0.1 Симметризация (Symmetrization)

Договоримся, что в случае использования верхнего индекса у выборок, он будет обозначать их мощность. Верхний индекс для упрощения понимания позже будем опускать.

$$\mathsf{E}_X R(X,F) = \mathsf{E}_{X,\varepsilon} \sup_{f\in F}\Big|\frac{1}{\ell}\sum_{i=1}^{l}\varepsilon_i f\big(X_i\big)\Big| = \text{(в нашей терминологии)} =$$

$$= \frac{1}{C_L^\ell}\frac{1}{C_\ell^{\ell/2}}\sum_{X\cup\bar{X}}\sum_{X=X_1^{\frac{\ell}{2}}\cup X_2^{\frac{\ell}{2}}}\sup_{f\in F}\Big|\frac{1}{\ell}\Big(\sum_{x\in X_1^{\frac{\ell}{2}}} f(x) - \sum_{x\in X_2^{\frac{\ell}{2}}} f(x)\Big)\Big| =$$

$$= \frac{1}{C_L^\ell}\frac{1}{C_\ell^{\ell/2}}\sum_{\substack{X_1^{\frac{\ell}{2}},\, X_2^{\frac{\ell}{2}}:\\ X_1^{\frac{\ell}{2}}\cap X_2^{\frac{\ell}{2}}=\emptyset}}\sup_{f\in F}\Big|\frac{1}{\ell}\Big(\sum_{x\in X_1^{\frac{\ell}{2}}} f(x) - \sum_{x\in X_2^{\frac{\ell}{2}}} f(x)\Big)\Big|.$$

Теперь оценим сверху математическое ожидание равномерного уклонения эмпирической ошибки от действительной ошибки при $\ell = k$:

$$\mathsf{E}\mathrm{Q}(X,F) = \frac{1}{C_L^\ell}\sum_{X\cup\bar{X}}\sup_{f\in F}\Big|\Big(\frac{1}{L}\sum_{x\in\mathbb{X}} f(x) - \frac{1}{\ell}\sum_{x\in X} f(x)\Big)\Big| = \frac{1}{2C_L^\ell}\sum_{X\cup\bar{X}}\sup_{f\in F}\Big|\frac{1}{\ell}\Big(\sum_{x\in\bar{X}} f(x) - \sum_{x\in X} f(x)\Big)\Big| = \ldots$$

Главная идея заключается в том, чтобы для каждого разбиения $X\cup\bar{X}$ представить $\sum_{x\in\bar{X}} f(x) - \sum_{x\in X} f(x)$ в виде суммы

$$\sum_{x\in X_1^{\frac{\ell}{2}}} f(x) - \sum_{x\in X_2^{\frac{\ell}{2}}} f(x) \; + \sum_{x\in\bar{X}\setminus X_1^{\frac{\ell}{2}}} f(x) - \sum_{x\in X\setminus X_2^{\frac{\ell}{2}}} f(x)$$

по всевозможным $X_1\subset\bar{X},\; X_2\subset X$.

$$\cdots = \frac{1}{2C_L^\ell}\frac{1}{C_\ell^{\frac{\ell}{2}}}\frac{1}{C_\ell^{\frac{\ell}{2}}}\sum_{X\cup\bar{X}}\sup_{f\in F}\left|\frac{1}{\ell}\sum_{\substack{X_1\subset\bar{X}\\ X_2\subset X}}\left(\sum_{x\in X_1} f(x) - \sum_{x\in X_2} f(x) \; + \sum_{x\in\bar{X}\setminus X_1} f(x) - \sum_{x\in X\setminus X_2} f(x)\right)\right| \leqslant$$

$$\leqslant \frac{2}{2C_L^\ell C_\ell^{\frac{\ell}{2}} C_\ell^{\frac{\ell}{2}}}\sum_{X\cup\bar{X}}\sum_{\substack{X_1\subset\bar{X}\\ X_2\subset X}}\frac{1}{\ell}\sup_{f\in F}\Big|\sum_{x\in X_1} f(x) - \sum_{x\in X_2} f(x)\Big| = \ldots$$

Остаётся заметить, что каждая пара (X_1, X_2) учитывается ровно $C_{L/2}^{\frac{\ell}{2}} = C_\ell^{\frac{\ell}{2}}$ раз, и мы получаем:

$$\cdots = \frac{C_\ell^{\frac{\ell}{2}}}{C_L^\ell C_\ell^{\frac{\ell}{2}} C_\ell^{\frac{\ell}{2}}}\sum_{\substack{X_1^{\frac{\ell}{2}},\, X_2^{\frac{\ell}{2}}:\\ X_1^{\frac{\ell}{2}}\cap X_2^{\frac{\ell}{2}}=\emptyset}}\frac{1}{\ell}\sup_{f\in F}\Big|\sum_{x\in X_1} f(x) - \sum_{x\in X_2} f(x)\Big| = R_\ell(F).$$

Итак

$$\mathsf{E}\mathrm{Q}(X,F) = \frac{1}{C_L^\ell}\sum_{X\cup\bar{X}}\sup_{f\in F}\Big|\Big(\frac{1}{L}\sum_{x\in\mathbb{X}} f(x) - \frac{1}{\ell}\sum_{x\in X} f(x)\Big)\Big| \leqslant \mathsf{E}\mathrm{R}(X,F).$$

0.2 Концентрация меры (concentration)

Теорема 0.1 (McDiarmid's inequality)**.** *Пусть функция* $g\colon X = (X_1, \ldots, X_\ell) \to \mathbb{R}$ *удовлетворяет условию ограниченных сумм с константами* $c_1, \ldots, c_\ell$*. Тогда*

$$\mathsf{P}\big\{Z - \mathsf{E}Z \geqslant t\big\} \leqslant e^{\frac{-2t^2}{C}};$$
$$\mathsf{P}\big\{|Z - \mathsf{E}Z| \geqslant t\big\} \leqslant 2e^{\frac{-2t^2}{C}},$$

где $C = \sum_{i=1}^{\ell} c_i^2$.

Применим эту теорему к случайной величине $\mathrm{Q}(X, F) - \mathrm{R}(X, F)$, которая удовлетворяет условию ограниченных сумм с константами $c_i = \frac{1}{\ell} + \frac{1}{2\ell} = \frac{3}{2\ell},\ i = 1, \ldots, l$. Получим следующее утверждение:

Утверждение 0.1. *С вероятностью не меньше* $1 - \delta$

$$\mathrm{Q}(X, F) \leqslant \mathrm{R}(X, F) + \frac{3}{2}\sqrt{\frac{\ln 1/\delta}{2\ell}},$$

или, что то же самое,

$$\sup_{f \in F}\Big|\frac{1}{L}\sum_{x \in \mathbb{X}} f(x) - \frac{1}{\ell}\sum_{x \in X} f(x)\Big| \leqslant \frac{1}{C_\ell^{\ell/2}} \sum_{X = X_1^{\frac{\ell}{2}} \cup X_2^{\frac{\ell}{2}}} \sup_{f \in F}\Big|\frac{1}{\ell}\Big(\sum_{x \in X_1^{\frac{\ell}{2}}} f(x) - \sum_{x \in X_2^{\frac{\ell}{2}}} f(x)\Big)\Big| + \frac{3}{2}\sqrt{\frac{\ln 1/\delta}{2\ell}}.$$

Доказательство. С вероятностью не меньше $1 - \delta$

$$\mathrm{Q}(X, F) - \mathrm{R}(X, F) + \mathsf{E}\mathrm{R}(X, F) - \mathsf{E}\mathrm{Q}(X, F) \leqslant \frac{3}{2}\sqrt{\frac{3\ln 1/\delta}{2\ell}};$$
$$\mathrm{Q}(X, F) \leqslant \mathrm{R}(X, F) + \underbrace{\big(\mathsf{E}\mathrm{Q}(X, F) - \mathsf{E}\mathrm{R}(X, F)\big)}_{\leqslant 0} + \frac{3}{2}\sqrt{\frac{\ln 1/\delta}{2\ell}}.$$

■