## 1. Bias-variance-noise decomposition

Нужно показать, что
$\mathbf{E}_{x,y}\mathbf{E}_{X^l}(y-a_{X^l}(x))^2 = \mathbf{E}_{x,y}(y-\mathbf{E}(y|x))^2+\mathbf{E}_{x,y}(\mathbf{E}(x|y)-\mathbf{E}_{X^l}a(x))^2+\mathbf{E}_{x,y}\mathbf{E}_{X^l}(a(x)-\mathbf{E}_{X^l}a(x))^2$

$\mathbf{E}_{x,y}\mathbf{E}_{X^l}(y-a_{X^l}(x))^2 = \mathbf{E}_{x,y}\mathbf{E}_{X^l}y^2 - 2\mathbf{E}_{x,y}\mathbf{E}_{X^l}ya_{X^l}(x) + \mathbf{E}_{x,y}\mathbf{E}_{X^l}a^2_{X^l}(x)$

$\mathbf{E}_{x,y}(y-\mathbf{E}(y|x))^2 = \mathbf{E}_{x,y}y^2 - 2\cdot\mathbf{E}_{x,y}y\mathbf{E}(x|y) + \mathbf{E}_{x,y}(\mathbf{E}_{x,y}(x|y))^2$

$\mathbf{E}_{x,y}(\mathbf{E}(x|y)-\mathbf{E}_{X^l}a(x))^2 = \mathbf{E}_{x,y}(\mathbf{E}(x|y))^2 - 2\cdot\mathbf{E}_{x,y}\cdot\mathbf{E}(x|y)\cdot\mathbf{E}_{X^l}a(x) + \mathbf{E}_{x,y}(\mathbf{E}_{X^l}a(x))^2$

$\mathbf{E}_{x,y}\mathbf{E}_{X^l}(a(x)-\mathbf{E}_{X^l}a(x))^2 = \mathbf{E}_{x,y}\mathbf{E}_{X^l}a(x)^2 - 2\cdot\mathbf{E}_{x,y}\mathbf{E}_{X^l}a(x)\cdot\mathbf{E}_{X^l}a(x) + \mathbf{E}_{x,y}(\mathbf{E}_{X_l}a(x))^2$
Видим, что всё сокращается, если просуммировать.

## 2. Смещение и разброс в бэггинге

$a(x) = \frac{1}{M}\sum\limits_{m=1}^{M} a_m(x).$

По условию все базовые алгоритмы распределены одинаково. Найдем смещение алгоритма $a(x)$:

$\mathbf{Bias^2}a(x) = \mathbf{E}_{x,y}(\mathbf{E}(y|x) - \mathbf{E}_{X^l}(a(x))^2$

Выразим $\mathbf{E}_{X^l}(a(x))$ через $a_1,\dots,a_m$:
$\mathbf{E}_{X^l}a(x) = \mathbf{E}_{X^l}(\frac{1}{M}\sum\limits_{m=1} Ma_m(x) = \frac{1}{M}M\cdot\mathbf{E}_{X^l}(a_1(x)) = \mathbf{E}_{X^l}(a_1(x)).$

Таким образом, смещение алгоритма не изменяется.

Теперь найдем разброс нового алгоритма, положив смещение алгоритма равным 0 (можно просто рассмотреть новый алгоритм, вычтя нужное число):

$\mathbf{Var}a(x) = \mathbf{E}_{x,y}\mathbf{E}_{X^l}(a(x)-\mathbf{E}_{X^l}a(x))^2 = \mathbf{E}_{x,y}\mathbf{E}_{X^l}(a(x)-\mathbf{E}_{X^l}a_1(x))^2 = \mathbf{E}_{x,y}\mathbf{E}_{X^l}a(x)^2$
Заменим $a(x)$ на $\frac{1}{M}\sum\limits_{m=1}^{M} a_m(x)$ в $\mathbf{E}_{X^l}a(x)^2$:

$\mathbf{E}_{X^l}a(x)^2 = \mathbf{E}_{X^l}(\frac{1}{M}\sum\limits_{m=1}^{M} a_m(x))^2 = \frac{1}{M^2}M\cdot\mathbf{E}_{X^l}a_1(x)^2 + \frac{1}{M^2}\sum\limits_{i\neq j}\mathbf{E}_{X^l}a_i(x)a_j(x) = \frac{1}{M}\mathbf{E}_{X^l}a_1(x)^2 + \frac{1}{M^2}\sum\limits_{i\neq j} r_{ij}\mathbf{E}_{X^l}a_i(x)\cdot\mathbf{E}_{X^l}a_j(x).$

Таким образом,

$\mathbf{Var}a(x) \leq \frac{1}{M}\mathbf{Var}a_1(x) + \frac{1}{M^2}\cdot M(M-1)\mathbf{Var}a_1(x) = \mathbf{Var}a_1(x)$
неравенство следует из оценки корреляций алгоритмов 1. Если хотя бы между несколькоми алгоритмами корреляция будет меньше 1, то получим строгое неравенство.

Т.е. разброс алгоритма уменьшается при бэггинге.

## 3. Корреляция ответов базовых алгоритмов

Пусть $\xi_1, \ldots, \xi_M$ – одинаково распределенные случайные величины с дисперсией $\sigma^2$, и $\forall i, j$ : $\frac{cov \xi_i \xi_j}{\sigma^2} = \rho > 0$.
Найдем дисперсию среднего этих случайных величин, считая, что их мат. ожидания 0 (в противном случаем будем просто вычтем мат. ожидание и будем работать с новыи случайными величинами, у которых мат. ожид. 0)

$\mathbf{D}(\frac{1}{m}\sum\limits_{i=1}^{M}\xi_i) = \frac{1}{M^2}\mathbf{E}(\xi_1 + \ldots + \xi_M)^2 = \frac{1}{M^2} \cdot M\sigma^2 + \frac{1}{M^2}\sum\limits_{i \neq j}\mathbf{cov}(\xi_i, \xi_j) = \frac{\sigma^2}{M} + \frac{1}{M^2} \cdot M(M-1)\rho\sigma^2 = \frac{\sigma^2}{M} + \rho\sigma^2 - \frac{\rho\sigma^2}{M} = \rho\sigma^2 + (1-\rho)\frac{\sigma^2}{M}$.