вторая лекция по вардропауту.

что происходит внутри нейросетей никто не знает. Это пугает.

на прошлой лекции: бернулли-дропаут. экв. гаусс-дроп. более удобный для анализа. Заметили, что функционал похож на первое слагаемое вар. вывода. ввели априорное распределение, что совпало. это дало возможность оптимизировать и по дисперсии.

$$p(t \mid x, w)p(w); p(w) \sim \prod_{i,j,l} \frac{1}{|w_{ij}^l|}$$

$$q(w \mid \mu, \alpha) = \prod_{i,j,l} \mathcal{N}(w_{ij}^l \mid \mu_{ik}, \alpha_l (\mu_{ij}^l)^2$$

$$-\mathrm{KL}(q(w \mid \mu, \alpha) \| p(w)) = -\sum_{i,j,k} \mathrm{KL}\left(q(w_{ij}^l \mid \mu_{ij}^l, \alpha_l \mu_{ij}^l) \| \frac{1}{|w_{ij}^l|}\right)$$

$$\approx \sum_{ijl} -\frac{1}{2} \log \frac{1+\alpha_l}{\alpha_l} + k_1 \sigma(k_2 + k_3 \log \alpha_l)) + \mathrm{Const}$$

рисунок с графиком приближения KL

а давайте на каждый вес ставить α_{ij}^l. мы не переобучаемся, тк просто подбираем лучшее апостериорное.

$$(\mu, \alpha) = \arg\max \left[\int q(w \mid \mu, \alpha) \log p(T \mid X, w) dw - \mathrm{KL}(q(w \mid \mu, \alpha) \| p(w))\right]$$
$$= \arg\max \mathcal{L}(\mu, \alpha)$$

Не работает. Стох. градиет по α – супер, по μ – плохо.

$$\frac{\partial}{\partial \mu_{ij}^l} \mathcal{L}(\mu, \alpha) = \frac{\partial}{\partial \mu_{ij}^l} \int q(w \mid \mu, \alpha) \log p(T \mid X, w) dW =$$

переход к минибатчу

$$\approx n \int r(\varepsilon) \frac{\partial}{\partial m u_{ij}^l} \log p(t_k \mid x_k, w(\varepsilon, \alpha, \varepsilon)) d\varepsilon$$

$$[w_{ij}^l = m u_{ij}^l + \sqrt{\alpha_{ij}^l} \mu_{ij}^l \varepsilon]$$

$$\approx n \frac{\partial}{\partial \mu_{ij}^l} \log p(t_k \mid x_k, w(\mu, \alpha, \hat{\varepsilon}))$$

$$[\hat{\varepsilon} \sim \mathcal{N}(\varepsilon, 0, I)]$$

$$n \frac{\partial}{\partial w_{ij}^l} \log p(t_k \mid x_k, w) \left(1 + \sqrt{\alpha_{ij}^l} \hat{\varepsilon}\right)$$

Мы расчитываем, что α_{ij}^l устремится к бесконечности. Но тогда у нас у второго множителя дисперсия взрывается, т.е. скорость сходимости нулевая.

Можно исправить?

На помощь приходит элегантная идея.

Давайте сменим параметризацию на эквивалентную, но в которой хороший градиент.

$$\mu_{ij}^l, \alpha_{ij}^l$$
$$\sigma_{ij}^l$$

Избавимся от α :

$$(\sigma_{ij}^l)^2 = \alpha_{ij}^l (\mu_{ij}^l)^2$$
$$\alpha_{ij}^l = \left(\frac{\sigma_{ij}^l}{\mu_{ij}^l}\right)^2$$

Теперь все супер, когда мы переходим от параметризации α_{ij}^l к σ_{ij}^l :

$$w_{ij}^l = \mu_{ij}^l + \sqrt{\alpha_{ij}^l}\mu_{ij}^l \varepsilon = \mu_{ij}^l + \sigma_{ij}^l \varepsilon$$
$$\frac{\partial w_{ij}^l}{\partial \mu_{ij}^l} = 1$$

Когда α^l один на слой, то в целом норм. Но когда на каждый вес – надо биться за дисперсию стох. града.

Как себя ведет дисперсия градиента:

$$\frac{\partial}{\partial \mu_{ij}^l} \mathcal{L}(\mu, \sigma) \approx n \frac{\partial}{\partial w_{ij}^l} \log p(t_k \mid x_k, w) \cdot 1$$

Не хватает KL. В новой параметризации появляется зависимость KL от него

$$\frac{\partial}{\partial \mu_{ij}^l} \mathcal{L}(\mu, \sigma) \approx n \frac{\partial}{\partial w_{ij}^l} \log p(t_k \mid x_k, w) \cdot 1 - \frac{\partial}{\partial \alpha_{ij}^l} \mathrm{KL}\left(q(w \mid \mu, \sigma) \| p(w)\right) \frac{\partial \alpha_{ij}^l}{\partial \mu_{ij}^l}$$

Математически эквивалентно, численно – нет. От выбора свободных переменных все сильно зависит.

Почти все веса исчезают. Схлопывается в дельта-функцию в нуле.

Все было хорошо, пока не пришли теоретики. Они написали статью "Variational Dropout is not Bayesian".

$\prod_{ijl} \frac{1}{|w_{ij}^l|}$ — несобственное. У него бесконечная масса в нуле. Честное апостериорное распределение должно быть тоже несобственным, что выгораживет исходную идею.

Любое несобственное распределение можно рассматривать как предел собственного

$$\lim_{a,b\to 0} \mathcal{G}(x \mid a, b)$$

Картинка

Но р-ие все равно близко к соб.......... Короче, нечем такую критику крыть

Прошло несколько месяцев, пока не поняли, как перевести это на байесовские рельсы. Оказывается, это не очень сложно.

$$-\mathrm{KL}(q(w \mid \mu, \alpha) \| \frac{1}{|w_{ij}^l|}) = \mathrm{Const} - \frac{1}{2} \log \frac{1+\alpha}{\alpha} + k_1 \sigma(k_2 + k_3 \log \alpha)$$

Давайте перейдем к RVM-подобному распределению:

$$p(w) = \prod_{ijl} \mathcal{N}\left(w_{ij}^l \mid 0, (\lambda_{ij}^l)^2\right)$$

$$\mathrm{KL}\left(\mathcal{N}\left(x \mid \mu, \sigma^2\right) \| \mathcal{N}\left(x \mid 0, \lambda^2\right)\right) = -\log \frac{\sigma}{\lambda} + \frac{\sigma^2 + \mu^2}{2\lambda^2}$$

$$\left[\mathrm{KL}\left(\mathcal{N}\left(w \mid \mu, \alpha\mu^2\right) \| \mathcal{N}(w|0, \lambda^2)\right)\right] = -\frac{1}{2} \log \frac{\alpha\mu^2}{\lambda^2} + \frac{\mu^2 + \alpha\mu^2}{2\lambda^2} \Bigg| \frac{\partial}{\partial \lambda}, = 0$$

$$\left[\frac{1}{\lambda} = \frac{\mu^2 + \alpha\mu^2}{\lambda^3}; \quad \lambda^2 = \mu^2(1+\alpha);\right]$$

$$= -\frac{1}{2} \log \frac{\alpha\mu^2}{(1+\alpha)\mu^2} + \frac{\mu^2(1+\alpha)}{2(1+\alpha)\mu^2} = \mathrm{Const} - \frac{1}{2} \log \frac{\alpha}{1+\alpha}$$

$$-\mathrm{KL}(\mathcal{N}\left(w \mid \mu, \alpha\mu^2\right) \| \mathcal{N}\left(w \mid 0, \lambda^2\right)) = \mathrm{Const} - \frac{1}{2} \log \frac{1+\alpha}{\alpha}$$

Отличается только на k_1, k_2, k_3, а они ни на что не влияют. Т.е. у нас не несобств., а вполне нормальная модель.

$$\log p(T \mid X, \alpha) \to \max_{\lambda}$$

$$\log p(T \mid X, \alpha) \geq \mathcal{L}(\mu, \alpha, \lambda) =$$

$$\int q(w \mid \mu, \alpha) \log p(T \mid X, w) dw - \mathrm{KL}(q(w \mid \mu, \alpha) \| p(w \mid \lambda)$$

Эмпирика - гаусс - вар вывод (несобственный) - вар вывод (собственный). Как-то так мы шли, и каждый раз нам открывалось больше понимания, возможностей. Как только мы перешли к байесу, мы получаем множество возможностей. Например, выбирать автоматически магнитуды. Даже индивидуально на каждый вес, на каждый нейрон.

Байес — это не панацея, тк все равно мы можем переобучиться эмпирическим байесом.

До сих пор ничего неожиданного не было.

Теперь можно шарить дисперсии на нейроны: λ_j^l. Так можно выкидывать целые нейроны, а это можно полезно использовать для реального ускорения.

Давайте оставлю индекс i

$$\sum_i \mathrm{KL}\left(\mathcal{N}\left(w_i, \alpha\mu_i^2\right) \| \mathcal{N}\left(w_i | 0, \lambda^2\right)\right) =$$

$$\sum_{i=1}^m \left(-\frac{1}{2}\log\frac{\alpha\mu_i^2}{\lambda^2} + \frac{\mu_i^2 + \alpha\mu_i^2}{2\lambda^2}\right)\Bigg| \frac{\partial}{\partial\lambda}, = 0$$

$$\sum_{i=1}^m \frac{1}{\lambda} - \frac{\mu_i^2(1+\alpha)}{\alpha^3} = 0$$

$$\lambda^2 m = \sum_{i=1}^m \mu_i^2(1+\alpha) = (1+\alpha)\sum_{i=1}^m \mu_i^2$$

$$\lambda^2 = \frac{1+\alpha}{m}\sum_{i=1}^m \mu_i^2$$

$$\sum_{i=1}^m \left[-\frac{1}{2}\log\frac{\alpha\mu_i^2 m}{(1+\alpha)\sum_k \mu_k^2} + \frac{\mu_i^2(1+\alpha)m}{2(1+\alpha)\sum_k \mu_k^2}\right] =$$

$$\frac{m}{2} - \sum_{i=1}^m \frac{1}{2}\log\mu_i^2 - \frac{m}{2}\log\frac{\alpha}{1+\alpha} - \frac{m}{2}\log m + \frac{m}{2}\log\sum_k \mu_k^2$$

То есть, то же можно получить оптимальное значение, запихнуть в оценку, и оптимизировать по старым приятным параметрам групповым разреживанием.

Вместе с групповостью мы уменьшили число гиперпараметров. Теперь количество параметров соизмеримо с числом объектов.

Это тоже не оч. круто.

А сейчас будет вау и парадоксально. Вернемся к исходной модели. Пусть α одна на слой снова.

$$q(w \mid \mu, \alpha) = \prod_{ijl} \mathcal{N}(w_{ij}^l \mid \mu_{ij}^l, \alpha^l(\mu_{ij}^l)^2)$$

Оптимайзим. Получаем, что есть 1-2 слоя, где α^l равен ИНФИНИТИ. То есть как? Можно вообще без нейросети?

Делаем отладочный эксперимент. Мониторим три вещи: качество на моде распределения, качество на сампле распределения, качество на ансамбле (10-20 samплов достаточно).

картинка про variance networks. when expectations don't meet your expectations

Возникает нормальное с нулевым мат. ожиданием и ненулевой дисперсией.

Думали, что ошибка. Делали много отладочных экспериментов. Один из них:

$$q(w) = \prod_{ijl} \mathcal{N}\left(w_{ij}^l \mid 0, \left(\sigma_{ij}^l\right)^2\right)$$

Работает. Отношение SNR — ужос, и получается, что сети умеют хранить информацию в дисперсиях. Конструкция работать не должна, а она обучается до соты.

И еще один эффект, который был обнаружен. NIPS 2018. Mode connectivity.

Берем нейросеть, обучаем два раза из двух разных инициализации. Получили w_1, w_2.

Что будет, если построить линию между ними. Пока фигня.

А может есть все же путь?

Будем оптимизировать такую штуку:

$$r(\varepsilon) \sim U(0,2);\ \int_0^2 r(\varepsilon) \log p(T \mid X, w(\varepsilon)) d\varepsilon \to \max_{w_0}$$

$$w(\varepsilon) = \begin{cases} (1-\varepsilon)w_1 + \varepsilon w_0, & \varepsilon < 1 \\ (2-\varepsilon)w_0 + (\varepsilon - 1)w_2, & \varepsilon > 1 \end{cases}$$

Такой путь всегда находится. WTF.

А что если переставлять нейроны местами? Вроде, не должно быть. Фиг там, есть опять какой-то путь.

Если мы нашли путь тупой ломаной, то значит их тонна. Есть гипотеза: множество глобальных минимумов образует единое связное многообразие какой-то размерности. Какая размерность? Не знаем, тк нет инструментов.