда правая часть уравнения (2) - это некоторая непрерывная оценка функции распределения вероятностей, построенная по выборке конечного объема.

2 Процедура синтеза оценки плотности распределения вероятностей в функциональном пространстве $L_2(-\infty, \infty)$

Пусть искомая плотность распределения вероятностей $p(x) \in L_2(-\infty, \infty)$. Будем искать $p(x)$ как решение уравнения (1). В качестве стабилизирующего функционала возьмем:

$$\Omega(f)\left\|\int_{-\infty}^{+\infty} g(x-t)f(t)dt\right\|^2_{L_2}, \qquad (5)$$

где $g(x-t)$ - интегрируемая на всей числовой оси функция.

Пусть

$$F_n(x) = \int_{-\infty}^{x} f_n(t)dt.$$

Согласно методу регуляризации, решение (1) может быть найдено путем минимизации в L_2 функционала (4), который в нашем случае

имеет вид:

$$R_{n,\alpha} = \left\| \int\limits_{-\infty}^{x} f(t)dt - \int\limits_{-\infty}^{x} f_n(t)dt \right\|_{L_2}^2 + \alpha_n \left\| \int\limits_{-\infty}^{+\infty} g(x-t)f(t)dt \right\|_{L_2}^2 \qquad (6)$$

Теорема 3 *Функция, на которой функционал (6) достигает минимума, представима в виде:*

$$f(x) = \int\limits_{-\infty}^{+\infty} K_{\alpha_n}(x-t)f_n(t)dt, \qquad (7)$$

где $K_{\alpha_n}(x-t)$ - обратное преобразование Фурье функции:

$$\frac{1}{1+\alpha_n u^2 \hat{g}(u)\hat{g}(-u)},$$

а $\hat{g}(u)$ - преобразование Фурье функции $g(t)$.

Доказательство. В силу равенства Парсеваля скалярное произведение функций в пространстве L_2 равно с точностью до множителя $\frac{1}{2\pi}$ скалярному произведению их преобразований Фурье. Поэтому функционал (6) можно записать, переходя к преобразованиям Фурье, в следующем виде:

$$\hat{R}_{n,\alpha}(f) = \left\| \Phi\left(\int\limits_{-\infty}^{x} f(t)dt \right) - \Phi\left(\int\limits_{-\infty}^{x} f_n(t)dt \right) \right\|_{L_2}^2 + \alpha_n \left\| \Phi\left(\int\limits_{-\infty}^{+\infty} g(x-t)f(t)dt \right) \right\|_{L_2}^2$$

Учитывая свойства преобразования Фурье свертки и тот факт, что $\Phi\left(\int\limits_{-\infty}^{x} f(t)dt \right) = \frac{1}{iu}\hat{f}(u) + \pi\delta(u)$, где u - аргумент, рассматриваемый функционал принимает вид:

$$\hat{R}_{n,\alpha}(f) \left\| \frac{1}{iu}\left(\hat{f}(u) - \hat{f}_n(u) \right) \right\|_{L_2}^2 + \alpha_n \left\| \hat{g}(u)\hat{f}(u) \right\|_{L_2}^2.$$

Для отыскания минимума такого функционала, найдем его производную Фреше по функции $\hat{f}(x)$:

$$\hat{R}_{n,\alpha}(\hat{f}+\hat{h}) - \hat{R}_{n,\alpha}(\hat{f}) \left\| \tfrac{1}{iu}\left(\hat{f}(u) - \hat{f}_n(u) \right) + \tfrac{1}{iu}\hat{h}(u) \right\|_{L_2}^2 +$$
$$+ \alpha_n \left\| \hat{g}(u)\hat{f}(u) + \hat{g}(u)\hat{h}(u) \right\|_{L_2}^2 - \left\| \tfrac{1}{iu}\left(\hat{f}(u) - \hat{f}_n(u) \right) \right\|_{L_2}^2 - \alpha_n \left\| \hat{g}(u)\hat{f}(u) \right\|_{L_2}^2 =$$

$$= \left\| \tfrac{1}{iu} \left(\hat{f}(u) - \hat{f}_n(u) \right) \right\|_{L_2}^2 + \left\| \tfrac{1}{iu} \hat{h}(u) \right\|_{L_2}^2 + 2 \left(\tfrac{1}{iu} \left(\hat{f}(u) - \hat{f}_n(u) \right), \tfrac{1}{iu} \hat{h}(u) \right) +$$
$$+ \alpha_n \left\| \hat{g}(u)\hat{f}(u) \right\|_{L_2}^2 + \alpha_n \left\| \hat{g}(u)\hat{f}(u) \right\|_{L_2}^2 + 2\alpha_n \left(\hat{g}(u)\hat{f}(u), \hat{g}(u)\hat{h}(u) \right) -$$
$$- \left\| \tfrac{1}{iu} \left(\hat{f}(u) - \hat{f}_n(u) \right) \right\|_{L_2}^2 - \alpha_n \left\| \hat{g}(u)\hat{f}(u) \right\|_{L_2}^2 =$$
$$= 2 \left(\tfrac{1}{u^2} \left(\hat{f}(u) - \hat{f}_n(u) \right), \hat{h}(u) \right) + 2\alpha_n \left(\hat{g}(u)\overline{\hat{g}(u)}\hat{f}(u), \hat{h}(u) \right) + o(\|h(x)\|) =$$
$$= 2 \left(\left[\tfrac{1}{u^2} \left(\hat{f}(u) - \hat{f}_n(u) \right) + \alpha_n \hat{g}(u)\hat{g}(-u)\hat{f}(u) \right], \hat{h}(u) \right) + o(\|h(x)\|).$$

Выражение в квадратных скобках и есть производная Фреше функционала $\hat{R}_{n,\alpha}$ по $\hat{f}(x)$, а следовательно, минимум достигается на решении уравнения:

$$\frac{1}{u^2} \left(\hat{f}(u) - \hat{f}_n(u) \right) + \alpha_n \hat{g}(u)\hat{g}(-u)\hat{f}(u) \equiv 0.$$

Решением последнего уравнения является функция:

$$\hat{f}(u) = \frac{\hat{f}_n(u)}{1 + u^2 \alpha_n \hat{g}(u)\hat{g}(-u)}.$$

Теперь, обозначив через $K_{\alpha_n}(x-t)$ обратное преобразование Фурье функции

$$\frac{1}{1 + \alpha_n u^2 \hat{g}(u)\hat{g}(-u)},$$

получим выражение нужного вида:

$$f(x) = \int_{-\infty}^{+\infty} \hat{f}(u) e^{iux} du \int_{-\infty}^{+\infty} K_{\alpha_n}(x-t) f_n(t) dt.$$

Теорема доказана.

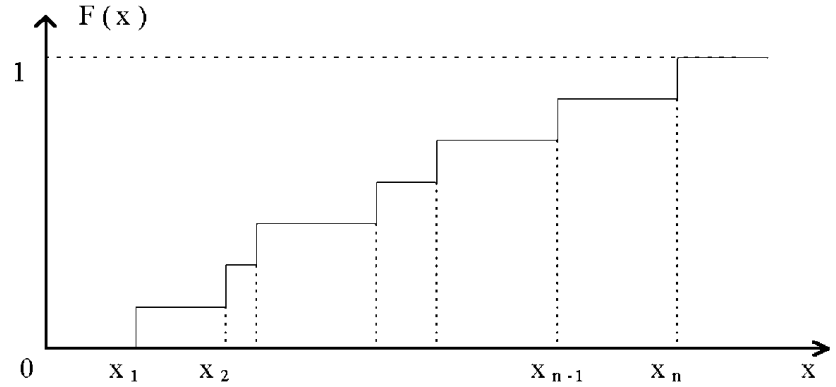
Согласно теореме 2 построенная оценка (7) будет сходится к искомой плотности распределения вероятностей в метрике $L_2(-\infty, +\infty)$.

3 Синтез оценки дискретной плотности распределения вероятностей на основе непрерывного полигона

Типичным способом нахождения оценки плотности распределения вероятностей является следующий метод [4].

Неизвестная функция распределения вероятностей $F(z)$ заменяется эмпирической функцией распределения вероятностей $F_n(z)$, найденной по заданной выборке $X = \{x_1, x_2, \ldots, x_n\}$. Далее с помощью метода

регуляризации решается некорректно поставленная задача для уравнения (2) с эмпирической функцией распределения вероятностей $F_n(z)$ в правой части. При таком способе нахождения функции плотности распределения вероятностей $p(x)$ в правой части уравнения (2) стоит разрывная функция $F_n(z)$, имеющая вид, представленный на рис. 1.



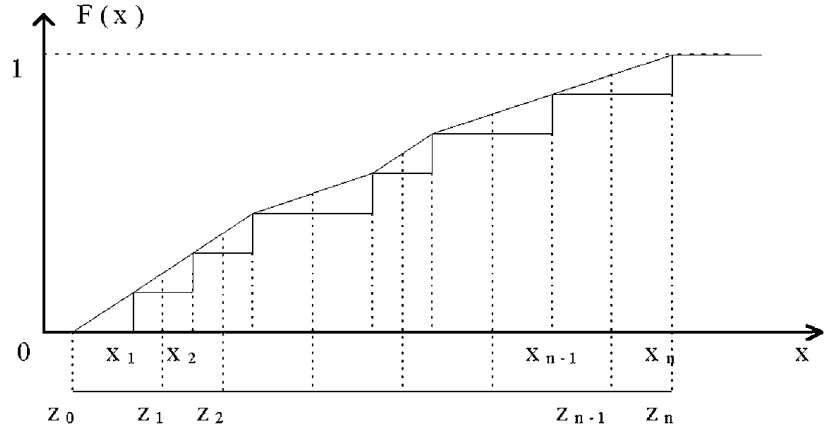
Случай разрывной функции $F_n(z)$, построенной по заданной выборке $X = \{x_1, x_2, \ldots, x_n\}$.

Если действовать таким способом, то в результате получается парзеновская оценка с экспоненциальным ядром, имеющая следующий вид [4]:

$$p(x) = \sum_{j=1}^{n} \frac{1}{2n\sqrt{\alpha}} e^{-\frac{|x-z_j|}{\sqrt{\alpha}}}. \qquad (8)$$

Заметим, что восстанавливая функцию распределения вероятностей $p(x)$ таким образом, при вычислении ее значения в каждой точке мы используем только информацию о том, сколько точек лежит левее заданной точки. Хотелось бы в функции-полигоне использовать еще и информацию о расстоянии между точками выборки. Этого можно достичь, если использовать вместо эмпирической функции распределения вероятностей $F_n(z)$ в правой части уравнения (2) некоторую непрерывную функцию.

В нашем случае будем строить специальный непрерывный полигон. Пусть имеется выборка $X = \{x_1, x_2, \ldots, x_n\}$. Рассмотрим изображенную на рис. 2 непрерывную функцию-полигон, которая является мажорантой разрывной функции $F_n(z)$, изображенной на рис. 1:



Вариант построения непрерывной функции $F(z)$, построенной по заданной выборке $X = \{x_1, x_2, \ldots, x_n\}$ и мажорирующей разрывную функцию $F_n(z)$.

1) по точкам выборки построим новую сетку $Z = \{z_0, z_1, \ldots, z_n\}$, где:

$$\begin{array}{l} z_0 = 2x_1 - x_2, \\ z_i = \frac{x_{i+1}+x_i}{2} + x_i, \ i = 1, \ldots, n-1, \\ z_n = x_n. \end{array}$$

2) значения функции в узлах сетки определим так:

$$\begin{array}{l} \tilde{F}_n(z_n) = \tilde{F}_n(z_{n-1}) + \frac{1}{n}, \\ \tilde{F}_n(z_0) = 0. \end{array}$$

3) по полученным точкам построим кусочно-линейную непрерывную функцию-полигон $\tilde{F}_n(x)$:

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{j=0}^{n-1} \left(j + \frac{x - z_j}{(z_{j+1} - z_j)} \right) [\theta(x - z_j) - \theta(x - z_{j+1})] + \theta(x - z_n). \quad (9)$$

Поскольку для рассматриваемого непрерывного полигона $\tilde{F}_n(x)$ и для эмпирической функции распределения $F_n(x)$, которая определяется формулой:

$$F_n(x) = \begin{cases} 0, & x < x_1, \\ \frac{k}{n}, & x_k \le x < x_{k+1}, \quad k = 1, 2, \ldots, n-1, \\ 1, & x \ge x_n. \end{cases}$$

справедливо неравенство:

$$\sup \left| F(x) - \tilde{F}_n(x) \right| \le \sup |F(x) - F_n(x)| + \frac{1}{n},$$

то справедлив следующий аналог теоремы Гливенко-Кантелли о сходимости эмпирической функции распределения вероятностей к истинной функции распределения вероятностей [8, 9].

Теорема 4 *Пусть $\tilde{F}_n(x)$ - функция, заданная формулой (9), $F(x)$ - функция распределения вероятностей случайной величины ξ. Тогда при $n \to \infty$ справедливо*

$$\mathrm{P}\left\{\sup_x \left|F(x) - \tilde{F}_n(x)\right| \underset{n\to\infty}{\longrightarrow} 0\right\} = 1$$

Пусть искомая плотность распределения вероятностей $f(x) \in L_2(-\infty, \infty)$. Будем искать $f(x)$ как решение уравнения (1) с функцией $\tilde{F}(x)$ в правой части.

Согласно методу регуляризации, решение (1) может быть найдено путем минимизации в L_2 функционала (4), который в нашем случае имеет вид:

$$R_{n,\alpha} = \left\| \int_{-\infty}^{x} f(t)dt - \tilde{F}_n(x) \right\|_{L_2}^2 + \alpha_n \|f(x)\|_{L_2}^2 \left\| Af(x) - \tilde{F}_n(x) \right\|_{L_2}^2 + \alpha_n \|f(x)\|_{L_2}^2,$$

где оператор A, согласно (2), задается следующей формулой:

$$Af(x) = \int_{-\infty}^{+\infty} \theta(x - z) f(z) dz.$$

Лемма 1 *Минимум функционала $R_{n,\alpha}$ достигается на решении уравнения:*

$$A^* A f(x) - A^* \tilde{F}_n(x) + \alpha_n f(x) \equiv 0, \tag{10}$$

где A - линейный оператор, а A^* - оператор, сопряженный к A.

Доказательство. Для отыскания минимума найдем производную Фреше от функционала

$$R_{n,\alpha} \left\| Af(x) - \tilde{F}_n(x) \right\|_{L_2}^2 + \alpha_n \|f(x)\|_{L_2}^2$$

по функции $f(x)$.

$$R(f+h) - R(f) \left\| Af(x) + Ah(x) - \tilde{F}_n(x) \right\|_{L_2}^2 + \alpha_n \|f(x) + h(x)\|_{L_2}^2 -$$
$$- \left\| Af(x) - \tilde{F}_n(x) \right\|_{L_2}^2 - \alpha_n \|f(x)\|_{L_2}^2 \left\| Af(x) - \tilde{F}_n(x) \right\|_{L_2}^2 +$$

$$+ \|Ah(x)\|^2_{L_2} + 2\left(Af(x) - \tilde{F}_n(x), Ah(x)\right) + \alpha_n\|f(x)\|^2_{L_2} +$$
$$+ \alpha_n\|h(x)\|^2_{L_2} + 2\alpha_n(f(x), h(x)) - \left\|Af(x) - \tilde{F}_n(x)\right\|^2_{L_2} - \alpha_n \|f(x)\|^2_{L_2} =$$
$$= 2\left(Af(x) - \tilde{F}_n(x), Ah(x)\right) + 2\alpha_n(f(x), h(x)) + o(\|h(x)\|) =$$
$$= 2\left(A^*\left(Af(x) - \tilde{F}_n(x)\right), h(x)\right) + 2\alpha_n(f(x), h(x)) + o(\|h(x)\|) =$$
$$= 2\left[\left\{A^*\left(Af(x) - \tilde{F}_n(x)\right) + \alpha_n f(x)\right\}, h(x)\right] + o(\|h(x)\|).$$

Выражение в фигурных скобках и есть производная Фреше функционала $R_{n,\alpha}$ по $f(x)$. Таким образом, минимум функционала $R_{n,\alpha}$ достигается на решении уравнения $A^*Af(x) - A^*\tilde{F}_n(x) + \alpha_n f(x) \equiv 0$, что и требовалось доказать.

Если в явном виде выписать оператор A, то уравнение (10) примет вид:

$$\int\limits_{-\infty}^{\infty} \theta(u - x)\left[\int\limits_{-\infty}^{\infty} \theta(u - \tau)f(\tau)d\tau - \tilde{F}_n(x)\right] du + \alpha_n f(x) = 0 \qquad (11)$$

Теорема 5 *Решением уравнения (11) является функция:*

$$p_n(x) = \frac{1}{2n}\left[\sum_{j=1}^{n-1} \operatorname{sign}(x - z_j)\lambda_j e^{-\frac{|x-z_j|}{\sqrt{\alpha_n}}} + \right.$$
$$+ \frac{1}{z_n - z_{n-1}} e^{-\frac{|x-z_n|}{\sqrt{\alpha_n}}} \operatorname{sign}(x - z_n) - \frac{1}{z_1 - z_0} e^{-\frac{|x-z_0|}{\sqrt{\alpha_n}}} \operatorname{sign}(x - z_0) +$$
$$\left. + 2\sum_{j=0}^{n-1} \frac{1}{(z_{j+1} - z_j)}(\theta(z_{j+1} - x) - \theta(z_j - x))\right],$$

где $\lambda_j = \frac{1}{(z_j - z_{j-1})} - \frac{1}{(z_{j+1} - z_j)}$.

Доказательство. Чтобы решить это уравнение, применим к уравнению (11) преобразование Фурье (в смысле обобщенных функций), учитывая, что преобразование Фурье свертки функций равно произведению преобразований Фурье этих функций, получаем уравнение:

$$\left(-\frac{1}{iu} + \pi\delta(u)\right)\left[\left(-\frac{1}{iu} + \pi\delta(u)\right)\hat{f}(u) - \Phi\left(\tilde{F}_n(x)\right)\right] + \alpha_n\hat{f}(u) = 0, \qquad (12)$$

где

$$\Phi\left(\tilde{F}_n(x)\right) \int\limits_{-\infty}^{\infty} \tilde{F}_n(x)e^{-ixu}du$$

- преобразование Фурье функции $\tilde{F}_n(x)$, $\hat{f}(u)$ - преобразование Фурье от неизвестной функции $f(x)$.

Преобразуем выражение (9) для $\tilde{F}_n(x)$ к следующему виду:

$$\tilde{F}_n(x)\frac{1}{n}\sum_{j=0}^{n-1}\left(j-\frac{z_j}{(z_{j+1}-z_j)}\right)[\theta(x-z_j)-\theta(x-z_{j+1})]+$$
$$+\frac{1}{n}\sum_{j=0}^{n-1}\frac{1}{(z_{j+1}-z_j)}x[\theta(x-z_j)-\theta(x-z_{j+1})]+\theta(x-z_n).$$

Пользуясь преобразованиями Фурье для функции Хевисайда [10]:

$$\Phi(\theta(x-a))=\pi\delta(u)-\frac{ie^{-iua}}{u},$$
$$\Phi(x\theta(x-a))=i\left[\pi\delta(1-u)+\frac{ie^{-iua}}{u^2}-\frac{ae^{-iua}}{u}\right],$$

где $\delta(u)$ - дельта функция Дирака, получаем преобразование Фурье рассматриваемой функции $\tilde{F}_n(x)$:

$$\Phi\left(\tilde{F}_n(x)\right)\frac{i}{nu}\sum_{j=0}^{n-1}\left(e^{-iuz_{j+1}}-e^{-iuz_j}\right)\left(j-\frac{i}{u(z_{j+1}-z_j)}\right)+$$
$$+\frac{i}{nu}\sum_{j=1}^{n-1}e^{-iuz_{j+1}}+\pi\delta(u)-\frac{i}{u}e^{-iu\alpha_n}.$$

Подставляем полученное преобразование Фурье в уравнение (12). Таким образом, (12) принимает вид:

$$\left(-\frac{1}{iu}+\pi\delta(u)\right)\left[-\frac{1}{iu}\hat{f}(u)+\pi\delta(u)\int\limits_{-\infty}^{\infty}f(x)e^{-iux}dx-\right.$$
$$-\frac{1}{n}\sum_{j=0}^{n-1}\frac{i}{u}\left(e^{-iuz_{j+1}}-e^{-iuz_j}\right)\left(j-\frac{i}{u(z_{j+1}-z_j)}\right)- \qquad (13)$$
$$\left.-\frac{1}{n}\sum_{j=1}^{n-1}\frac{i}{u}e^{-iuz_{j+1}}-\pi\delta(u)+\frac{i}{u}e^{-iuz_n}\right]+\alpha_n\hat{f}(u)=0$$

В силу свойств δ - функции Дирака и плотности распределения вероятностей:

$$\delta(u)f(u)=\delta(u)f(0),$$
$$\pi\delta(u)\hat{f}(u)=\pi\delta(u)\int\limits_{-\infty}^{+\infty}f(x)e^{-iux}dx=\pi\delta(u)\int\limits_{-\infty}^{+\infty}f(x)dx,$$

поскольку мы ищем решение этого уравнения в классе плотностей распределения вероятностей, последний интеграл равен единице, т.е. $\pi\delta(u)\hat{f}=\pi\delta(u)$.

Члены в квадратных скобках, не содержащие величины $\hat{f}(u)$, могут быть преобразованы к виду:

$$-\frac{1}{nu^2}\sum_{j=1}^{n-1}\frac{i}{u}\lambda_j e^{-iuz_j} - \frac{1}{nu^2}\left(\frac{1}{z_n - z_{n-1}}e^{-iuz_n} - \frac{1}{z_1 - z_0}e^{-iuz_0}\right).$$

После такого преобразования уравнение (13) примет вид:

$$\frac{1}{u^2}\hat{f}(u) + \pi\delta(u)\frac{1}{iu}\hat{f}(u) + \alpha_n\pi\hat{f}(u) -$$
$$-\left(\frac{1}{iu} + \pi\delta(u)\right)\frac{1}{nu^2}\left[\sum_{j=1}^{n-1}\lambda_j e^{-iuz_j} + \left(\frac{1}{z_n - z_{n-1}}e^{-iuz_n} - \frac{1}{z_1 - z_0}e^{-iuz_0}\right)\right] = 0$$

Это уравнение эквивалентно следующему:

$$\hat{f}(u) - i\pi\delta(u)u\hat{f}(u) + \alpha_n u^2\hat{f}(u) +$$
$$+\frac{1}{n}\left(\frac{1}{iu} + \pi\delta(u)\right)\left[\sum_{j=1}^{n-1}\lambda_j e^{-iuz_j} + \left(\frac{1}{z_n - z_{n-1}}e^{-iuz_n} - \frac{1}{z_1 - z_0}e^{-iuz_0}\right)\right] = 0.$$

В силу свойств функции Дирака:

$$i\pi\delta(u)u\hat{f}(u) = 0,$$
$$\pi\delta(u)\left[\sum_{j=1}^{n-1}\lambda_j e^{-iuz_j} + \left(\frac{1}{z_n - z_{n-1}}e^{-iuz_n} - \frac{1}{z_1 - z_0}e^{-iuz_0}\right)\right] = 0,$$

а оставшиеся члены дают следующее выражение для $\hat{f}(u)$:

$$\hat{f}(u) = -\frac{1}{inu(1+\alpha_n u^2)}\left[\sum_{j=1}^{n-1}\lambda_j e^{-iuz_j} + \left(\frac{1}{z_n - z_{n-1}}e^{-iuz_n} - \frac{1}{z_1 - z_0}e^{-iuz_0}\right)\right]. \qquad (14)$$

Применяя обратное преобразование Фурье к функции $\hat{f}(u)$ и учитывая, что:

$$\Phi_{inv}\left(\frac{e^{-iua}}{u(1+\alpha u^2)}\right) - \frac{i}{2}\left(-1 + 2\theta(a-x) + e^{\frac{1}{\sqrt{\alpha}}(-x+a)}\theta(x-a) - e^{\frac{1}{\sqrt{\alpha}}(-x+a)}\theta(a-x)\right),$$

получаем решение уравнения (11):

$$p_n(x) = \frac{1}{2n}\left[\sum_{j=1}^{n-1}\operatorname{sign}(x - z_j)\lambda_j e^{-\frac{|x-z_j|}{\sqrt{\alpha_n}}} + \right.$$
$$+\frac{1}{z_n - z_{n-1}}e^{-\frac{|x-z_n|}{\sqrt{\alpha_n}}}\operatorname{sign}(x - z_n) - \frac{1}{z_1 - z_0}e^{-\frac{|x-z_0|}{\sqrt{\alpha_n}}}\operatorname{sign}(x - z_0) + \qquad (15)$$
$$\left.+2\sum_{j=0}^{n-1}\frac{1}{(z_{j+1} - z_j)}(\theta(z_{j+1} - x) - \theta(z_j - x))\right],$$

что и требовалось доказать.

4 О сходимости полученной оценки

Оценка (15) может быть получена как оценка вида

$$f(x) = \int\limits_{-\infty}^{+\infty} K_{\alpha_n}(x-t) f_n(t) dt.$$

Если рассматривать $g(t) = \delta(t)$, тогда стабилизирующий функционал (5) примет вид:

$$\Omega(f) = \left\| \int\limits_{-\infty}^{+\infty} \delta(x-t) f(t) dt \right\|_{L_2}^2 \|f(x)\|_{L_2}^2.$$

Тогда $\hat{g}(x-t) = \Phi(\delta(x-t)) = e^{-itu}$, а следовательно, ядро $K_{\alpha_n}(t)$ в оценке (7) имеет вид:

$$K_{\alpha_n}(t) = \int\limits_{-\infty}^{+\infty} \frac{e^{iut}}{1+\alpha u^2} du = \frac{1}{2\sqrt{\alpha}} e^{-\frac{|t|}{\sqrt{\alpha}}}.$$

Если функция $f_n(x)$ дает представление полигона (9) в виде:

$$\tilde{F}_n(x) = \int\limits_{-\infty}^{x} f_n(t) dt,$$

то тогда оценка (15) может быть представлена в виде:

$$f(x) = \int\limits_{-\infty}^{+\infty} K_{\alpha_n}(x-t) f_n(t) dt \frac{1}{2\sqrt{\alpha}} \int\limits_{-\infty}^{+\infty} e^{-\frac{|x-t|}{\sqrt{\alpha}}} f_n(t) dt. \tag{16}$$

Для оценок такого вида из теоремы 2 следует сходимость к искомой плотности в метрике L_2.

В работе Надарая [11] было показано, что для оценки неизвестной плотности распределения вероятностей, представимой в виде:

$$f(x) = \frac{1}{h} \int\limits_{-\infty}^{+\infty} K\left(\frac{x-t}{h}\right) f_n(t) dt, \tag{17}$$

где $K(x)$ - некоторая плотность распределения вероятностей, $h \to 0$, при $n \to \infty$, справедлива следующая теорема:

Теорема 6 *Пусть $K(x)$ - функция с ограниченным изменением, плотность распределения вероятностей $f(x)$ равномерно непрерывна и ряд $\sum_{n=1}^{\infty} e^{-\alpha n h^2}$ сходится при любом $\alpha > 0$. Тогда при $n \to \infty$ с вероятностью единица*

$$V_n = \sup_{-\infty<x+\infty} |f_n(x) - f(x)| \to 0$$

В нашем случае для оценки (16) эта теорема дает сходимость в метрике $C_{[-\infty,\infty]}$. Таким образом, при выполнении соответствующих условий, рассматриваемые оценки сходятся в L_2 и $C_{[-\infty,\infty]}$.

5 Сравнение оценок функции плотности распределения вероятностей

Как уже отмечалось ранее, при использовании метода стохастической регуляризации, если в правой части уравнения (1) использовать эмпирическую функцию распределения вероятностей, изображенную на рис. 1, то в результате получается оценка (8), представляющая собой парзеновскую оценку с экспоненциальным ядром.

Сравнительный анализ оценки (8) и оценки (15) показал, что оценка (15), построенная на основе непрерывного полигона (9) для выборок из любых распределений в некотором смысле лучше, чем оценка (8). В частности, оценка (15) дает заметно лучшее приближение в случае, когда в восстанавливаемой плотности распределения вероятностей имеются "узкие"пики или "тяжелые хвосты". Экспериментальное сравнение проводилось на выборках из нормального распределения, гамма распределения и распределения Коши. Преимущество новой оценки проявляется в большей степени при малых объемах случайной выборки. Чем меньше объем выборки, тем большее преимущество дает новая оценка (15).

Многие авторы отмечают (см., например [12]), что даже самые незначительные изменения ширины колокола h могут "драматически"изменить парзеновскую оценку (17). В случае же предложенной в работе оценки (15), отклонения значения параметра регуляризации α_n от оптимального даже на порядок не приводят к таким драматическим изменениям, причем это справедливо в том числе и при очень малых объемах заданной выборки.

Интересно было бы сравнить оценку (15) с парзеновской оценкой с ядром Епанечникова, которая имеет вид:

$$p(x) = \begin{cases} \frac{1}{n\sqrt{\alpha}} \sum_{j=1}^{n} \frac{3}{4}\left(1 - (t_j)^2\right), & |t_j| \le 1 \\ 0, & |t_j| > 1 \end{cases}, \text{ где } t_j = \frac{x - z_j}{\sqrt{\alpha}}. \qquad (18)$$

При условии, что истинная плотность распределения вероятностей разлагается в ряд Тейлора в любой точке числовой оси [13], данная оценка является парзеновской оценкой с оптимальной в некотором смысле формой ядра $K\left(\frac{x-t}{h}\right)$. Сравнительный анализ двух оценок показал, что новая оценка действительно имеет преимущества. В частности, оценка (15) гораздо устойчивее к изменениям параметра α_n, оценка (15) дает более качественное приближение при малых объемах выборки. Улучшение особенно заметно, когда восстанавливаемая плотность распределения вероятностей не является унимодальной.

Так же интересно сравнить (15) с одной из наиболее часто применяемых парзеновских оценок - оценки с гауссовым ядром:

$$p(x)\frac{1}{\sqrt{2\pi}}\frac{1}{n\sqrt{\alpha}}\sum_{j=1}^{n}e^{-\frac{t_j^2}{2}},\ \text{где}\ t_j=\frac{x-z_j}{\sqrt{\alpha}}. \tag{19}$$

Оценка (15) демонстрирует те же самые преимущества относительно парзеновской оценки с гауссовым ядром, что и по отношению к двум другим рассматриваемым оценкам.

Детальный анализ всех четырех оценок показывает, что оценка (15) для выборок малого и очень малого объема дает заметное улучшение относительно существующих. Этими преимуществами являются устойчивость оценки относительно ее параметров, а также возможность более успешного ее применения к выборкам малого объема (порядка 20-40 элементов).

6 Заключение

Исследован метод стохастической регуляризации для случая, когда в качестве полигона берется специальным способом сконструированная функция. В результате получена новая оценка функции плотности распределения вероятностей по эмпирическим данным.

Произведен сравнительный анализ новой оценки и наиболее часто употребляемых парзеновских оценок (с гауссовым ядром, экспоненциальным ядром и ядром Епанечникова). Предложенная оценка функции плотности распределения вероятностей по эмпирическим данным дает заметное преимущество относительно классических парзеновских оценок. Этими преимуществами являются устойчивость оценки относительно ее параметров, а также возможность более успешного ее применения к выборкам малого объема (порядка 20-40 элементов). Чем меньше заданная выборка, тем более заметное улучшение дает новая оценка.

Исследована сходимость новой оценки. Практическая значимость полученных результатов заключается в применимости к решению комплексной проблемы

идентификации и управления битовым широкополосным пачечным трафиком в широкополосных цифровых сетях интегрального обслуживания.

Список литературы

[1] Page E.S. Continuous inspection schemes. - Biometrika, 1954, v. 41, N2, pp. 100-114.

[2] Стефанюк А.Р. Об оценивании отношения правдоподобия. - Статистические проблемы управления. Вып. 83, Вильнюс. ИМК АН ЛитССР, 1986.

[3] Stefanyuk A.R., Morgenstern W. "Analysis methods for population characteristics: heterogeneity detection"Proceedings of the international Conf. "Modelling of Noncommunicable diseases: Methodological Issues"(19-21 Sept, 1994, Heildelberg, Germany) Heildelberg, 1996, pp. 17-27.

[4] Vapnik V. Estimation of Dependences based on Empirical Data. Springer-Verlag. New York - Heidelberg - Berlin. 1982.

[5] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. Москва. Наука. 1986.

[6] Айду Ф.А., Вапник В.Н. Оценивание плотности вероятностей на основе метода стохастической регуляризации. Автоматика и Телемеханика. N4, 1988, стр. 84-97.

[7] Карандеев Д.А., Стефанюк А.Р. Выбор параметров настройки алгоритма при восстановлении функции плотности вероятности по эмпирическим данным. Автоматика и Телемеханика, N10, 1996, стр. 95-111.

[8] Рао С. Линейные статистические методы и их применения. Москва. Наука. 1968.

[9] Смирнов Н.В. Теория вероятностей и математическая статистика. Избранные труды. Москва. Наука. 1970.

[10] Крейн С.Г. Функциональный анализ. Москва, Наука. 1972.

[11] Надарая Э.А. О непараметрических оценках плотности вероятностей и регрессии. Теория вероятностей и ее применения, т. 10, вып. 1, стр. 199-203, 1965.

[12] Silverman B.W. Choosing the window width when estimating a density. Biometrika, N 1, pp. 1-11, 1978.

[13] Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности. Теория вероятностей и ее применение, т. 14, вып. 1, стр. 156-161, 1969.