Стандартные эвристики для метода стохастического градиента

Применимы все те же эвристики, что и в обычном SG:

- инициализация весов;
- порядок предъявления объектов;
- оптимизация величины градиентного шага;
- регуляризация (сокращение весов);

Кроме того, появляются новые проблемы:

- выбор функций активации в каждом нейроне;
- выбор числа слоёв и числа нейронов;
- выбор значимых связей;

1. Начальное приближение — послойное обучение сети.
Нейроны настраиваются как отдельные линейные алгоритмы

- либо по случайной подвыборке $X' \subseteq X^\ell$;
- либо по случайному подмножеству входов;
- либо из различных случайных начальных приближений;

тем самым обеспечивается *различность* нейронов.

2. Выбивание из локальных минимумов (jogging of weights).

3. Адаптивный градиентный шаг (метод скорейшего спуска).

4. Метод сопряжённых градиентов и chunking — разбиение суммы $Q(w) = \sum\limits_{i=1}^{\ell} \mathscr{L}_i(w)$ по подмножествам объектов (chunks).

Метод Ньютона-Рафсона (второго порядка):

$$w := w - \eta\bigl(\mathscr{L}_i''(w)\bigr)^{-1}\mathscr{L}_i'(w),$$

где $\bigl(\mathscr{L}_i''(w)\bigr) = \bigl(\frac{\partial^2 \mathscr{L}_i(w)}{\partial w_{jh} \partial w_{j'h'}}\bigr)$ — гессиан, размера $\bigl(H(n+M+1)+M\bigr)^2$.

Эвристика. Считаем, что гессиан диагонален:

$$w_{jh} := w_{jh} - \eta \left(\frac{\partial^2 \mathscr{L}_i(w)}{\partial w_{jh}^2} + \mu \right)^{-1} \frac{\partial \mathscr{L}_i(w)}{\partial w_{jh}},$$

η — темп обучения,
μ — параметр, предотвращающий обнуление знаменателя.

Отношение η/μ есть темп обучения на ровных участках функционала $\mathscr{L}_i(w)$, где вторая производная обнуляется.