Глубокое обучение III: инициализация, нормализация, градиентный спуск

Сергей Николенко

uData School — Киев
28 августа 2018 г.

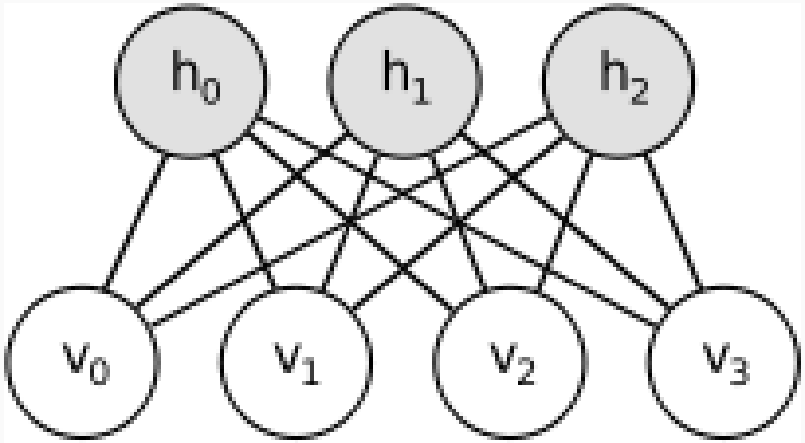
Random facts:

- 28 августа 1851 г. открыто железнодорожное движение между Санкт-Петербургом и Москвой
- 28 августа 1898 г. Калеб Брэдхем переименовал свой «Brad's Drink» в «Pepsi-Cola»
- 28 августа 1789 г. Вильям Гершель открыл спутник Сатурна Энцелад, а 28 августа 1993 г. аппарат «Галилео» открыл Дактиль, вращающийся вокруг 243 Ида — первый известный спутник астероида

ИНИЦИАЛИЗАЦИЯ ВЕСОВ

Предобучение без учителя

- Революция глубокого обучения началась с *предобучением без учителя* (unsupervised pretraining).
- Главная идея: добраться до хорошей области пространства весов, затем уже сделать fine-tuning градиентным спуском.
- Ограниченные машины Больцмана (restricted Boltzmann machines):



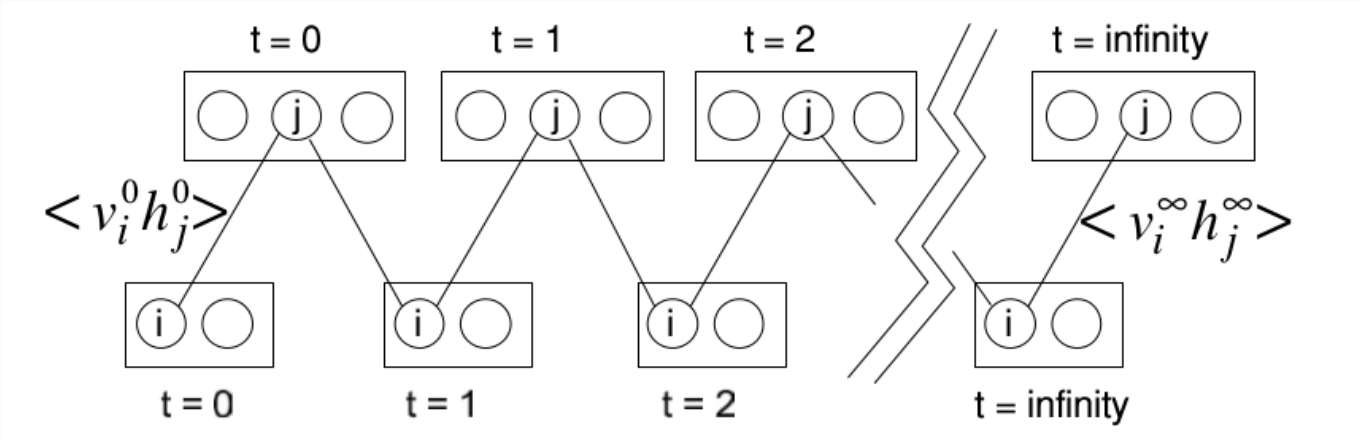
Предобучение без учителя

- Это ненаправленная графическая модель, задающая распределение

$$p(\mathbf{v}) = \sum_h p(\mathbf{v}, h) = \frac{1}{Z} \sum_h e^{-E(\mathbf{v},h)}, \text{ где}$$

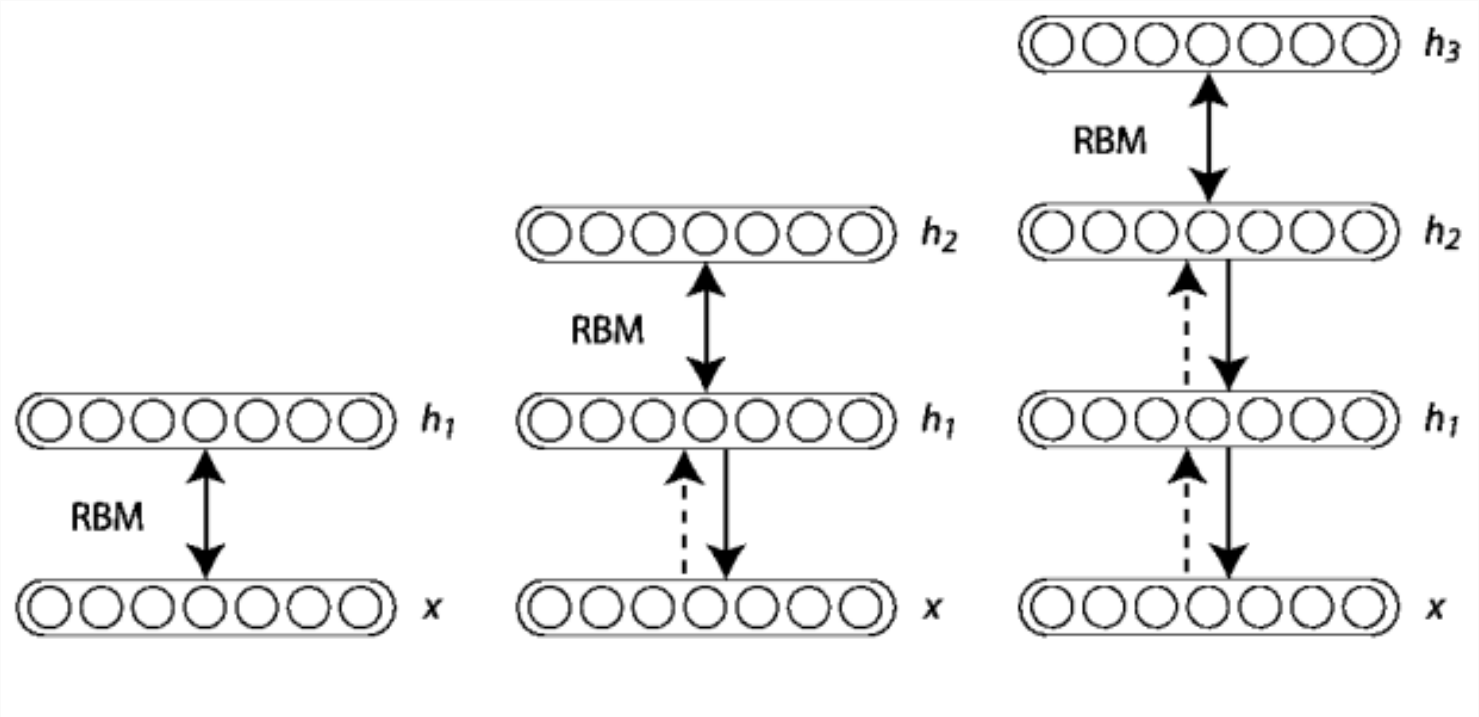
$$E(\mathbf{v}, h) = -\mathbf{b}^\top \mathbf{v} - \mathbf{c}^\top h - h^\top W \mathbf{v}.$$

- Обучают алгоритмом Contrastive Divergence (приближение к сэмплированию по Гиббсу).



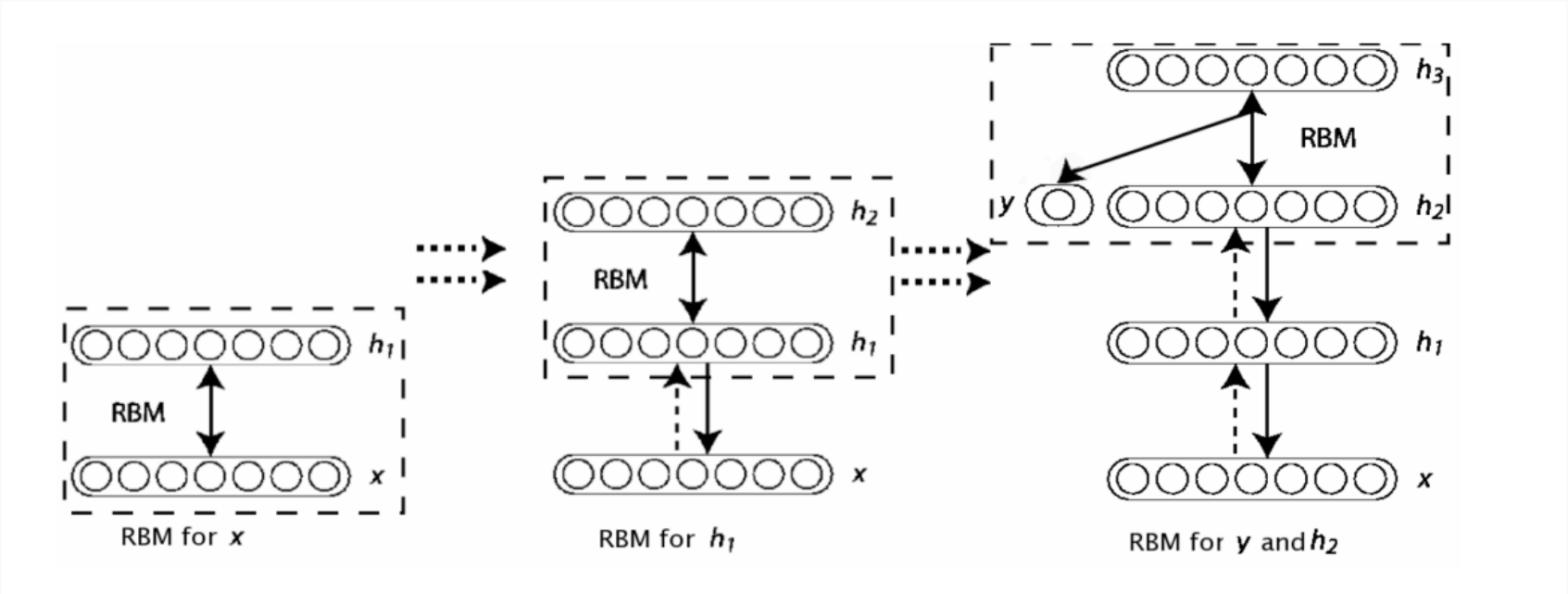
Предобучение без учителя

- Из RBM можно сделать глубокие сети, поставив одну на другую:



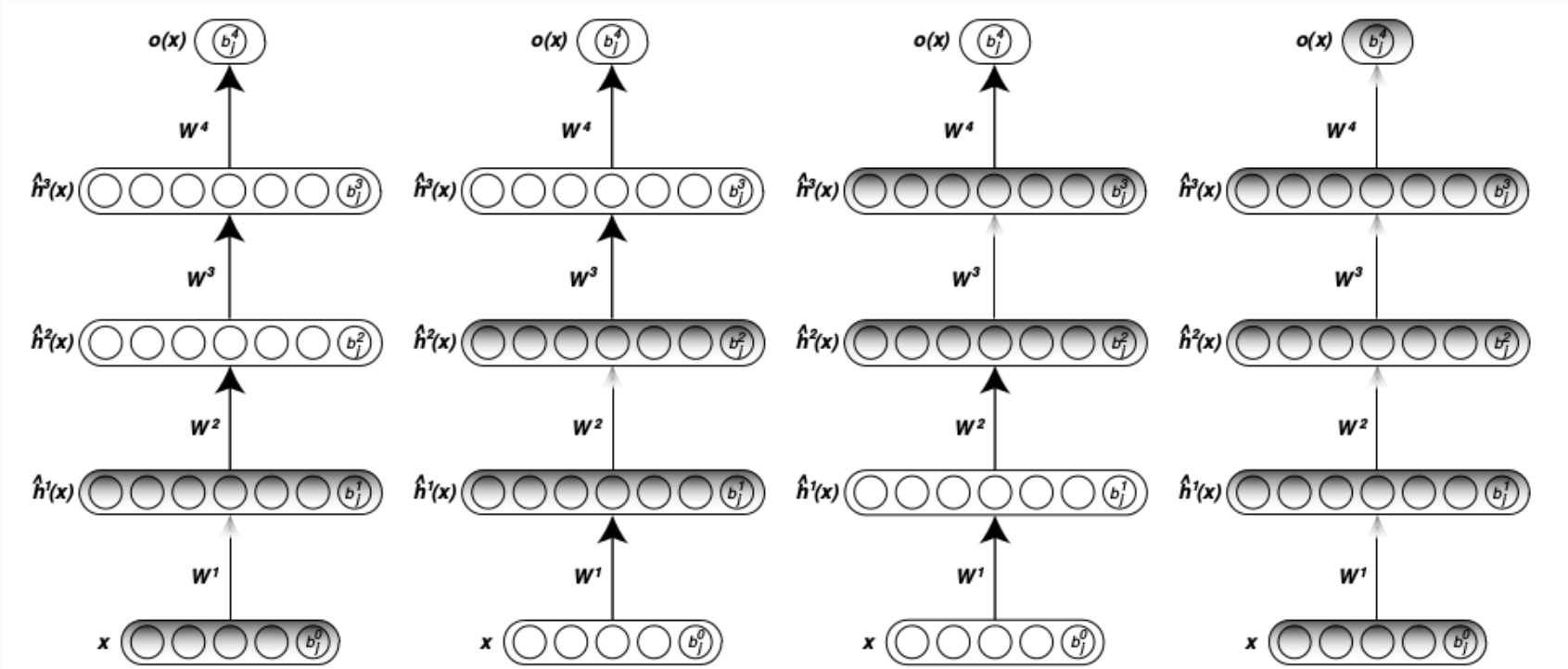
Предобучение без учителя

- И вывод можно вести последовательно, уровень за уровнем:



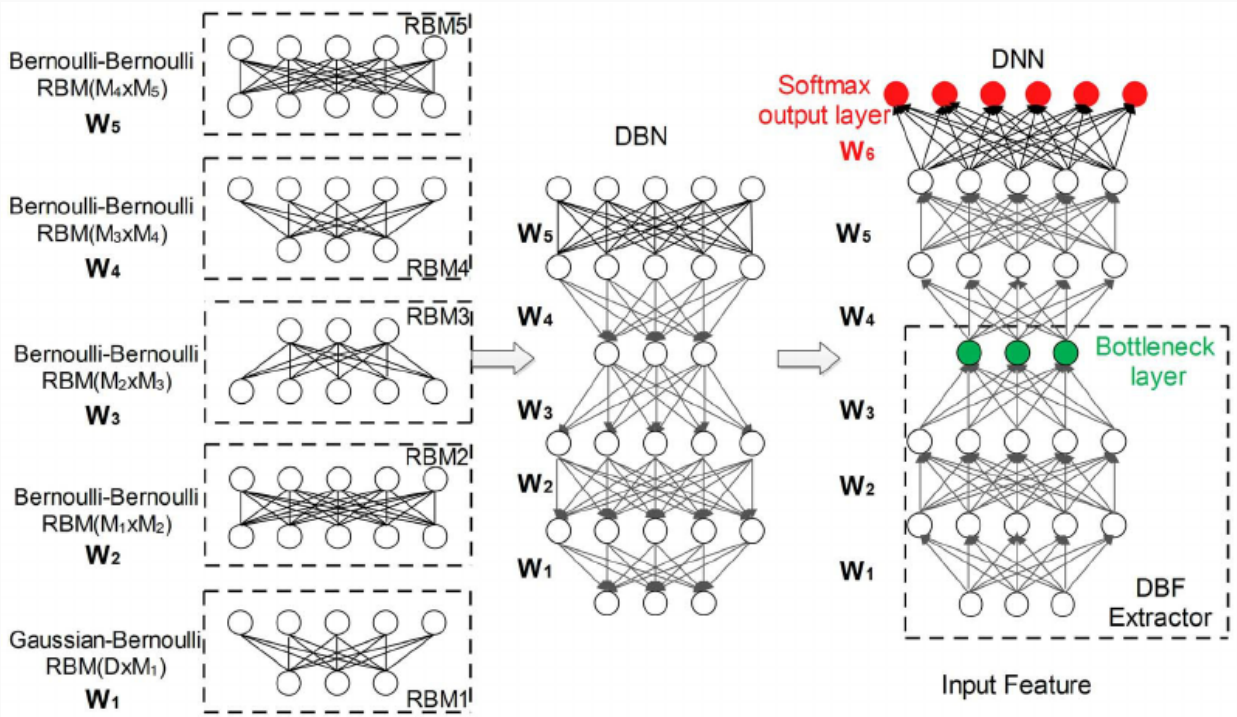
Предобучение без учителя

- А потом уже дообучать градиентным спуском (fine-tuning).



Предобучение без учителя

- Этот подход привёл к прорыву в распознавании речи.



Предобучение без учителя

- Но обучать глубокие сети из RBM довольно сложно, они хрупкие, и вычислительно тоже нелегко.
- И сейчас уже не очень-то и нужны сложные модели вроде RBM для того, чтобы попасть в хорошую начальную область.
- Инициализация весов — важная часть этого.

Инициализация Хавьера

- *Xavier initialization* (Glorot, Bengio, 2010).
- Рассмотрим простой линейный нейрон:

$$y = \mathbf{w}^\top \mathbf{x} + b = \sum_i w_i x_i + b.$$

- Его дисперсия равна

$$\begin{aligned}\operatorname{Var}[y_i] = \operatorname{Var}[w_i x_i] = \mathbb{E}[w_i^2 x_i^2] - (\mathbb{E}[w_i x_i])^2 = \\ = \mathbb{E}[x_i]^2 \operatorname{Var}[w_i] + \mathbb{E}[w_i]^2 \operatorname{Var}[x_i] + \operatorname{Var}[w_i] \operatorname{Var}[x_i].\end{aligned}$$

Инициализация Хавьера

- Его дисперсия равна

$$\operatorname{Var}[y_i] = \operatorname{Var}[w_i x_i] = \mathbb{E}\left[w_i^2 x_i^2\right] - \left(\mathbb{E}[w_i x_i]\right)^2 = \\ = \mathbb{E}[x_i]^2 \operatorname{Var}[w_i] + \mathbb{E}[w_i]^2 \operatorname{Var}[x_i] + \operatorname{Var}[w_i] \operatorname{Var}[x_i].$$

- Для нулевого среднего весов

$$\operatorname{Var}[y_i] = \operatorname{Var}[w_i] \operatorname{Var}[x_i].$$

- И если w_i и x_i инициализированы независимо из одного и того же распределения,

$$\operatorname{Var}[y] = \operatorname{Var}\left[\sum_{i=1}^{n_{\text{out}}} y_i\right] = \sum_{i=1}^{n_{\text{out}}} \operatorname{Var}[w_i x_i] = n_{\text{out}} \operatorname{Var}[w_i] \operatorname{Var}[x_i].$$

- Иначе говоря, дисперсия на выходе пропорциональна дисперсии на входе с коэффициентом $n_{\text{out}} \operatorname{Var}[w_i]$.

Инициализация Хавьера

- До (Glorot, Bengio, 2010) стандартным способом инициализации было

$$w_i \sim U\left[-\frac{1}{\sqrt{n_{\text{out}}}}, \frac{1}{\sqrt{n_{\text{out}}}}\right].$$

- См., например, *Neural Networks: Tricks of the Trade*.
- Так что с дисперсиями получается

$$\text{Var}\left[w_i\right] = \frac{1}{12}\left(\frac{1}{\sqrt{n_{\text{out}}}} + \frac{1}{\sqrt{n_{\text{out}}}}\right)^2 = \frac{1}{3n_{\text{out}}}, \text{ и}$$

$$n_{\text{out}}\text{Var}\left[w_i\right] = \frac{1}{3},$$

и после нескольких уровней сигнал совсем умирает; аналогичный эффект происходит и в backprop.

Инициализация Хавьера

- Инициализация Хавьера — давайте попробуем уменьшить изменение дисперсии, т.е. взять

$$\operatorname{Var}[w_i] = \frac{2}{n_{\text{in}} + n_{\text{out}}};$$

для равномерного распределения это

$$w_i \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_{\text{in}} + n_{\text{out}}}}, \frac{\sqrt{6}}{\sqrt{n_{\text{in}} + n_{\text{out}}}}\right].$$

- Но это работает только для симметричных активаций, т.е. не для ReLU...

Инициализация Хавьера

- ...до работы (He et al., 2015). Вернёмся к

$$\operatorname{Var}[w_i x_i] = \mathbb{E}[x_i]^2 \operatorname{Var}[w_i] + \mathbb{E}[w_i]^2 \operatorname{Var}[x_i] + \operatorname{Var}[w_i] \operatorname{Var}[x_i]$$

- Мы теперь можем обнулить только второе слагаемое:

$$\operatorname{Var}[w_i x_i] = \mathbb{E}[x_i]^2 \operatorname{Var}[w_i] + \operatorname{Var}[w_i] \operatorname{Var}[x_i] = \operatorname{Var}[w_i] \mathbb{E}[x_i^2], \text{ и}$$

$$\operatorname{Var}\left[y^{(l)}\right] = n_{\text{in}}^{(l)} \operatorname{Var}\left[w^{(l)}\right] \mathbb{E}\left[\left(x^{(l)}\right)^2\right].$$

Инициализация Хавьера

- Мы теперь можем обнулить только второе слагаемое:

$$\mathrm{Var}\left[y^{(l)}\right] = n_{\mathrm{in}}^{(l)} \mathrm{Var}\left[w^{(l)}\right] \mathbb{E}\left[\left(x^{(l)}\right)^2\right].$$

- Предположим, что $x^{(l)} = \max(0, y^{(l-1)})$, и у $y^{(l-1)}$ симметричное распределение вокруг нуля. Тогда

$$\mathbb{E}\left[\left(x^{(l)}\right)^2\right] = \frac{1}{2}\mathrm{Var}\left[y^{(l-1)}\right], \quad \mathrm{Var}\left[y^{(l)}\right] = \frac{n_{\mathrm{in}}^{(l)}}{2}\mathrm{Var}\left[w^{(l)}\right]\mathrm{Var}\left[y^{(l-1)}\right].$$

Инициализация Хавьера

- И это приводит к формуле для дисперсии активации ReLU; теперь нет никакого n_{out}:

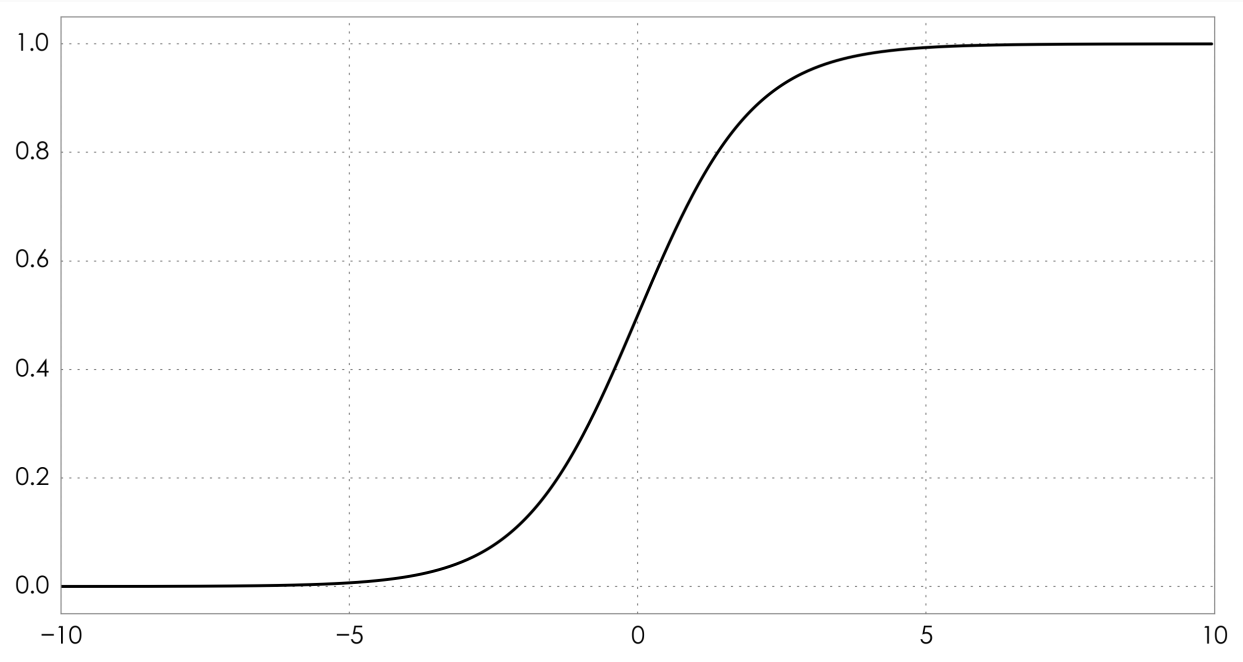
$$\operatorname{Var}[w_i] = 2/n_{\text{in}}^{(l)}.$$

- Кстати, равномерную инициализацию делать не обязательно, можно и нормальное распределение:

$$w_i \sim \mathcal{N}\left(0, \sqrt{2/n_{\text{in}}^{(l)}}\right).$$

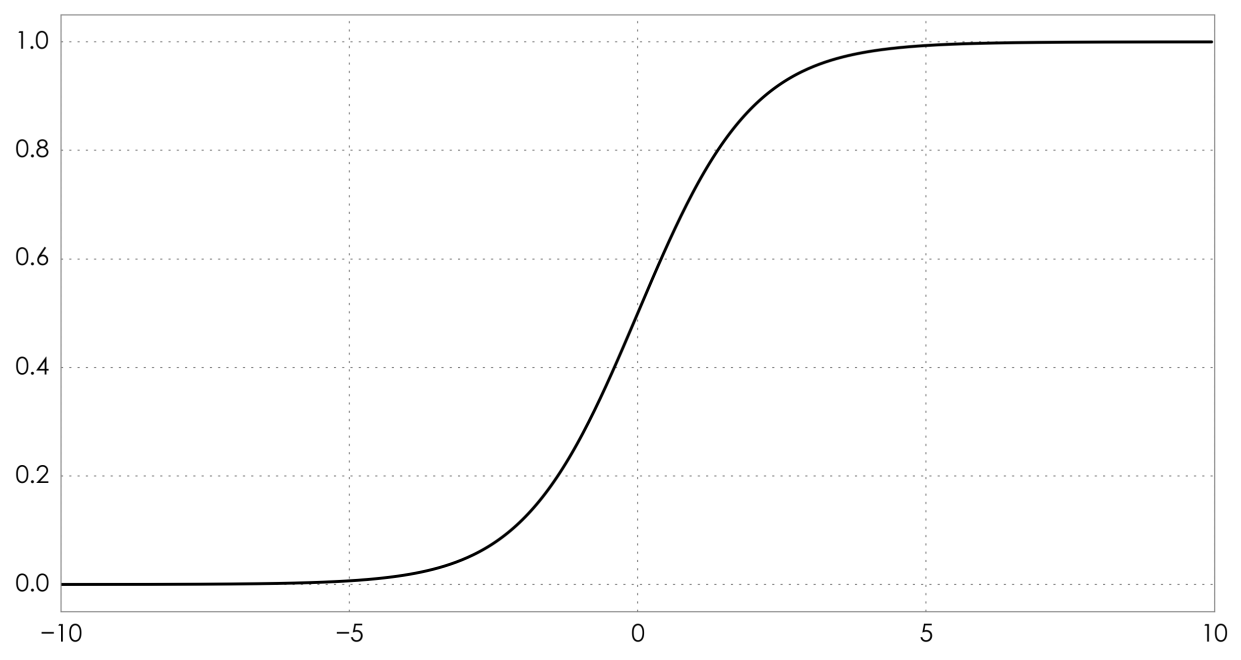
О сигмоидах

- Кстати, о (Glorot, Bengio, 2010) – ещё одна важная идея.
- Эксперименты показали, что $\sigma(x) = \frac{1}{1+e^{-x}}$ работает в глубоких сетях довольно плохо.



О сигмоидах

- *Насыщение*: если $\sigma(x)$ уже «обучилась», т.е. даёт большие по модулю значения, то её производная близка к нулю и «поменять мнение» трудно.



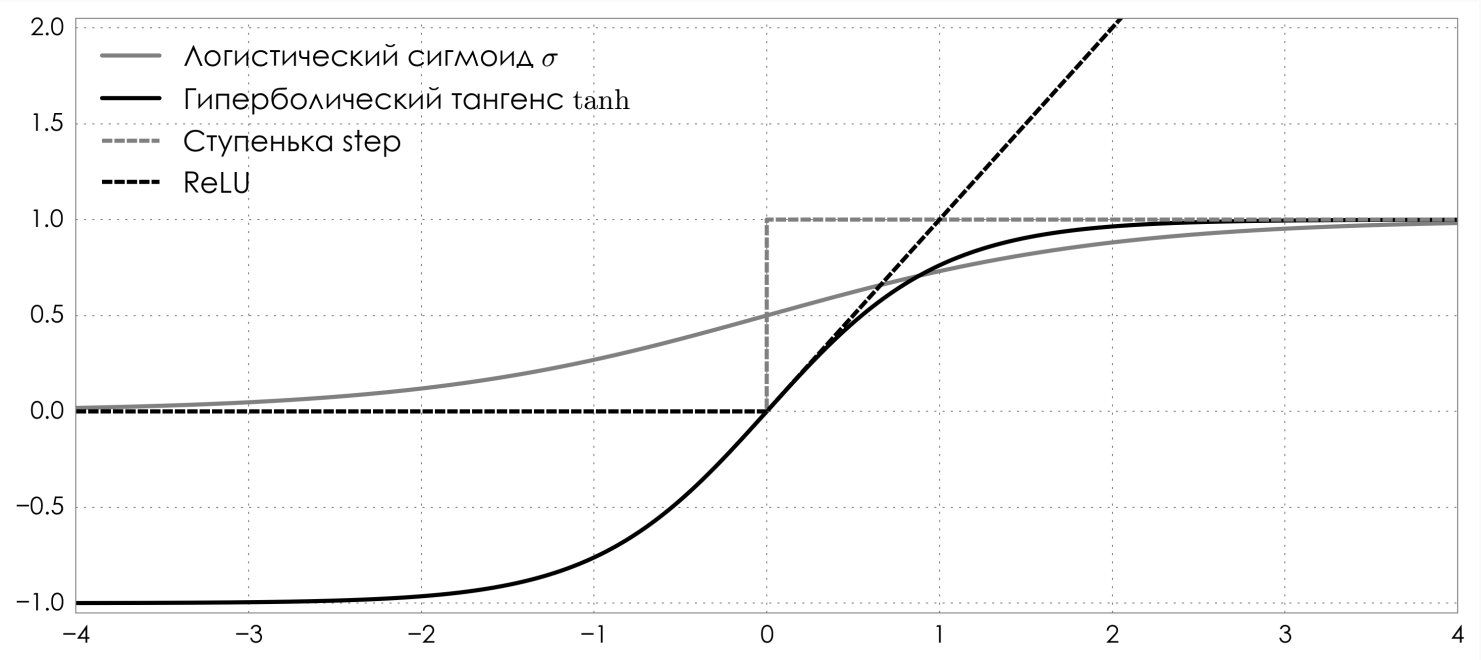
- Но ведь другие тоже насыщаются? В чём разница?

О сигмоидах

- Рассмотрим последний слой сети $h(W\mathbf{a} + \mathbf{b})$, где $\mathbf{a}$ — выходы предыдущего слоя, $\mathbf{b}$ — свободные члены, h — функция активации последнего уровня, обычно $\mathrm{softmax}$.
- Когда мы начинаем оптимизировать сложную функцию потерь, поначалу выходы h не несут полезной информации о входах, ведь первые уровни ещё не обучены.
- Тогда неплохим приближением будет константная функция, выдающая средние значения выходов.
- Это значит, что $h(W\mathbf{a} + \mathbf{b})$ подберёт подходящие свободные члены $\mathbf{b}$ и постарается обнулить слагаемое Wh, которое поначалу скорее шум, чем сигнал.

О сигмоидах

- Иначе говоря, в процессе обучения мы постараемся привести выходы предыдущего слоя к нулю.
- Здесь и проявляется разница: у $\sigma(x) = \frac{1}{1+e^{-x}}$ область значений $(0, 1)$ при среднем $\frac{1}{2}$, и при $\sigma(x) \to 0$ будет и $\sigma'(x) \to 0$.
- А у $\tanh$ наоборот: когда $\tanh(x) \to 0$, $\tanh'(x)$ максимальна.



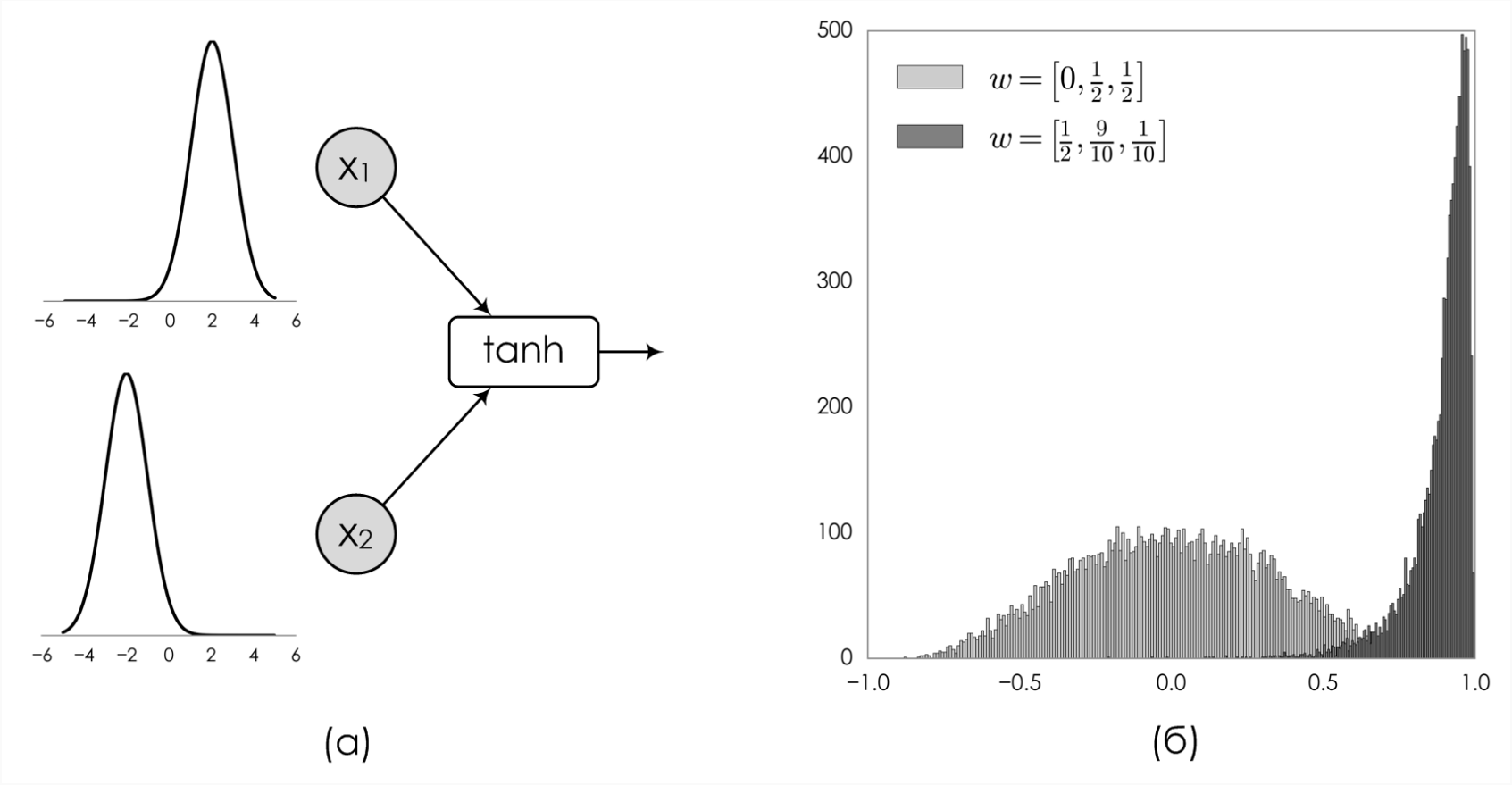
Нормализация по мини-батчам

Нормализация по мини-батчам

- Ещё одна важная проблема в глубоких сетях: *внутренний сдвиг переменных* (internal covariate shift).
- Когда меняются веса слоя, меняется распределение его выходов.
- Это значит, что следующему уровню придётся всё начинать заново, он же не ожидал таких входов, не видел их раньше!
- Более того, нейроны следующего уровня могли уже и насытиться, и им теперь сложно быстро обучиться заново.
- Это серьёзно мешает обучению.

Нормализация по мини-батчам

- Вот характерный пример:



- Что делать?

Нормализация по мини-батчам

- Можно пытаться нормализовать входы каждого уровня.
- Не работает: рассмотрим для простоты уровень с одним только bias b и входами u:

$$\hat{\mathbf{x}} = \mathbf{x} - \mathbb{E}[\mathbf{x}], \text{ где } \mathbf{x} = u + b.$$

- На следующем шаге градиентного спуска получится $b := b + \Delta b$...
- ...но $\hat{\mathbf{x}}$ не изменится:

$$u + b + \Delta b - \mathbb{E}[u + b + \Delta b] = u + b - \mathbb{E}[u + b].$$

- Так что всё обучение сведётся к тому, что b будет неограниченно расти — не очень хорошо.

Нормализация по мини-батчам

- Можно пытаться добавить нормализацию как отдельный слой:

$$\hat{\mathbf{x}} = \mathrm{Norm}(\mathbf{x}, \mathcal{X}).$$

- Это лучше, но теперь этому слою на входе нужен весь датасет $\mathcal{X}$!
- И на шаге градиентного спуска придётся вычислить $\frac{\partial \mathrm{Norm}}{\partial \mathbf{x}}$ и $\frac{\partial \mathrm{Norm}}{\partial \mathcal{X}}$, да ещё и матрицу ковариаций

$$\mathrm{Cov}[\mathbf{x}] = \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \left[\mathbf{x}\mathbf{x}^\top\right] - \mathbb{E}\left[\mathbf{x}\right] \mathbb{E}\left[\mathbf{x}\right]^\top.$$

- Это точно не сработает.

Нормализация по мини-батчам

- Решение в том, чтобы нормализовать каждый вход отдельно, и не по всему датасету, а по текущему кусочку; это и есть *нормализация по мини-батчам* (batch normalization).
- После нормализации по мини-батчам получим

$$\hat{x}_k = \frac{x_k - \mathbb{E}[x_k]}{\sqrt{\mathrm{Var}[x_k]}},$$

  где статистики подсчитаны по текущему мини-батчу.
- Ещё одна проблема: теперь пропадают нелинейности!
- Например, σ теперь практически всегда близка к линейной.

Нормализация по мини-батчам

- Чтобы это исправить, нужно добавить гибкости уровню batchnorm.
- В частности, нужно разрешить ему обучаться иногда *ничего не делать* со входами.
- Так что вводим дополнительные параметры (сдвиг и растяжение):

$$y_k = \gamma_k \hat{x}_k + \beta_k = \gamma_k \frac{x_k - \mathbb{E}[x_k]}{\sqrt{\mathrm{Var}[x_k]}} + \beta_k.$$

- γ_k и β_k — это новые переменные, тоже будут обучаться градиентным спуском, как веса.

Нормализация по мини-батчам

- И ещё добавим ϵ в знаменатель, чтобы на ноль не делить.
- Теперь мы можем формально описать слой батч-нормализации — для очередного мини-батча $B = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$:
  - вычислить базовые статистики по мини-батчу

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i, \quad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \mu_B)^2,$$

  - нормализовать входы

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mu_b}{\sqrt{\sigma_B^2 + \epsilon}},$$

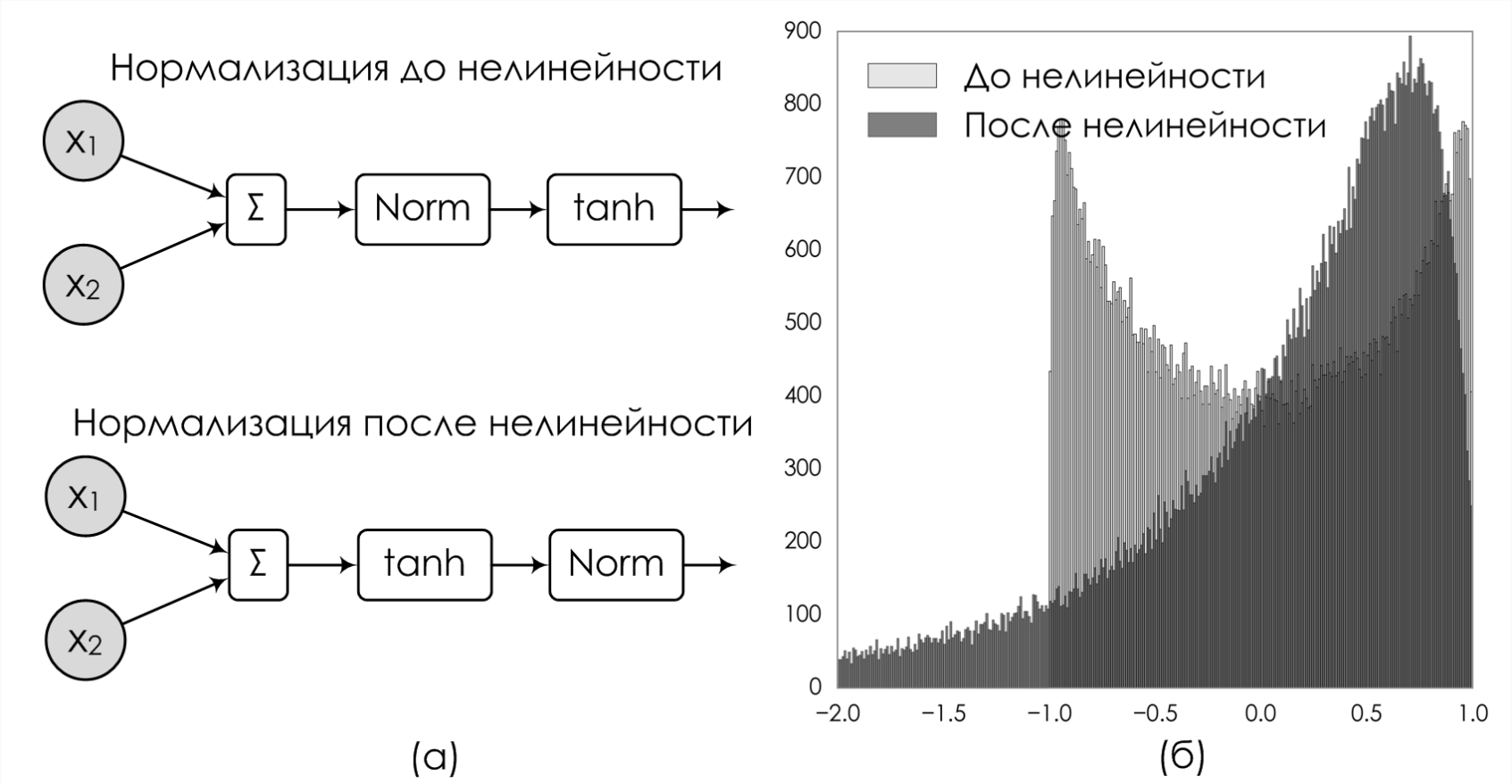
  - вычислить результат

$$\mathbf{y}_i = \gamma\mathbf{x}_i + \beta.$$

- Через всё это совершенно стандартным образом пропускаются градиенты, в том числе по γ и β.

Нормализация по мини-батчам

- Последнее замечание: важно, куда поместить слой batchnorm.
- Можно до, а можно после нелинейности.



Варианты

- Нормализация по мини-батчам сейчас стала фактически стандартом.
- Очередная очень крутая история, примерно как дропаут.
- Но мысль идёт и дальше:
  - (Laurent et al., 2016): BN не помогает рекуррентным сетям;
  - (Cooijmans et al., 2016): recurrent batch normalization;
  - (Salimans and Kingma, 2016): нормализация весов для улучшения обучения — давайте добавим веса как

$$h_i = f\left(\frac{\gamma}{\|\mathbf{w}_i\|}\mathbf{w}_i^\top \mathbf{x} + b_i\right);$$

    тогда мы будем перемасштабировать градиент, стабилизировать его норму и приближать его матрицу ковариаций к единичной, что улучшает обучение.

ВАРИАНТЫ ГРАДИЕНТНОГО СПУСКА

Метод моментов

- «Ванильный» стохастический градиентный спуск:

$$\theta_t = \theta_{t-1} - \eta \nabla E(\mathbf{x}_t, \theta_{t-1}, y_t).$$

- Всё зависит от скорости обучения η.
- Первая мысль — пусть η уменьшается со временем:
  - линейно (linear decay):

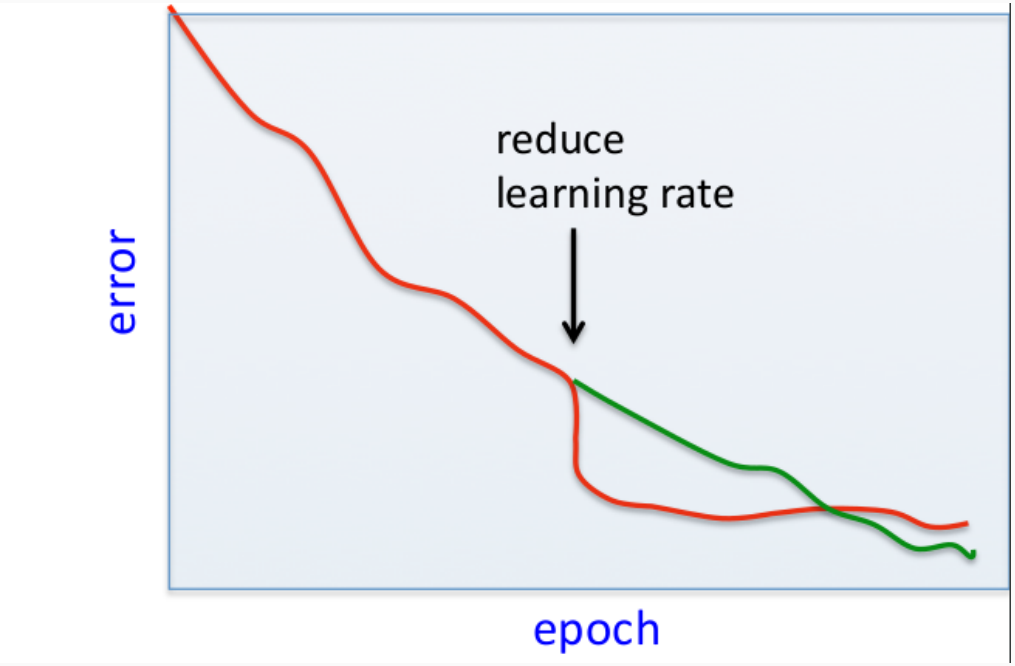
$$\eta = \eta_0 \left(1 - \frac{t}{T}\right);$$

  - или экспоненциально (exponential decay):

$$\eta = \eta_0 e^{-\frac{t}{T}}.$$

Метод моментов

- Скорость обучения лучше не уменьшать слишком быстро.



- Но это в любом случае никак не учитывает собственно E; лучше быть *адаптивным*.

Метод моментов

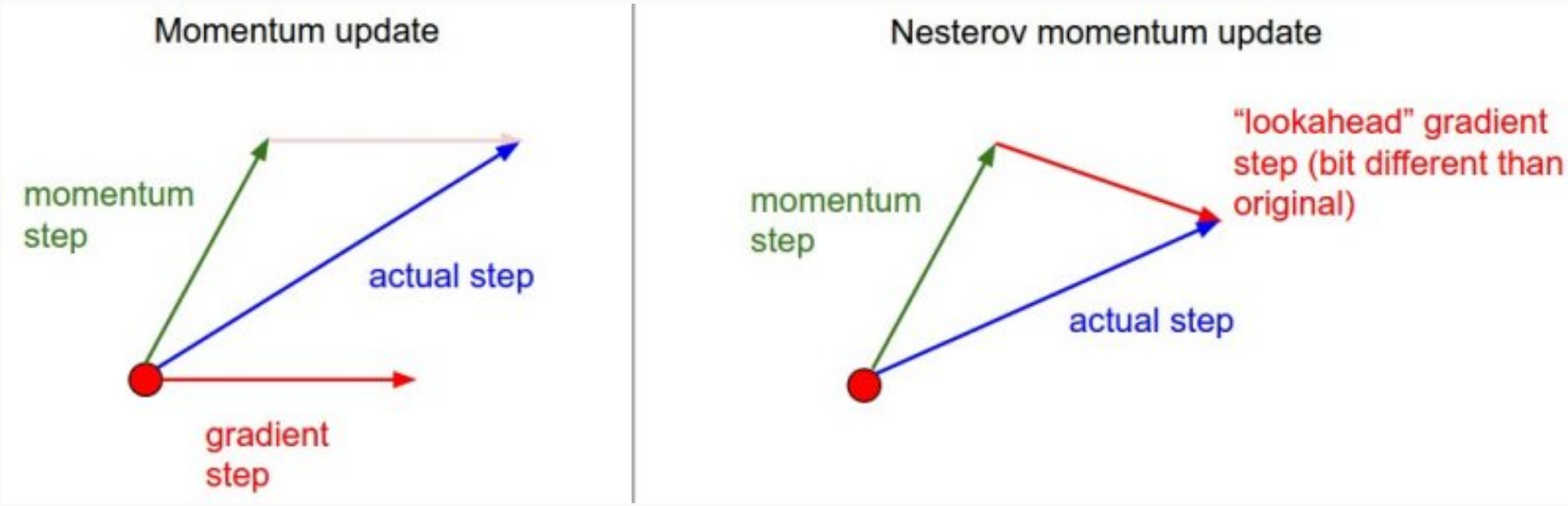
- *Метод моментов* (momentum): сохраним часть скорости, как у материальной точки.
- С инерцией получается

$$u_t = \gamma u_{t-1} + \eta \nabla_\theta E(\theta),$$
$$\theta = \theta - u_t.$$

- И теперь мы сохраняем γu_{t-1}.

Метод моментов

- Но на самом деле мы уже знаем, что попадём в γu_{t-1} на промежуточном шаге.
- Давайте прямо там, на полпути, и вычислим градиент!



Метод моментов

- *Метод Нестерова* (Nesterov's momentum):

$$u_t = \gamma u_{t-1} + \eta \nabla_\theta E(\theta - \gamma u_{t-1})$$

- Можно ли ещё лучше?..

Метод моментов

- ...ну, можно попробовать методы второго порядка.
- Метод Ньютона:

$$E(\theta) \approx E(\theta_0) + \nabla_\theta E(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\top H(E(\theta))(\theta - \theta_0).$$

- Когда работает, это обычно гораздо быстрее, и нет никакой η, ничего настраивать не надо.
- Но нужно считать гессиан $H(E(\theta))$, и это нереально.

Метод моментов

- Есть, правда, приближения.
- L-BFGS (limited memory Broyden–Fletcher–Goldfarb–Shanno):
  - строим аппроксимацию к H^{-1};
  - для этого сохраняем последовательно апдейты аргументов функции и градиентов и выражаем через них H^{-1}.
- Интересный открытый вопрос: можно ли заставить L-BFGS работать для deep learning?
- Но пока не получается («в лоб» было бы нужно считать градиент по всему датасету, а по мини-батчам непонятно как).

Адаптивные методы градиентного спуска

- Но дальше улучшить всё равно можно.
- Заметим, что до сих пор скорость обучения была одна во всех направлениях.
- Идея: давайте быстрее двигаться по тем параметрам, которые не сильно меняются, и медленнее по быстро меняющимся параметрам.

Адаптивные методы градиентного спуска

- *Adagrad*: давайте накапливать историю этой скорости изменений и учитывать её.
- Обозначая $g_{t,i} = \nabla_{\theta_i} L(\theta)$, получим

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i},$$

  где G_t – диагональная матрица с $G_{t,ii} = G_{t-1,ii} + g_{t,i}^2$, которая накапливает общее значение градиента по всей истории обучения.
- Так что скорость обучения всё время уменьшается, но с разной скоростью для разных θ_i.

Адаптивные методы градиентного спуска

- Проблема: G всё увеличивается и увеличивается, и скорость обучения иногда уменьшается слишком быстро.
- *Adadelta* (Zeiler, 2012) – та же идея, но две новых модификации.
- Во-первых, историю градиентов мы теперь считаем с затуханием:

$$G_{t,ii} = \rho G_{t-1,ii} + (1-\rho) g_{t,i}^2.$$

- А всё остальное здесь точно так же:

$$u_t = -\frac{\eta}{\sqrt{G_{t-1} + \epsilon}} \mathbf{g}_{t-1}.$$

Адаптивные методы градиентного спуска

- Во-вторых, надо бы «единицы измерения» привести в соответствие.
- В предыдущих методах была проблема:
  - в обычном градиентном спуске или методе моментов «единицы измерения» обновления параметров $\Delta\theta$ — это единицы измерения градиента, т.е. если веса в секундах, а целевая функция в метрах, то градиент будет иметь размерность «метр в секунду», и мы вычитаем метры в секунду из секунд;
  - а в Adagrad получалось, что значения обновлений $\Delta\theta$ зависели от отношений градиентов, и величина обновлений вовсе безразмерная.

Адаптивные методы градиентного спуска

- Эта проблема решается в методе второго порядка: обновление параметров $\Delta\theta$ пропорционально $H^{-1}\nabla_\theta f$, то есть размерность будет

$$\Delta\theta \propto H^{-1}\nabla_\theta f \propto \frac{\frac{\partial f}{\partial \theta}}{\frac{\partial^2 f}{\partial \theta^2}} \propto \text{ размерность } \theta.$$

- Чтобы привести *Adadelta* в соответствие, нужно домножить на ещё одно экспоненциальное среднее, но теперь уже от квадратов обновлений параметров, а не от градиента.
- Настоящее среднее мы не знаем, аппроксимируем предыдущими шагами:

$$\mathbb{E}\left[\Delta\theta^2\right]_t = \rho\mathbb{E}\left[\Delta\theta^2\right]_{t-1} + (1-\rho)\Delta\theta^2, \text{ где}$$

$$u_t = -\frac{\sqrt{\mathbb{E}\left[\Delta\theta^2\right]_{t-1} + \epsilon}}{\sqrt{G_{t-1} + \epsilon}} \cdot g_{t-1}.$$

Адаптивные методы градиентного спуска

- Следующий вариант – *RMSprop* из курса Хинтона.
- Практически то же, что *Adadelta*, только *RMSprop* не делает вторую поправку с изменением единиц и хранением истории самих обновлений, а просто использует корень из среднего от квадратов (вот он где, RMS) от градиентов:

$$u_t = -\frac{\eta}{\sqrt{G_{t-1} + \epsilon}} \cdot g_{t-1}.$$

Адаптивные методы градиентного спуска

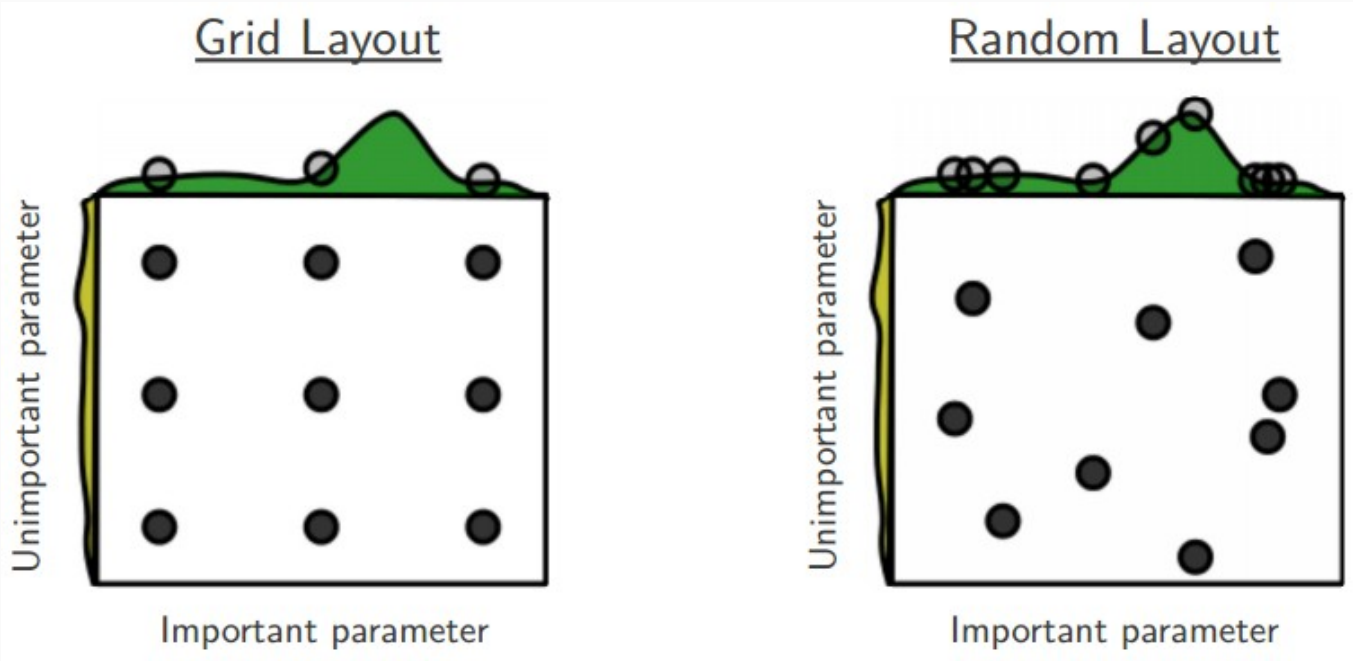
- И последний алгоритм – *Adam* (Kingma, Ba, 2014).
- Модификация *Adagrad* со сглаженными версиями среднего и среднеквадратичного градиентов:

$$
\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \\
u_t &= \frac{\eta}{\sqrt{v + \epsilon}} m_t.
\end{aligned}
$$

- (Kingma, Ba, 2014) рекомендуют $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.
- *Adam* практически не требует настройки, используется на практике очень часто.
- ...[анимированные примеры]...

Практические замечания

- Ещё практические замечания об оптимизации гиперпараметров:
  - одного валидационного множества достаточно;
  - лучше гиперпараметры искать на логарифмической шкале;
  - и лучше случайным поиском, а не по сетке (Bergstra and Bengio).



Спасибо!

Спасибо за внимание!