$$x = z_0 \to^{w_1} h_1 \to^{f} z_2 \to^{w_2} h_2 \to \cdots \to z_T = y$$

$$h^{t+1} = w^{t+1} z_t; z_t = f(h_t)$$

$$\xi_{ij}^t \sim \text{Ber}(\xi_j^t | p_t); 1 - p_t \text{ – dropout rate}$$

$$h_{t_i} = \sum_{j=1} w_{ij}^t z_j^{t-1} \xi_{ij}^t$$

$$\int q(\xi|p) \log p(T|X, w, \xi) d\xi \to \max_w$$

$$Eh_{ti} = p_t \sum_j w_{ij}^t z_j^t$$

$$Dh_{ti} = p_t(1 - p_t) \sum_j (w_{ij}^t z_j^t)^2$$

$$D(\alpha X + \beta Y) = p_t(1 - p_t)(\alpha^2 + \beta^2)$$

h_{ti} приближенно нормальное:

$$h_{ti} \approx N(h_{ti} | \mu_{ti}, \sigma_{ti}^2)$$

$$Eh_{ti} = p_t \sum_j w_{ij}^t z_j^t = \mu_{ti}$$

$$Dh_{ti} = p_t(1 - p_t) \sum_j (w_{ij}^t z_j^t)^2 = \sigma_{ti}^2$$

$$\xi_{ti}^t \sim N(\xi_{ij}^t | p_t, p_t(1 - p_t))$$

давайте объединять $w_{ij}^t \xi_{ij}^t$ с весами.

$$w_{ij}^t \sim N(w_{ij}^t | \mu_{ij}^t, (\sigma_{ij}^t)^2)$$

$$\mu_{ij}^t = p_t w_{ij}^t$$

$$\sigma_{ij}^t = p_t(1 - p_t)(w_{ij}^t)^2 = \frac{1 - p_t}{p_t}(\mu_{ij}^t)^2 = \alpha_t(\mu_{ij}^t)^2$$

$$\hat{w}_{ij}^t \sim N\left(\hat{w}_{ij}^t | \mu_{ij}^t, (\sigma_{ij}^t)^2\right) = N\left(\hat{w}_{ij}^t | \mu_{ij}^t, \alpha_t(\mu_{ij}^t)^2\right)$$

Во что превратился наш функционал. У нас становится мат. ожидание по этим весам.

$$\int q(\xi|p) \log p(T|X, w, \xi) d\xi \to \max_w$$

$$\int q(w|\mu, \alpha) \log p(T|X, w) dw \to \max_\mu; \alpha_t = \frac{1 - p_t}{p_t}$$

$\alpha_t = 0$ – шума много, $\alpha_t = 1$ – шума нет

$$q(w|\mu, \alpha) = \prod_{i,j,t} N(w_{ij}^t|\mu_{ij}^t, \alpha_t(\mu_{ij}^t)^2$$

Если мы оптимизируем по α, то это бессмысленно, можно сразу ставить ml. Впрыск шума ухудшает loss, поэтому оптимизация приведет к $\alpha = 0$
Это по прежнему не объясняет, почему dropout работает. Люди много лет ломали голову над объяснением того, почему dropout работает, пока не пришли байезиане. Max Welling.
Напоминает байес, но нет kl.
Как бы мы действовали?

$$p(T|x, w)p(w)$$
$$p(w|X, T) \approx q(w|\phi) = \arg\min_{\phi} KL\left(q(w|\phi)\|p(w|X, T)\right) =$$
$$= \arg\max_{\phi} \left[\int q(w|\phi) \log p(T|X, w)dw - KL\left(q(w|\phi)\|p(w)\right)\right]$$
$$q(w|\mu, \alpha) = \prod_{i,j,t} N(w_{ij}^t|\mu_{ij}^t; \alpha_t(\mu_{ij}^t)^2$$

Если гаусс-дропаут ответ — то какой вопрос? А что если удастся ввести априорное на w так, что KL не зависит от μ, только от α. Тогда опт. станет эквивалентной оптимизации гаусс-дропаута.
Это существует: лог-юниформ.

$$p(w) \propto \prod_{i,j,t} \frac{1}{w_{i,j}^t}$$

несобственное, но это не плохо. (типичный пример $\lim_{\sigma\to\infty} N(x|0, \sigma) = U(R)$)

$$\lim_{a,b\to 0} G(x|a, b) = p(x) \propto \frac{1}{x}; x > 0 \text{ – log-uniform}$$
$$p(\log x) = U(R)$$

$$P\{x \in [3.14, 3.15]\} = P\{x \in [1020, 1030]\} = \ldots$$

Если у нас лог-ю, то все эти вероятности одинаковые. Отсутствие информации по масштабу, только на число значащих цифр.
Фактически мы ввели штраф на число значащих цифр. Не настраиваем

веса слишком точно. Как выглядит KL:

$$KL(q(w|\mu,\alpha)\|p(w)) = \sum KL(N(w_{ij}^t|\mu_{ij}^t,\alpha_t(\mu_{ij}^t)^2)\|\frac{1}{|w_{ij}^t|}) + \text{Const} =$$

$$\sum_{i,j,t}\int N(w|\mu,\alpha(\mu)^2)\log N(w|\mu,\alpha(\mu)^2)dw + \int N(w|\mu,\alpha_t(\mu)^2)\log|w|dw$$

$$= \sum_{i,j,t} -\log\sqrt{2\pi\dots} - \frac{1}{2}\log\alpha - \log|\mu| + \int N(\varepsilon|0,1)log|\mu + \sqrt{\alpha(\mu)^2}|d\varepsilon =$$

$$\sum_{i,j,t} -\log\sqrt{2\pi\dots} - \frac{1}{2}\log\alpha - \log|\mu| + \int N(\varepsilon|0,1)\log|\mu||1 + \sqrt{\alpha_t}\varepsilon|d\varepsilon$$

$$\sum_{i,j,t} -\log\sqrt{2\pi e} - \log|\mu| + \log|\mu| + \int N(\varepsilon|0,1)\log|1 + \sqrt{\alpha}\varepsilon d\varepsilon$$

$$= f(\alpha)$$

Вывод: гаусс (и обычный) дропаут — байесовская процедура с хитрым априорным распределением. Группа Гала более коряво и гаусс притянула к байесовым рельсам, но объяснять не буду.

"Вар. Дропаут не Байесовский— исправляемо, но неприятно. На следующей неделе будет.

Хорошо. И что? А дальше что делать полезного? Ок, мы делаем б. ансамблирование, понятно, почему улучшается качество.

Можно настраивать α. Во. Без KL это было бессмысленно, теперь все круто.

Мы обощаем вариационное приближение. Мы не переобучаемся, тк улучшаем апостериорное. Опт. по α не только возможна, но и необходима — получим более точное приближение. Об этом тоже было в статье Веллинга. Смотрите, мы показали, что ВД это байес, поэтому можно оптимизировать. Но там нестабильности и т.д., поэтому мы ограничили оптимизацию α от 0 до 1 (иначе все плохо). На следующей лекции эта проблема сниматся.

Почему α одна на слой? Давайте на каждый вес. Тогда все еще точнее. Начинается разреживание. Остается очень-очень мало весов. Почти все веса исчезают (99.9%).

Введение избыточных параметров облегчает оптимизацию. За это платим гигантскими сетями. Оказывается, вот один способ устранения этой избыточности.

Бернулли=Гаусс=ВарВывод. Это все работает. Еще достоинство. Хорошо. Как модифицировать? Как только мы получили, что это байес, то теперь все понятно. Н-р, меняем вар. семейство, априорное и т.д. Можно сравнить с бустингом. Сначала были инженеры, потом пришли ученые и сказали, что это такой-то функционал. У него можно менять а-б-в-г-..., и получить что-то более крутое, чем было до этого.

Предположим мы решаем

$$\arg\max_{\phi}\left[\int q(w|\phi)\log p(T|X,w)dw - KL\left(q(w|\phi)\|p(w)\right)\right]$$

Проблемы — с первым слагаемым.

$$\sum_{k=1}^{n} \int N(w|\mu, \alpha) \log p(t_k|x_k, w) dw$$

минибатчи? ок. а веса? dsvi? неа. нужен сампл из 100000000-мерного распределения. можно проще? да — локальная репараметризация.

Выход зависит не напрямую, а через ашки.

$$h_{ti}|h_{t(i-1)} = \sum_{j=1}^{m} w_{ij}^t z_j^{t-1}$$

$$h_{ti}|h_{t(i-1)} \sim N(h_{ti}| \sum_{j=1}^{m} \mu_{ij}^t z_j^{t-1}, \alpha_t \sum_{j=1}^{m} \mu_{ij}^t z_j^{t-1}$$

$$\sum_{k=1}^{n} \int N(w|\mu, \alpha) \log p(t_k|x_k, w) dw = \sum_{k=1}^{n} \int r(h|w, x_k) \log p(t_k|x_k, w) dh =$$

$$\left[r(h|w, x_k) = \prod_{t=1}^{T} r(h_t|h^{t-1}, w) \right] =$$

$$\left[\prod_{t=1}^{n} \prod_{i=1}^{m} N \left(h_{ti} \middle| \sum_j \mu_{ij}^t z_j^{t-1}, \alpha_t \sum_j \left(\mu_j^t z_j^{t-1} \right)^2 \right) \right] =$$

$$\sum_{k=1}^{n} \int N(\varepsilon|0, I) \log p\left(t_k|x_k, h(\varepsilon, x_k, \mu, \alpha)\right) dh$$

LRT. Еще уменьшается дисперсия. С одной стороны — численная хитрость. С другой – мы получили более эффективную процедуру. Можно пользоваться тем, что размерность h намного меньше размерности весов.

Зашумление — новый способ регуляризации. Где еще было? Аугментации. Сам SGD. Показано, что без SGD мы быстро переобучаемся. Аналогия со слепым человеком. Если дисперсия высокая, то мы детали не видим; только широкие холмы, узкие не видим. Интуиция — это правильно, нам нужно искать широкие холмы. Все такие зашумления нам помогают в этом. Если мы уменьшаем дисперсию стох. градиента, то мы получаем сильное переобучение. Мы этого не хотим, только если мы не байесиане. А если мы оптимизируем ELBO, то нам это и нужно — мы становимся ближе к честному апостериорному.