Локальные меры сложности и быстрые скорости сходимости в теории статистического обучения

Толстихин Илья
ВЦ РАН

Апрель 2014

Локальный анализ

- Современный подход в теории статистического обучения
- Часто дает оценки оптимальных порядков
- Позволяет получать вычислимые по данным оценки

История:

- **Первые оценки «быстрого» порядка** $o(n^{-1/2})$
  (Vapnik and Chervonenkis, 1974)
- **Неравенство Талаграна**
  (Talagrand, 1996), (Bousquet, 2002)
- **Основы локального анализа**
  (Koltchinskii and Panchenko, 1999; Massart, 2000)
- **Локальные Радемахеровские сложности**
  (Bartlett et al., 2005; Koltchinskii, 2006)

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Определения и постановка задачи

- Множества **объектов** $\mathcal{X}$ и **ответов** $\mathcal{Y}$
  - Классификация: $\mathcal{Y} = \{0, 1\}$,
  - Регрессия: $\mathcal{Y} = \mathbb{R}$
- Неизвестное вероятностное распределение P на $\mathcal{X} \times \mathcal{Y}$
  - Частный случай: $\exists f \colon \mathcal{X} \to \mathcal{Y}$, такая что

$$P(Y = f(X)|X) = 1$$

    почти наверное для $X \sim P_X$.
- **Обучающая выборка** $S = \{(X_i, Y_i)\}_{i=1}^n$ — i.i.d. из P
- Ограниченная **функция потерь** $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$
  - Бинарная: $\ell(y', y'') = \mathbb{1}\{y \neq y'\}$,
  - Квадратичная: $\ell(y', y'') = (y' - y'')^2$.
- Класс отображений $\mathcal{H} = \{h \colon \mathcal{X} \to \mathcal{Y}\}$
  - Полиномы степени m для задач регрессии
  - Гиперплоскости для задач классификации

Определения и постановка задачи

- Множества **объектов** $\mathcal{X}$ и **ответов** $\mathcal{Y}$
- Неизвестное вероятностное распределение P на $\mathcal{X} \times \mathcal{Y}$
- **Обучающая выборка** $S = \{(X_i, Y_i)\}_{i=1}^n$ — i.i.d. из P
- Ограниченная **функция потерь** $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$
- Класс отображений $\mathcal{H} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$
- Потери при использовании $h \in \mathcal{H}$:

$$\ell_h(X, Y) = \ell\big(Y, h(X)\big), \quad (X, Y) \in \mathcal{X} \times \mathcal{Y}.$$

- **Средний риск** отображения $h \in \mathcal{H}$:

$$L(h) = \mathbb{E}_{(X,Y)\sim P}\left[\ell_h(X, Y)\right].$$

Задача: Минимизация среднего риска:

$$L(h) \to \min_{h \in \mathcal{H}}$$

Определения и постановка задачи

$$L(h) = \mathbb{E}_{(X,Y)\sim P}\left[\ell_h(X,Y)\right] = \mathbb{E}_{(X,Y)\sim P}\left[\ell\big(Y, h(X)\big)\right].$$

Байесовское отображение g^* и **Байесовский риск** L^*:

$$L^* = L(g^*) = \min_{g\colon \mathcal{X}\to\mathcal{Y}} L(g)$$

- Задача регрессии $\mathcal{Y} = \mathbb{R}$, $\ell(y', y'') = (y' - y'')^2$:

$$g^*(x) = \mathbb{E}[Y|X = x]$$

- Задача классификации $\mathcal{Y} = \{0, 1\}$, $\ell(y', y'') = \mathbb{1}\{y' \neq y''\}$:

$$g^*(x) = \mathbb{1}\left\{P(Y = 1|X = x) > 1/2\right\}.$$

Определения и постановка задачи

Задача: Минимизация среднего риска:

$$L(h) \to \min_{h \in \mathcal{H}} \qquad \text{(RM)}$$

Решение обозначим $h^* \in \mathcal{H}$. Очевидно, $L(h^*) \geq L^*$.

Проблема: Распределение P неизвестно! $\Rightarrow$ Не можем решить (RM)

- **Эмпирический риск** отображения $h \in \mathcal{H}$:

$$L_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell_h(X_i, Y_i).$$

- Если h не зависит от S, то $L_n(h) \to L(h)$ почти наверное (ЗБЧ).

Задача: Минимизация эмпирического риска:

$$L_n(h) \to \min_{h \in \mathcal{H}} \qquad \text{(ERM)}$$

Решение обозначим $\hat{h}_n \in \mathcal{H}$.

Определения и постановка задачи

Выбор множества $\mathcal{H}$

- Ошибки оценивания и аппроксимации:

$$L(\hat{h}_n) - L^* = \Big(L(\hat{h}_n) - \inf_{h \in \mathcal{H}} L(h)\Big) + \Big(\inf_{h \in \mathcal{H}} L(h) - L^*\Big).$$

- Чем «больше» $\mathcal{H}$, тем больше первая и меньше вторая.
  - Способ выбора $\mathcal{H}$ изучают задачи **выбора модели**.
  - Мы не будем их рассматривать. См. (Massart, 2007)

Насколько хорошо $\hat{h}_n$ (ERM) приближает оптимальную h^* (RM)?

- **Обобщающая способность**:

$$L(\hat{h}_n) - L_n(\hat{h}_n)$$

- **Избыточный риск**

$$\mathcal{E}(\hat{h}_n) = L(\hat{h}_n) - L(h^*) = L(\hat{h}_n) - \min_{h \in \mathcal{H}} L(h)$$

- Обе величины — случайные

Определения и постановка задачи

Задача: получение вероятностных оценок следующего вида:

- **Оценки обобщающей способности**:
  - Для всех $\epsilon \geq 0$:

$$\mathbb{P}\left\{L(\hat{h}_n) - L_n(\hat{h}_n) \geq \epsilon\right\} \leq B(\epsilon, n, \mathcal{H})$$

  - Для всех $\delta \geq 0$ с вероятностью не меньше $1 - \delta$:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq C(n, \mathcal{H}, \delta)$$

- **Оценки избыточного риска**: для $\epsilon \geq 0$
  - Для всех $\epsilon \geq 0$:

$$\mathbb{P}\left\{L(\hat{h}_n) - L(h^*) \geq \epsilon\right\} \leq B(\epsilon, n, \mathcal{H})$$

  - Для всех $\delta \geq 0$ с вероятностью не меньше $1 - \delta$:

$$L(\hat{h}_n) - L(h^*) \leq C(n, \mathcal{H}, \delta)$$

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Подход Вапника–Червоненкиса

- Равномерные по классу $\mathcal{H}$ отклонения:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq \sup_{h \in \mathcal{H}} |L(h) - L_n(h)|.$$

- Кроме того, для избыточного риска справедливо:

$$\begin{aligned}
\mathcal{E}(\hat{h}_n) &= \mathcal{E}(\hat{h}_n) + L_n(\hat{h}_n) - L_n(h^*) - \big(L_n(\hat{h}_n) - L_n(h^*)\big) \\
&\leq \mathcal{E}(\hat{h}_n) - \big(L_n(\hat{h}_n) - L_n(h^*)\big) \\
&= L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(h^*) - L(h^*) \\
&\leq 2 \sup_{h \in \mathcal{H}} |L(h) - L_n(h)|.
\end{aligned}$$

- Будем оценивать супремум эмпирического процесса:

$$\sup_{h \in \mathcal{H}} |L(h) - L_n(h)|.$$

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Типичные результаты VC-подхода

Для любого $\delta \geq 0$ с вероятностью не меньше $1 - \delta$ справедливо:

- Для конечного класса $\mathcal{H} = \{h_1, \ldots, h_N\}$:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \text{ и } \mathcal{E}(\hat{h}_n) = O\left(\sqrt{\frac{\log N + \log \frac{1}{\delta}}{n}}\right)$$

- Для классификации и $\mathcal{H}$ с конечной VC-размерностью h:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \text{ и } \mathcal{E}(\hat{h}_n) = O\left(\sqrt{\frac{h\log \frac{2ne}{h} + \log \frac{2}{\delta}}{n}}\right)$$

- Для произвольных классов $\mathcal{H}$ и функций потерь:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \text{ и } \mathcal{E}(\hat{h}_n) = O\left(2\mathbb{E}[R_n(\mathcal{H})] + 4\sqrt{\frac{2\log \frac{4}{\delta}}{n}}\right),$$

где $\mathbb{E}[R_n(\mathcal{H})]$ — Радемахеровская сложность класса $\mathcal{H}$.
Ее типичный порядок равен $O(n^{-1/2})$

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Идея «быстрой» скорости сходимости

- $O(n^{-1/2})$ — «медленная», а $o(n^{-1/2})$ — «быстрая» скорости
- Откуда берется быстрая? Вернемся к неравенству:

$$\begin{aligned}\mathcal{E}(\hat{h}_n) &\le L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(h^*) - L(h^*) \\ &\le 2 \sup_{h \in \mathcal{H}} |L(h) - L_n(h)|.\end{aligned} \tag{OLD}$$

- Но мы видели, что $\mathcal{E}(\hat{h}_n) \sim O(n^{-1/2})$.Значит, для $r \sim O(n^{-1/2})$:

$$\hat{h}_n, h^* \in \mathcal{H}(r) = \{h \in \mathcal{H} \colon L(h) - L(h^*) \le r\}.$$

- Возвращаясь к (OLD), получим:

$$\mathcal{E}(\hat{h}_n) \le 2 \sup_{h \in \mathcal{H}(r)} |L(h) - L_n(h)|. \tag{NEW}$$

- И так далее. . . (Можем рекурсивно повторять)
- Получаем скорости быстрее, если **дисперсии** функций в $\mathcal{H}(r)$ убывают с уменьшением r

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Неравенство Бернштейна

Теорема (Неравенство Бернштейна)

Пусть $\xi_1, \ldots, \xi_n$ — последовательность независимых случайных величин, таких что $\mathbb{E}[\xi_i] = 0$ и $\xi_i \leq 1$ (с вероятностью 1) для $i = 1, \ldots, n$. Положим

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}[\xi_i].$$

Тогда для всех $\delta \geq 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \leq \sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}.$$

Следствие Для каждого отображения $h \in \mathcal{H}$ и любого $\delta \geq 0$ с вероятностью не меньше $1 - \delta$:

$$L(h) - L_n(h) \leq \sqrt{\frac{2\mathrm{Var}[\ell_h(X, Y)] \log \frac{1}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}.$$

Н-во Бернштейна для конечного класса $\mathcal{H}$

Следствие Для каждого отображения $h \in \mathcal{H}$ и любого $\delta \geq 0$ с вероятностью не меньше $1 - \delta$:

$$L(h) - L_n(h) \leq \sqrt{\frac{2\mathrm{Var}[\ell_h(X,Y)] \log \frac{1}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}.$$

- Пусть $\mathcal{H} = \{h_1, \ldots, h_N\}$.
- Покажем, что для всех $\delta \geq 0$ с вероятностью не меньше $1 - \delta$:

$$\forall h \in \mathcal{H}: \quad L(h) - L_n(h) \leq \sqrt{\frac{2\mathrm{Var}[\ell_h(X,Y)] \log(N/\delta)}{n}} + \frac{2 \log(N/\delta)}{3n}$$

Доказательство:

$$\mathbb{P}\left\{\exists h \in \mathcal{H}\colon L(h) - L_n(h) \geq \sqrt{\frac{2\mathrm{Var}[\ell_h(X,Y)] \log(N/\delta)}{n}} + \frac{2 \log(N/\delta)}{3n}\right\}$$

$$= \mathbb{P}\left\{\bigcup_{j=1,\ldots,N} L(h_j) - L_n(h_j) \geq \sqrt{\frac{2\mathrm{Var}[\ell_{h_j}(X,Y)] \log(N/\delta)}{n}} + \frac{2 \log(N/\delta)}{3n}\right\}$$

$$\leq \sum_{j=1}^{N} \mathbb{P}\left\{L(h_j) - L_n(h_j) \geq \sqrt{\frac{2\mathrm{Var}[\ell_{h_j}(X,Y)] \log(N/\delta)}{n}} + \frac{2 \log(N/\delta)}{3n}\right\} \leq N \cdot \frac{\delta}{N} = \delta.$$

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Задачи без шума и оценки порядка $O(n^{-1})$

Теорема

Пусть $\mathcal{H} = \{h_1, \ldots, h_N\}$. *Тогда для всех* $\delta \geq 0$ *с вер-ю не меньше* $1 - \delta$:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq \sqrt{2L_n(\hat{h}_n)\frac{\log N + \log \frac{1}{\delta}}{n}} + 4\frac{\log N + \log \frac{1}{\delta}}{n} \qquad \text{(T.1)}$$

Также с вер-ю не меньше $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) \leq 3\sqrt{2L(h^*)\frac{\log N + \log \frac{2}{\delta}}{n}} + 7\frac{\log N + \log \frac{2}{\delta}}{n}. \qquad \text{(T.2)}$$

Пусть $g^* \in \mathcal{H}$ *и* $L(g^*) = 0$. *Тогда с вер-ю не меньше* $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) = L(\hat{h}_n) - L_n(\hat{h}_n) = L(\hat{h}_n) \leq 4\frac{\log N + \log \frac{1}{\delta}}{n}. \qquad \text{(T.3)}$$

Задачи без шума и оценки порядка $O(n^{-1})$

Доказательство:

- Из Следствия: Для всех $\delta \geq 0$ с вер-ю $\geq 1-\delta$:

$$\forall h \in \mathcal{H}: \quad L(h) - L_n(h) \leq \sqrt{\frac{2\mathrm{Var}[\ell_h(X,Y)]\log(N/\delta)}{n}} + \frac{2\log(N/\delta)}{3n}$$

- Заметим, что

$$\mathrm{Var}[\ell_h(X,Y)] \leq \mathbb{E}\left[\left(\ell_h(X,Y)\right)^2\right] \leq \mathbb{E}[\ell_h(X,Y)] = L(h).$$

- Получаем:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq \sqrt{\frac{2L(\hat{h}_n)\log(N/\delta)}{n}} + \frac{2\log(N/\delta)}{3n}$$

- Воспользовавшись $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, получаем:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq \sqrt{2L_n(\hat{h}_n)\frac{\log(N/\delta)}{n}} + 4\frac{\log(N/\delta)}{n}.$$

Задачи без шума и оценки порядка $O(n^{-1})$

Доказательство (часть 2):

- Для всех $\delta \geq 0$ с вер-ю $\geq 1 - \delta$:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq \sqrt{2L_n(\hat{h}_n)\frac{\log(N/\delta)}{n}} + 4\frac{\log(N/\delta)}{n}. \qquad (*)$$

- Но $g^* \in \mathcal{H}$ и $Y = g^*(X)$ п.н. $\Rightarrow L_n(\hat{h}_n) = 0$ п.н.
- Поскольку

$$\mathcal{E}(\hat{h}_n) = L(\hat{h}_n) - L(h^*) = L(\hat{h}_n) - L(g^*) = L(\hat{h}_n) = L(\hat{h}_n) - L_n(\hat{h}_n),$$

мы завершаем доказательство (T1) и (T3).

- Для оценки избыточного риска заметим, что:

$$L_n(\hat{h}_n) \leq L_n(h^*) = L_n(h^*) - L(h^*) + L(h^*). \qquad (**)$$

- С учетом (*), нер-ва Бернштейна и нер-ва Буля:

$$\mathcal{E}(\hat{h}_n) \leq \sqrt{2L_n(\hat{h}_n)\frac{\log\frac{2N}{\delta}}{n}} + 4\frac{\log\frac{2N}{\delta}}{n} + \sqrt{\frac{2L(h^*)\log\frac{2}{\delta}}{n}} + \frac{2\log\frac{2}{\delta}}{3n}.$$

- Вновь воспользовавшись (**) завершаем док-во T2.

Задачи без шума и оценки порядка $O(n^{-1})$

Обобщение для бесконечного класса $\mathcal{H}$:

Теорема ((Vapnik and Chervonenkis, 1974))

Рассмотрим задачу с бинарной функцией потерь и класс $\mathcal{H}$ с конечной VC-размерностью h. Тогда для всех $\delta \geq 0$ с вер-ю не меньше $1 - \delta$:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leq 2\sqrt{L_n(\hat{h}_n)\frac{2h\log(n+1) + \log\frac{4}{\delta}}{n}} + 4\frac{2h\log(n+1) + \log\frac{4}{\delta}}{n}.$$

Также для некоторой $C > 0$ с вер-ю не меньше $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) \leq C\left(\sqrt{L(h^*)\frac{h\log n + \log\frac{1}{\delta}}{n}} + \frac{h\log n + \log\frac{1}{\delta}}{n}\right).$$

Пусть $g^* \in \mathcal{H}$ и $L(g^*) = 0$. Тогда с вер-ю не меньше $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) \leq 4\frac{2h\log(n+1) + \log\frac{4}{\delta}}{n}.$$

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Соотношения между дисперсией и риском

$$\mathcal{E}(\hat{h}_n) \leq C\left(\sqrt{L(h^*)\frac{h\log n + \log(1/\delta)}{n}} + \frac{h\log n + \log(1/\delta)}{n}\right).$$

Проблемы прошлого подхода

- Слишком жесткие ограничения на P, особенно $L^* = 0$
- Не дают промежуточных порядков между $O(n^{-1/2})$ и $O(n^{-1})$

Как справиться с ними?

- Оказывается, $L(h^*)$ — не лучший множитель перед $n^{-1/2}$
- $L(h^*)$ не убывает к нулю при $n \to \infty$.
- Зато $L(\hat{h}_n) - L(h^*)$ обычно убывает.
- Как бы нам получить $L(\hat{h}_n) - L(h^*)$ перед $n^{-1/2}$?
- Рассмотрим **класс избыточных потерь** $\mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}$
- Для $f = \ell_h - \ell_{h^*}$ справедливо:

$$\mathbb{E}[f(X,Y)] = L(h) - L(h^*)$$

Соотношения между дисперсией и риском

- Рассмотрим **класс избыточных потерь** $\mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}$
- Для $f = \ell_h - \ell_{h^*}$ справедливо:

$$\mathbb{E}[f(X,Y)] = L(h) - L(h^*) = \mathcal{E}(h)$$

- Пусть $\mathcal{H} = \{h_1, \ldots, h_N\}$
- Повторим шаги прошлого доказательства (с н-вом Бернштейна)
- Возникнет проблема:
  - Для $f \in \mathcal{F}^*$ не выполнено $f(X,Y) \in [0,1]$
  - Возможно, $f(X,Y) \leq 0$ с вер-ю больше 0
  - $\Rightarrow$ не можем оценить сверху дисперсию мат. ожиданием
  - В прошлом док-ве ключевую роль играло соотношение

$$\mathrm{Var}[\ell_h(X,Y)] \leq \mathbb{E}[\ell_h(X,Y)] = L(h)$$

- Нас интересует соотношение для $f \in \mathcal{F}^*$ вида:

$$\mathrm{Var}\left[f(X,Y)\right] \leq \mathbb{E}\left[\left(f(X,Y)\right)^2\right] \leq c\big(\mathbb{E}[f(X,Y)]\big)^\alpha = c\big(L(h) - L(h^*)\big)^\alpha,$$

где $c > 0$ и $\alpha \in (0,1]$.

Задачи с шумом: конечный класс $\mathcal{H}$

$$\operatorname{Var}[f(X,Y)] \le \mathbb{E}\left[\big(f(X,Y)\big)^2\right] \le c\big(\mathbb{E}[f(X,Y)]\big)^{\alpha} = c\big(L(h) - L(h^*)\big)^{\alpha}, \quad (*)$$

где $f \in \mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}$, $c > 0$ и $\alpha \in (0, 1]$.

Оказывается, (*) достаточно для быстрых скоростей сходимости:

Теорема

Пусть $\mathcal{H} = \{h_1, \ldots, h_N\}$ и для всех $h \in \mathcal{H}$ выполнено (*). Тогда для всех $\delta \ge 0$ с вер-ю не меньше $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) = O\left(\left(\frac{\log N + \log \frac{1}{\delta}}{n}\right)^{\frac{1}{2-\alpha}}\right).$$

Задачи с шумом: конечный класс $\mathcal{H}$

$$\mathrm{Var}\,[f(X,Y)] \le \mathbb{E}\left[\big(f(X,Y)\big)^2\right] \le c\Big(\mathbb{E}[f(X,Y)]\Big)^{\alpha} = c\Big(L(h) - L(h^*)\Big)^{\alpha}, \qquad (*)$$

где $f \in \mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}$, $c > 0$ и $\alpha \in (0, 1]$.

Доказательство:

- Для $f \in \mathcal{F}^*$ обозначим $\hat{E}_n[f] = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$
  Так, для $\hat{f}_n = \ell_{\hat{h}_n} - \ell_{h^*}$ имеем: $\hat{E}_n[\hat{f}_n] = L_n(\hat{h}_n) - L_n(h^*) \le 0$
- Для всех $\delta \ge 0$ с вер-ю не меньше $1 - \delta$:

$$\forall f \in \mathcal{F}^*: \quad \mathbb{E}[f] - \hat{\mathbb{E}}_n[f] \le \sqrt{\frac{2\mathrm{Var}[f(X,Y)]\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}.$$

- С учетом (*) для $\hat{f}_n = \ell_{\hat{h}_n} - \ell_{h^*} \in \mathcal{F}^*$ получаем:

$$\mathcal{E}(\hat{h}_n) \le \sqrt{\frac{2\big(\mathcal{E}(\hat{h}_n)\big)^{\alpha}\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}.$$

- Завершаем док-во, «решая» неравенство относительно $\mathcal{E}(\hat{h}_n)$ ■

Задачи с шумом: конечный класс $\mathcal{H}$

$$\operatorname{Var}[f(X,Y)] \le \mathbb{E}\left[\big(f(X,Y)\big)^2\right] \le c\Big(\mathbb{E}[f(X,Y)]\Big)^{\alpha} = c\Big(L(h) - L(h^*)\Big)^{\alpha}, \qquad (*)$$

где $f \in \mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}$, $c > 0$ и $\alpha \in (0, 1]$.

Оказывается, (*) достаточно для быстрых скоростей сходимости:

Теорема

Пусть $\mathcal{H} = \{h_1, \ldots, h_N\}$ и для всех $h \in \mathcal{H}$ выполнено (*). Тогда для всех $\delta \ge 0$ с вер-ю не меньше $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) = O\left(\left(\frac{\log N + \log\frac{1}{\delta}}{n}\right)^{\frac{1}{2-\alpha}}\right).$$

- Оценка всегда лучше $O(n^{-1/2})$
- Если $\alpha = 1$, получаем $O(n^{-1})$
- Не требуем ни $g^* \in \mathcal{H}$, ни $L^* = 0$

Вопрос: когда выполняется (*)?

Классификация: условия малого шума

- Напомним, что Байесовское отображение имеет вид:

$$g^*(x) = \operatorname{sgn}\left\{\eta(x) - 1/2\right\} = \operatorname{sgn}\left\{P(Y = 1|X = x) - 1/2\right\}$$

  - $L(g^*) = 0 \Leftrightarrow \eta(x) \in \{0, 1\}$: получаем скорость $O(n^{-1})$
  - Никаких ограничений на $\eta(x)$: получаем скорость $O(n^{-1/2})$
  - А что в промежуточных случаях?

- **Условие Массара** (Massart and Nédélec, 2006):
  Существует $h \geq 0$, такая что с вер-ю 1 выполенно:

$$|2\eta(X) - 1| > h \qquad \text{(M)}$$

  Интерпретация: $\eta(x)$ не приближается к 0.5

- **Условие Маммена-Цыбакова** (Mammen and Tsybakov, 1999):
  Существуют $\alpha \in [0, 1]$ и $B > 0$, такие что для всех $t \geq 0$:

$$\mathbb{P}\{|2\eta(X) - 1| \leq t\} \leq Bt^{\frac{\alpha}{1-\alpha}}. \qquad \text{(MT)}$$

  Интерпретация: $\eta(x)$ не слишком часто приближается к 0.5

- (M) — частный случай условия (MT)

Классификация: условия малого шума

- Существует $h \geq 0$, такая что с вер-ю 1 выполенно:

$$|2\eta(X) - 1| > h \tag{M}$$

- Существуют $\alpha \in [0, 1]$ и $B > 0$, такие что для всех $t \geq 0$:

$$\mathbb{P}\{|2\eta(X) - 1| \leq t\} \leq Bt^{\frac{\alpha}{1-\alpha}}. \tag{MT}$$

Теорема ((Boucheron et al., 2005))

Рассмотрим бинарную функцию потерь. Пусть $g^* \in \mathcal{H}$.

- *Если выполнено условие (M), то для всех $f \in \mathcal{F}^*$:*

$$\operatorname{Var}[f(X,Y)] \leq \mathbb{E}\left[\left(f(X,Y)\right)^2\right] \leq \frac{1}{h}\mathbb{E}[f(X,Y)].$$

- *Если выполнено условие (MT), то $\exists c > 0$, такая что для всех $f \in \mathcal{F}^*$:*

$$\operatorname{Var}[f(X,Y)] \leq \mathbb{E}\left[\left(f(X,Y)\right)^2\right] \leq c\left(\mathbb{E}[f(X,Y)]\right)^{\alpha}.$$

Классификация: условия малого шума

- Существует $h \geq 0$, такая что с вер-ю 1 выполенно:

$$|2\eta(X) - 1| > h \tag{M}$$

- Существуют $\alpha \in [0, 1]$ и $B > 0$, такие что для всех $t \geq 0$:

$$\mathbb{P}\{|2\eta(X) - 1| \leq t\} \leq Bt^{\frac{\alpha}{1-\alpha}}. \tag{MT}$$

Получаем следствие:

Следствие Рассмотрим задачу с $\mathcal{Y} = \{-1, 1\}$ и бинарной функцией потерь. Пусть $\mathcal{H} = \{h_1, \ldots, h_N\}$ и $g^* \in \mathcal{H}$ (но не обязательно $L^* = 0$). Пусть, кроме того, выполнено условие (M). Тогда для всех $\delta \geq 0$ с вер-ю не меньше $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) = O\left(\frac{\log N + \log \frac{1}{\delta}}{n}\right).$$

Регрессия: выпуклый класс $\mathcal{H}$

Теорема (*Lee et al. (1998)*)

Рассмотрим задачу с $\mathcal{Y} \subseteq [0,1]$, $\ell\colon (y', y'') \to (y' - y'')^2$ *и* **выпуклым** *множеством* $\mathcal{H}$, *таким что* $h(x) \in [0,1]$ *для всех* $x \in \mathcal{X}, h \in \mathcal{H}$. *Тогда для всех* $f \in \mathcal{F}^*$ *справедливо:*

$$\operatorname{Var}\left[f(X,Y)\right] \leq \mathbb{E}\left[f(X,Y)\right)^2\Big] \leq 8\mathbb{E}[f(X,Y)].$$

Доказательство:

- Функция $\ell(y,\cdot)$ — 2-Липшицева на отрезке $[0,1]$:

  $$|(a-x)^2-(b-x)^2| = |(a-b)(a+b-2x)| \leq 2|a-b|,\ a,b,x \in [0,1]$$

- Поэтому для $f \in \mathcal{F}^*$:

$$\begin{aligned}
\operatorname{Var}[f(X,Y)] &\leq \mathbb{E}\left[\left(f(X,Y)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(Y-h(X)\right)^2 - \left(Y-h^*(X)\right)^2\right)^2\right] \\
&\leq 4\mathbb{E}\left[\left(h(X)-h^*(X)\right)^2\right].
\end{aligned}$$

Регрессия: выпуклый класс $\mathcal{H}$

Доказательство (часть 2):

- Итак,

$$\mathrm{Var}[f(X,Y)] \leq 4\mathbb{E}\left[\big(h(X)-h^*(X)\big)^2\right].$$

- Воспользуемся тождеством:

$$\frac{\big(Y-h(X)\big)^2+\big(Y-h^*(X)\big)^2}{2} = \left(\frac{h(X)+h^*(X)-2Y}{2}\right)^2 + \frac{1}{4}\big(h(X)-h^*(X)\big)^2.$$

- Взяв математические ожидания от обеих сторон, получаем:

$$\frac{L(h)+L(h^*)}{2} = L\left(\frac{h(X)+h^*(X)}{2}\right) + \mathbb{E}\left[\frac{1}{4}\big(h(X)-h^*(X)\big)^2\right].$$

- Поскольку $\mathcal{H}$ — выпуклое множество и $L(h^*) = \min_{h\in\mathcal{H}} L(h)$:

$$\frac{L(h)+L(h^*)}{2} \geq L(h^*) + \mathbb{E}\left[\frac{1}{4}\big(h(X)-h^*(X)\big)^2\right]$$

или

$$\mathbb{E}\left[\big(h(X)-h^*(X)\big)^2\right] \leq 2\big(L(h)-L(h^*)\big)\blacksquare$$

Результат обобщен на $\ell(y',y'') = (y'-y'')^p$, $p \geq 2$, (Mendelson, 2002).

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Неравенство Талаграна

Теорема ((Bousquet, 2002))

Пусть $X_1, \dots, X_n$ — i.i.d., принимающие значения в $\mathcal{X}$. Рассмотрим класс функций $\mathcal{G}$ из $\mathcal{X}$ в $\mathbb{R}$. Пусть $\mathbb{E}[f(X)] = 0$ и $f(X) \in [-c, c]$ для всех $f \in \mathcal{G}$. Обозначим $\sigma_{\mathcal{F}}^2 = \sup_{f \in \mathcal{G}} \operatorname{Var}[f(X_i)]$. Тогда для всех $\delta \geq 0$ с вер-ю не меньше $1 - \delta$:

$$Z \leq \mathbb{E}[Z] + \sqrt{2(n\sigma_{\mathcal{F}}^2 + 2c\mathbb{E}[Z]) \log \frac{1}{\delta}} + \frac{c \log \frac{1}{\delta}}{3},$$

где $Z = \sup_{f \in \mathcal{G}} \sum_{i=1}^n f(X_i)$.

- В случае $\mathcal{G} = \{f_0\}$ мы получаем в точности н-во Бернштейна:
$$\mathbb{E}[Z] = \sum_{i=1}^n \mathbb{E}[f_0(X_i)] = 0$$
- Н-во Талаграна — равномерное по классу функций обобщение
- Мы можем применить его к
$$Z' = \sup_{h \in \mathcal{H}} L(h) - L_n(h) = \sup_{h \in \mathcal{H}} \sum_{i=1}^n \frac{1}{n} \Big(\mathbb{E}[\ell_h(X, Y)] - \ell_h(X_i, Y_i) \Big)$$

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Локальный анализ

- Мы рассмотрим наиболее общую постановку:
  - Бесконечный класс $\mathcal{H}$
  - $L(g^*) \neq 0$
  - g^* не обязательно принадлежит $\mathcal{H}$
- Локальный подход был развит в работах
  (Koltchinskii and Panchenko, 1999)
  (Massart, 2000)
  (Bartlett et al., 2005)
  (Koltchinskii, 2006)

Определения:

- Отображение $\psi\colon [0,\infty) \to [0,\infty)$ — **подкоренное**, если оно (а)не убывает, (б)неотрицательно, (в)$r \to \psi(r)/\sqrt{r}$ не возрастает
  Свойство:
  Подкоренная функция имеет единственную неподвижную точку.
- $$\mathcal{H}(r) = \left\{ h \in \mathcal{H} \colon \mathbb{E}\left[\left(\ell_h(X,Y) - \ell_{h^*}(X,Y) \right)^2 \right] \leq r \right\}$$
  — функции из $\mathcal{H}$ с малыми дисперсиями избыточных потерь

Локальный анализ: главный результат

$$\mathcal{H}(r) = \left\{ h \in \mathcal{H} \colon \mathbb{E}\left[\left(\ell_h(X,Y) - \ell_{h^*}(X,Y) \right)^2 \right] \leq r \right\}$$

Теорема ((Bartlett et al., 2005))

Пусть существует константа $B > 0$, *такая что для всех* $h \in \mathcal{H}$:

$$\mathbb{E}\left[\left(\ell_h(X,Y) - \ell_{h^*}(X,Y) \right)^2 \right] \leq B\big(L(h) - L(h^*) \big).$$

Пусть, кроме того, существует подкоренное отображение $\psi_n(r)$, *такое что:*

$$B \cdot \mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} \big(L(h) - L_n(h) \big) - \big(L(h^*) - L_n(h^*) \big) \right] \leq \psi_n(r).$$

Пусть $r_n^* = \psi_n(r_n^*)$. *Тогда для всех* $\delta > 0$ *с вер-ю не меньше* $1 - \delta$:

$$\mathcal{E}(\hat{h}_n) \leq 151 \frac{r_n^*}{B} + \frac{(6 + 7B) \log \frac{1}{\delta}}{n}.$$

Локальный анализ: главный результат

Обсуждение результата

- Теорема фактически утверждает, что $\mathcal{E}(\hat{h}_n) \sim r_n^*$
- Модуль непрерывности эмпирического процесса в точке h^*

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} \big(L(h) - L_n(h)\big) - \big(L(h^*) - L_n(h^*)\big)\right].$$

  Его неподвижная точка r_n^* также играет важную роль в задачах M-оценок (van de Geer, 2000)

- Во многих интересных случаях $r_n^* \sim o(n^{-1/2})$
- Константы не оптимальны. Возможно, могут быть улучшены.

Локальный анализ: главный результат

Напомним:

$$\mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}, \quad \hat{\mathbb{E}}_n[f] = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$$

Основные шаги доказательства

1. Зафиксируем $r > 0$ и введем нормированный класс:

$$\mathcal{G}_r = \left\{ \frac{r}{\Delta(r, f)} f,\ f \in \mathcal{F}^* \right\},$$

   где нормировочный множитель $\Delta(r, f)$ выбран так, чтобы дисперсии функций множества $\mathcal{G}_r$ не превосходили r.
2. С помощью н-ва Талаграна для $V_r = \sup_{g \in \mathcal{G}_r}(\mathbb{E}[g] - \hat{\mathbb{E}}_n[g])$, всех $\delta \geq 0$ с вер-ю не меньше $1 - \delta$ получим следующую верхнюю оценку:

$$V_r \leq 2\mathbb{E}[V_r] + \sqrt{\frac{2r \log \frac{1}{\delta}}{n}} + \frac{8 \log \frac{1}{\delta}}{3n}.$$

3. С помощью *пилинга* (peeling) получим $\mathbb{E}[V_r] \leq 5\psi_n(r)/B$. Поскольку ψ_n — подкоренная, то для всех $r \geq r_n^*$ справедливо $\psi_n(r) \leq \sqrt{r r_n^*}$. Откуда мы получаем, что для $r \geq r_n^*$:

$$V_r \leq \sqrt{r}\left(\frac{10\sqrt{r_n^*}}{B} + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}\right) + \frac{8 \log \frac{1}{\delta}}{3n}.$$

Локальный анализ: главный результат

$$V_r = \sup_{g \in \mathcal{G}_r} (\mathbb{E}[g] - \hat{\mathbb{E}}_n[g])$$

Основные шаги доказательства (продолжение...)

3. Для $r \geq r_n^*$:

$$V_r \leq \sqrt{r} \left(\frac{10\sqrt{r_n^*}}{B} + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \right) + \frac{8 \log \frac{1}{\delta}}{3n}.$$

4. Выбрав определенным образом $r_0 > r_n^*$, для любой константы $K > 1$:

$$V_{r_0} \leq \frac{r_0}{KB}.$$

Воспользовавшись определением V_r, мы получаем, что для всех $\delta > 0$ с вер-ю не менее $1 - \delta$:

$$\forall f \in \mathcal{F}^*, \ \forall K > 1: \quad \mathbb{E}[f] - \hat{\mathbb{E}}_n[f] \leq \frac{\max\big(r_0, \mathbb{E}[f^2]\big)}{r_0} \frac{r_0}{KB} = \frac{\max\big(r_0, \mathbb{E}[f^2]\big)}{KB}.$$

5. Рассмотреть два случая: $\mathbb{E}[f^2] > r_0$ и $\mathbb{E}[f^2] \leq r_0$. С помощью соотношения на дисперсию получим:

$$\forall f \in \mathcal{F}^*: \quad \mathbb{E}[f] - \hat{\mathbb{E}}_n[f] \leq 151 \frac{r_n^*}{B} + \frac{(6 + 7B) \log \frac{1}{\delta}}{n}.$$

Наконец, заметив, что для $\hat{f}_n = \ell_{\hat{h}_n}(X) - \ell_{h^*}(X)$ выполнено $\hat{\mathbb{E}}_n \hat{f} \leq 0$, мы завершаем доказательство.

Локальный анализ: главный результат

Введем определения:

$$\mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}, \quad E_n f = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i), \quad Ef = \mathbb{E}\left[\left(f(X,Y)\right)^2\right]$$

Подробное доказательство: ШАГ 1

- Выберем $\lambda > 1$ и $r > 0$ и введем нормированную версию класса избыточных потерь $\mathcal{F}^*$:

$$\mathcal{G}_r = \left\{ \frac{r}{w(r,f)} f : f \in \mathcal{F}^* \right\},$$

  где $w(r,f) = \min\{r\lambda^k : k \in \mathbb{N}, r\lambda^k \geq Ef^2\}$.

- Рассмотрим два случая.
  - Если для $f \in \mathcal{F}^*$ выполнено $Ef^2 \leq r$, то $w(r,f) = r$; таким f соответствуют $g \in \mathcal{G}_r$, такие что $g = f$. Значит, $\mathrm{Var}[g] = \mathrm{Var}[f] \leq Ef^2 \leq r$.
  - Если для $f \in \mathcal{F}^*$ выполнено $Ef^2 > r$, то $w(r,f) = \lambda^k r$; таким f соответствуют $g \in \mathcal{G}_r$, такие что $g = f/\lambda^k$ и $Ef^2 \in (r\lambda^{k-1}, r\lambda^k]$. Значит, $\mathrm{Var}[g] = \mathrm{Var}[f]/\lambda^{2k} \leq r$.
- Таким образом, для всех $g \in \mathcal{G}_r$ справедливо $Var[g] \leq r$.

Локальный анализ: главный результат

Введем определения:

$$\mathcal{F}^* = \{\ell_h - \ell_{h^*} : h \in \mathcal{H}\}, \quad E_n f = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i), \quad Ef = \mathbb{E}\left[\left(f(X,Y)\right)^2\right]$$

Подробное доказательство: ШАГ 2

- Рассмотрим

$$V_r = \sup_{g \in \mathcal{G}_r} (Eg - E_n g)$$

Заметим, что $g(X,Y) \in [-1,1]$ для всех $g \in \mathcal{G}_r$. Также

$$\frac{1}{2}(Eg - E_n g) = \frac{1}{n}\sum_{i=1}^n \frac{Eg - g(X_i, Y_i)}{2},$$

$Eg - g(X_i, Y_i) \in [-1,1]$ и $\mathbb{E}\left[Eg - g(X_i, Y_i)\right] = 0$.

- Значит, мы можем применить н-во Талаграна:

$$V_r \le 2\mathbb{E}[V_r] + \sqrt{\frac{2r\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n}.$$

Локальный анализ: главный результат

Подробное доказательство: ШАГ 3

- Определим

$$\mathcal{F}^*(x,y) = \{f \in \mathcal{F}^* : x \le Ef^2 \le y\}.$$

Заметим, что $\mathrm{Var}[f(X,Y)] \le Ef^2 \le BEf \le B$ для всех $f \in \mathcal{F}^*$.

- Заметим, что для любых множеств A и B справедливо:

$$\mathbb{E}\left[\sup_{g\in A\cup B} Eg - E_n g\right] \le \mathbb{E}\left[\sup_{g\in A} Eg - E_n g\right] + \mathbb{E}\left[\sup_{g\in B} Eg - E_n g\right]$$

(выпуклость sup и н-во Йенсена)

- Пусть k — наименьшее целое, такое что $r\lambda^{k+1} \ge B$.

$$\mathbb{E}[V_r] = \mathbb{E}\left[\sup_{g\in\mathcal{G}_r} Eg - E_n g\right] \le \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - E_n f\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r,B)} \frac{r}{w(r,f)}(Ef - E_n f)\right]$$

$$\le \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - E_n f\right] + \sum_{j=0}^{k} \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r\lambda^j, r\lambda^{j+1})} \frac{r}{w(r,f)}(Ef - E_n f)\right]$$

$$\le \frac{\psi_n(r)}{B} + \frac{1}{B}\sum_{j=0}^{k} \lambda^{-j}\mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r\lambda^j, r\lambda^{j+1})} Ef - E_n f\right]$$

$$\le \frac{\psi_n(r)}{B} + \frac{1}{B}\sum_{j=0}^{k} \lambda^{-j}\psi_n(r\lambda^{j+1}) \le \frac{\psi_n(r)}{B}\left(1 + \sqrt{\lambda}\sum_{j=0}^{k}\lambda^{-j/2}\right) \le 5\frac{\psi_n(r)}{B},$$

поскольку $\psi_n(r)$ — подкоренное, для любого $\beta \ge 1$ имеем $\psi_n(\beta r) \le \sqrt{\beta}\psi_n(r)$.

- Наконец, для всех $r \ge r_n^*$ выполнено:

$$\mathbb{E}[V_r] \le \frac{5}{B}\psi_n(r) \le \frac{5}{B}\sqrt{r r_n^*}.$$

Локальный анализ: главный результат

Подробное доказательство: ШАГ 4

- Получили, что для всех $r \geq r_n^*$:

$$V_r \leq \sqrt{r}\left(\frac{10\sqrt{r_n^*}}{B} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + \frac{8\log\frac{1}{\delta}}{3n}.$$

- Выберем $r_0 \geq r_n^*$ так, чтобы

$$V_{r_0} \leq \frac{r}{\lambda B K}.$$

Для этого нам достаточно взять в качестве r_0 больший из корней уравнения

$$\sqrt{r}\left(\frac{10\sqrt{r_n^*}}{B} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + \frac{8\log\frac{1}{\delta}}{3n} = \frac{r}{\lambda B K}. \qquad (*)$$

Легко показать, что для такого r_0 выполнено $r_0 \geq r_n^*$ и получить верхнюю оценку на r_0.

- Вспомнив определение

$$V_r = \sup_{g \in \mathcal{G}_r} (Eg - E_n g),$$

получаем, что

$$\forall f \in \mathcal{F}^*,\ \forall K > 1: \quad Ef - E_n f \leq \frac{w(r_0, f)}{r_0}\frac{r_0}{\lambda K B} = \frac{w(r_0, f)}{\lambda K B}.$$

Локальный анализ: главный результат

Подробное доказательство: ШАГ 5

- Рассмотрим два случая:
  - Если для $f \in \mathcal{F}^*$ выполнено $Ef^2 \le r_0 \Rightarrow w(r_0, f) = r_0$. С учетом (*):

$$Ef \le E_n f + \frac{r_0}{\lambda K B} \le E_n f + 151 \frac{r_n^*}{B} + \frac{(6+7B)\log(1/\delta)}{n}$$

  - Если для $f \in \mathcal{F}^*$ выполнено $Ef^2 > r_0 \Rightarrow w(r_0, f) = \lambda^i r_0,\ Ef^2 \in (r_0\lambda^{i-1}, r_0\lambda^i]$:

$$Ef - E_n f \le \frac{w(r_0, f)}{\lambda B K} = \frac{\lambda^i r_0}{\lambda B K} = \frac{\lambda(\lambda^{i-1} r_0)}{\lambda B K} \le \frac{Ef^2}{BK} \le \frac{Ef}{K}$$

    или $Ef \le K/(K-1) E_n f$.

- Вместе мы получаем:

$$\forall f \in \mathcal{F}^*, K > 1: \quad Ef \le \frac{K}{K-1} E_n f + 151 \frac{r_n^*}{B} + \frac{(6+7B)\log(1/\delta)}{n}$$

- Для $\hat{f}_n = \ell_{\hat{h}_n} - \ell_{h^*} \in \mathcal{F}^*$ выполнено

$$E_n \hat{f}_n = L_n(\hat{h}_n) - L_n(h^*) \le 0.$$

■

План доклада

Введение: Теория Статистического Обучения (SLT)

- Определения и поставноки задач
- Подход Вапника–Червоненкиса
- Скорости сходимости порядка $O(n^{-1/2})$ («медленные»)

Раздел II: Первые примеры оценок «быстрого» порядка $o(n^{-1/2})$

- Общая идея
- Неравенство Бернштейна и учет дисперсии
- Задачи без шума и оценки порядка $O(n^{-1})$
- Задачи с шумом и конечным классом отображений: условия ограниченного шума Массара и Маммена-Цыбакова

Раздел III: Локальный анализ

- Неравенство Талаграна
- Задачи с шумом и бесконечным классом отображений
- Пример оценок порядка $O(n^{-1})$ в общих предположениях

Как выбирать $\psi_n(r)$ и оценивать r_n^*?

- Задача классификации с бинарной функцией потерь
- **Функция роста** класса $\mathcal{H}$:

$$S_{\mathcal{H}}(n) = \sup_S \Big| \big\{ \big(\ell_h(X_1, Y_1), \ldots, \ell_h(X_n, Y_n)\big) : h \in \mathcal{H} \big\} \Big|.$$

- **VC-размерность** класса $\mathcal{H}$:

$$\max\{h \in \mathbb{N} \colon S_{\mathcal{H}}(n) = 2^h\}.$$

  Иначе VC-размерность $= \infty$.

- Если VC-размерность h класса $\mathcal{H} \leq \infty$, то лемма Сауэра-Шелаха-Вапника-Червоненкиса гласит:

$$\forall n \geq h \colon \quad S_{\mathcal{H}}(n) \leq \left(\frac{e \cdot n}{h}\right)^h$$

Как выбирать $\psi_n(r)$ и оценивать r_n^*?

Теорема ((Massart, 2000))

Рассмотрим задачу с бинарной функцией потерь ℓ. Пусть, кроме того, класс отображений $\mathcal{H}$ имеет конечную VC-размерность $h < \infty$. Тогда справедливо:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}^* :\, \mathbb{E}[f^2] \le r} \mathbb{E}[f(X,Y)] - \hat{E}_n[f]\right] \le 24\sqrt{\frac{r \log S_{\mathcal{H}}(n)}{n}} \le \underbrace{24\sqrt{\frac{rh \log\left(\frac{en}{h}\right)}{n}}}_{\psi_n(r)}$$

Функция $\psi_n(r)$, очевидно, является подкоренной, и существует константа $C > 0$, такая что для ее неподвижной точки r_n^* справедливо:

$$r_n^* \le C \frac{h \log\left(\frac{en}{h}\right)}{n}.$$

Как выбирать $\psi_n(r)$ и оценивать r_n^*?

Существует $h \geq 0$, такая что с вер-ю 1 выполенно:

$$|2\eta(X) - 1| > h \qquad \text{(M)}$$

Следствие Рассмотрим задачу с бинарной функцией потерь ℓ и классом отображений $\mathcal{H}$, имеющим конечную VC-размерность $h < \infty$. Пусть, кроме того, Байесовское отображение g^* принадлежит классу $\mathcal{H}$, и выполнено условие на шум Массара (M). Тогда для всех $\delta > 0$ с вер-ю не меньше $1 - \delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) = O\left(\frac{h \log\left(\frac{en}{h}\right)}{n}\right).$$

- Этот результат справедлив для бесконечных классов $\mathcal{H}$
- Не требует $L^* = 0$

Вариации на тему . . .

- Результат может быть обобщен для условия Маммена-Цыбакова
- Ядра (Mendelson, 2003; Bartlett et al., 2005):
  - $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ — симметр. полож. опред. и $k(X', X'') \leq 1$
  - Пусть $\mathcal{C}_k$ — RKHS, соответствующий ядру k
  - $\{\lambda_1 \geq \lambda_2 \geq \dots\}$ — собственные значения линейного интегрального оператора $A_k f(x) = \int_{\mathcal{X}} K(x,z) f(z) dP(z)$.
  - Пусть $\mathcal{H} = \{f \in \mathcal{C}_k \colon \|f\|_k \leq 1\}$
  - Пусть ℓ — Липшицева

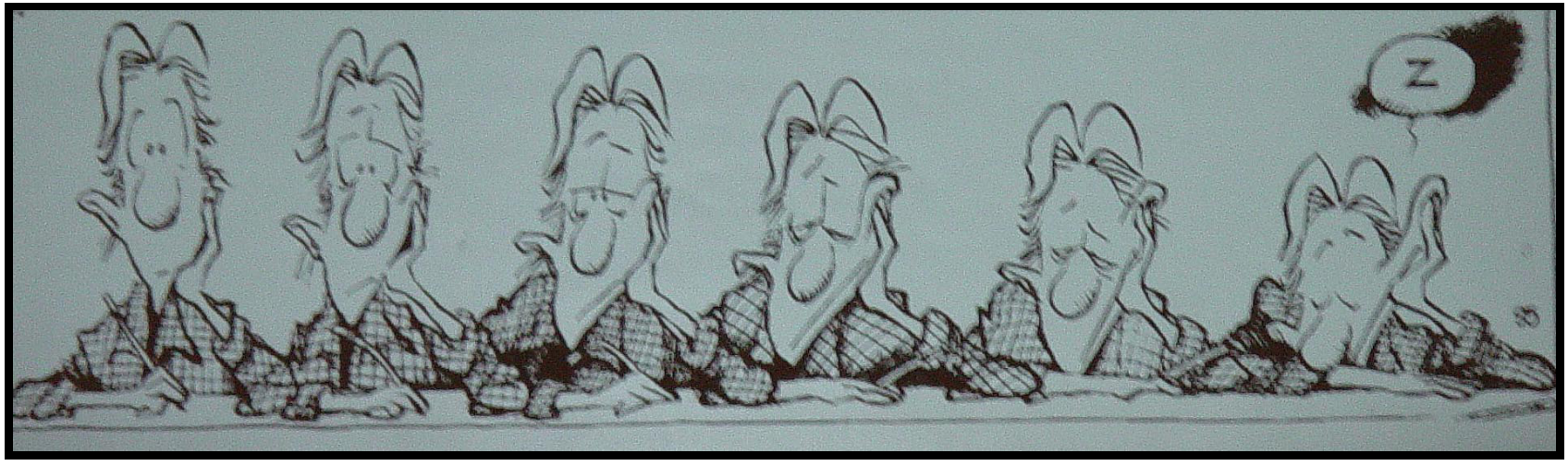
  Тогда

$$r_n^* \sim \min_{0 \leq p \leq n} \left(\frac{p}{n} + \sqrt{\frac{1}{n} \sum_{i>p} \lambda_i} \right)$$

- Локальный подход в трансдуктивной постановке SLT (Tolstikhin, Blanchard, Kloft, COLT 2014)
- С помощью локальной Радемахеровской сложности можно получить аналоги оценок, вычислимые по данным (Bartlett et al., 2005) (Koltchinskii, 2006)

Outro

Спасибо за внимание!
Вопросы?



P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 2005.

O Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I*, 334:495–500, 2002.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6): 2593–2656, 2006.

Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. In D. E. Gine and J.Wellner, editors, *High Dimensional Probability, II*, pages 443–457. Birkhauser, 1999.

W. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44, 1998.

Enno Mammen and Alexandre Tsybakov. Smooth discrimination analisys. *Annals of Statistics*, 27:1808–1829, 1999.

P Massart. *Concentration Inequalities and Model Selection: École D'Été de Probabilités de Saint-Flour 2003*. Ecole d'été de probabilités de Saint-Flour. Springer, 2007.

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Pascal Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9(6):245–303, 2000.

Shahar Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48, 2002.

Shahar Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759–771, December 2003.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126, 1996.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. Theory of pattern recognition. *Nauka, Moscow (in Russian)*, 1974. German

translation: W.N.Wapnik, A.Ya.Tschervonenkis (1979), Theorie der Zeichenerkennug, Akademia, Berlin.