# Вариационный вывод в моделях глубокого обучения

Бахтеев Олег

June 16, 2020

# Сложность модели: зачем?

| Model | image size | # parameters | Mult-Adds | Top 1 Acc. (%) | Top 5 Acc. (%) |
|---|---|---|---|---|---|
| Inception V2 [29] | 224×224 | 11.2 M | 1.94 B | 74.8 | 92.2 |
| **NASNet-A (5 @ 1538)** | **299×299** | **10.9 M** | **2.35 B** | **78.6** | **94.2** |
| Inception V3 [59] | 299×299 | 23.8 M | 5.72 B | 78.0 | 93.9 |
| Xception [9] | 299×299 | 22.8 M | 8.38 B | 79.0 | 94.5 |
| Inception ResNet V2 [57] | 299×299 | 55.8 M | 13.2 B | 80.4 | 95.3 |
| **NASNet-A (7 @ 1920)** | **299×299** | **22.6 M** | **4.93 B** | **80.8** | **95.3** |
| ResNeXt-101 (64 x 4d) [67] | 320×320 | 83.6 M | 31.5 B | 80.9 | 95.6 |
| PolyNet [68] | 331×331 | 92 M | 34.7 B | 81.3 | 95.8 |
| DPN-131 [8] | 320×320 | 79.5 M | 32.0 B | 81.5 | 95.8 |
| **SENet [25]** | **320×320** | **145.8 M** | **42.3 B** | **82.7** | **96.2** |
| **NASNet-A (6 @ 4032)** | **331×331** | **88.9 M** | **23.8 B** | **82.7** | **96.2** |

Zoph et al., 2017. Сложность моделей отличается почти в два раза при одинаковом качестве.

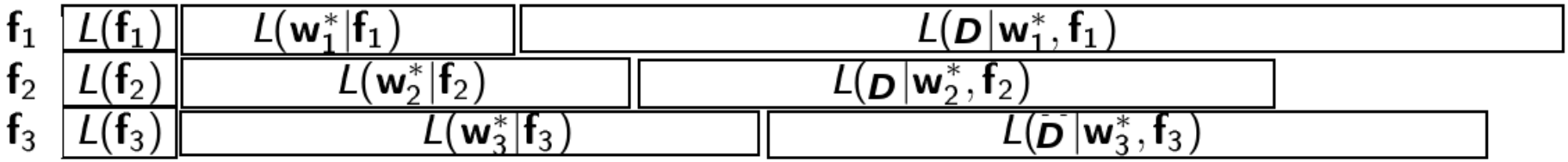
# Принцип минимальной длины описания

$$\mathrm{MDL}(\mathbf{f}, \mathfrak{D}) = L(\mathbf{f}) + L(\mathfrak{D}|\mathbf{f}),$$

где $\mathbf{f}$ — модель, $\mathfrak{D}$ — выборка, $L$ — длина описания в битах.

$$\mathrm{MDL}(\mathbf{f}, \mathfrak{D}) \sim L(\mathbf{f}) + L(\mathbf{w}^*|\mathbf{f}) + L(\mathfrak{D}|\mathbf{w}^*, \mathbf{f}),$$

$\mathbf{w}^*$ — оптимальные параметры модели.



# MDL и Колмогоровская сложность

**Колмогоровская сложность** — длина минимального кода для выборки на предварительно заданном языке.

**Теорема инвариантности**

Для двух сводимых по Тьюрингу языков колмогоровская сложность отличается не более чем на константу, не зависящую от мощности выборки.

**Отличия от MDL**:

- Колмогоровская сложность невычислима.
- Длина кода может зависеть от выбранного языка. Для небольших выборок теорема инвариантности не дает адекватных результатов.

# Связанный байесовский вывод

*Первый уровень:* выбираем оптимальные параметры:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

*Второй уровень:* выбираем модель, доставляющую максимум обоснованности модели.

Обоснованность модели ("Evidence"):

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

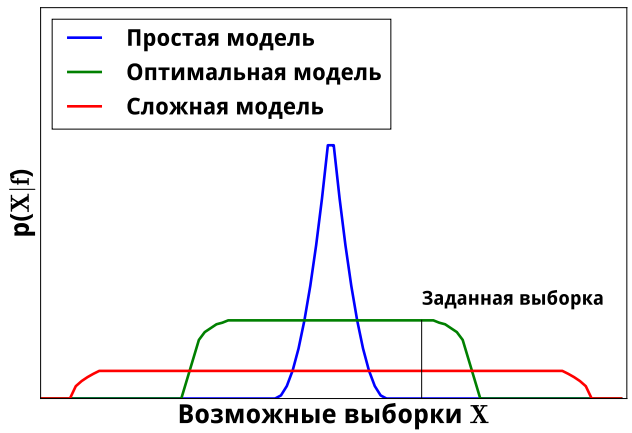
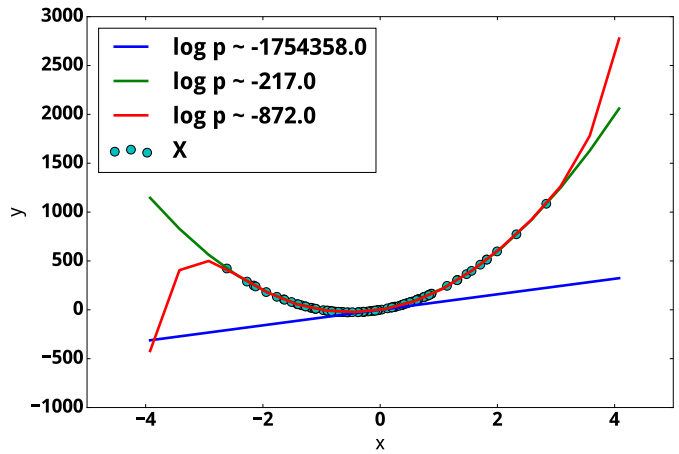
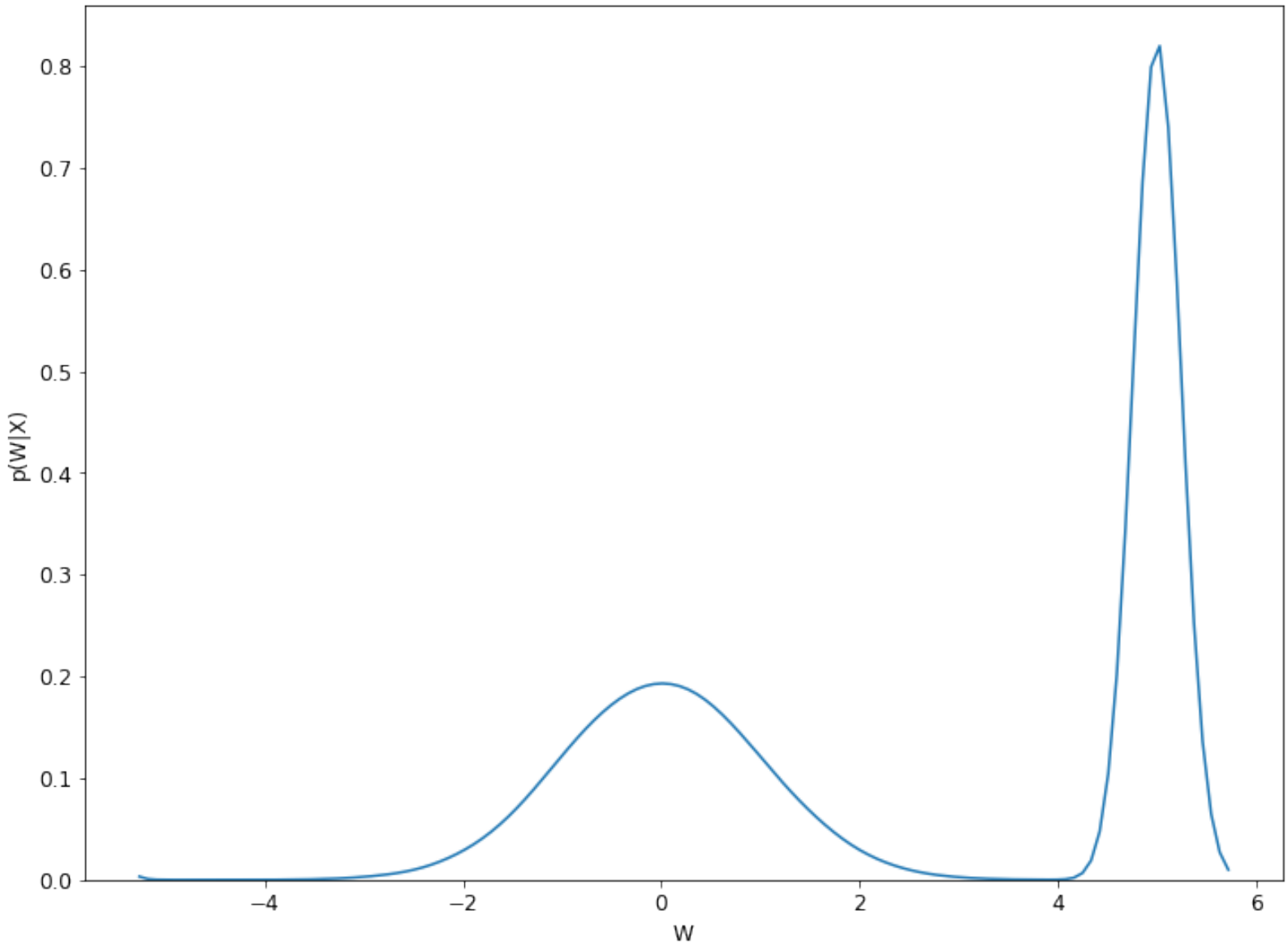



Схема выбора модели




Пример: полиномы

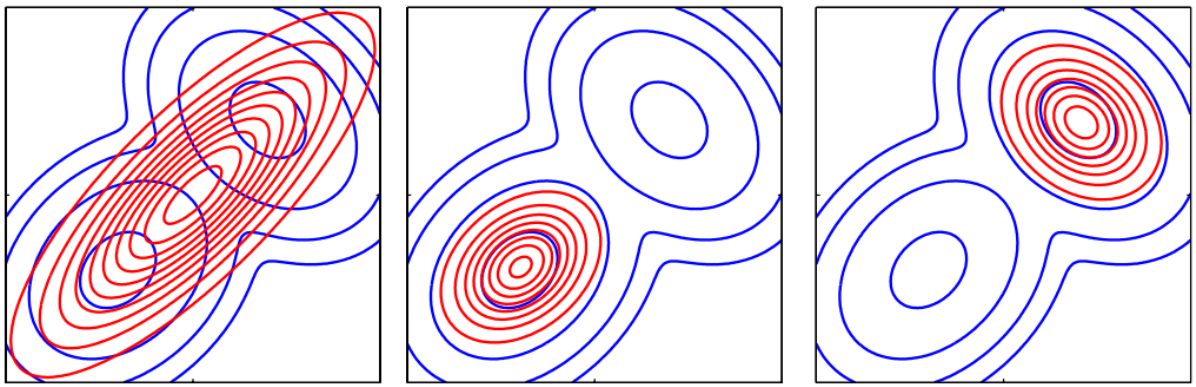
# MAP и Evidence



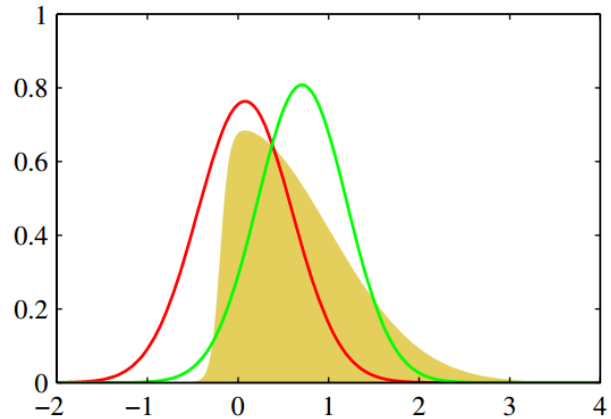
# Вариационная оценка, ELBO

**Вариационная оценка Evidence**, Evidence lower bound — метод нахождения приближенного значения аналитически невычислимого распределения $p(\mathbf{w}|\mathfrak{D},\mathbf{h})$ распределением $q(\mathbf{w}) \in \mathfrak{Q}$. Получение вариационной нижней оценки обычно сводится к задаче минимизации

$$\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D})) = -\int_{\mathbf{w}} q(\mathbf{w})\log\frac{p(\mathbf{w}|\mathfrak{D})}{q(\mathbf{w})}d\mathbf{w} = \mathrm{E}_{\mathbf{w}}\log p(\mathfrak{D}|\mathbf{w}) - \mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h}))$$



Вариационный вывод и expectation propogation (Bishop)




Аппроксимация Лапласа и вариационная оценка, зеленая линия (Bishop)

# Получение вариацонной нижней оценки

Максимизация вариационной нижней оценки

$$\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h})}{q(\mathbf{w})} d\mathbf{w}$$

эквивалентна минимизации расстояния Кульбака–Лейблера между распределением $q(\mathbf{w}) \in \mathfrak{Q}$ и апостериорным распределением параметров $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})$:

$$\hat{q} = \arg\max_{q \in \mathfrak{Q}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h})}{q(\mathbf{w})} d\mathbf{w} \Leftrightarrow \hat{q} = \arg\min_{q \in \mathfrak{Q}} \mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})\big),$$

$$\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})\big) = \int_{\mathbf{w}} q(\mathbf{w}) \log \left( \frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})} \right) d\mathbf{w}.$$

# ELBO: нормальное распределение

Пусть $q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q)$.
Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{h}) d\mathbf{w} - D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathbf{y}_i|\mathbf{x}_i, \hat{\mathbf{w}}) - D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) \to \max_{\mathbf{A_q}, \boldsymbol{\mu_q}}, \quad \hat{\mathbf{w}} \sim q.$$

В случае, если априорное распределение параметров $p(\mathbf{w}|\mathbf{h})$ является нормальным:

$$p(\mathbf{w}|\mathbf{h}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}),$$

дивергенция $D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})$ вычисляется аналитически:

$$D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) = \frac{1}{2}\big(\mathrm{tr}(\mathbf{A}^{-1}\mathbf{A}_q) + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathsf{T}}\mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - n + \ln|\mathbf{A}| - \ln|\mathbf{A}_q|\big).$$

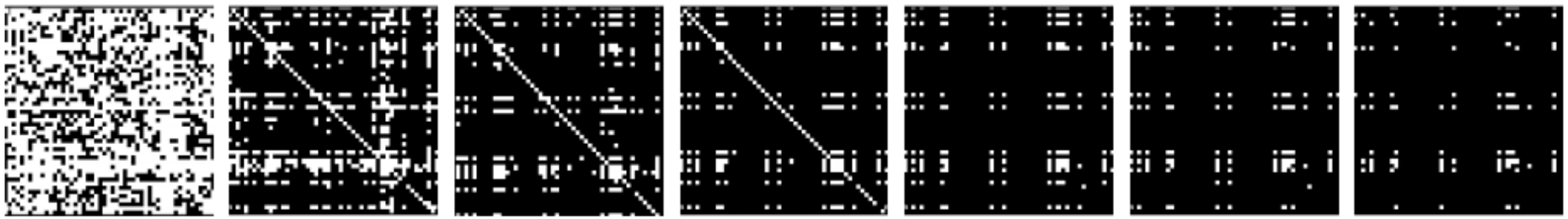
# Graves, 2011

**Априорное распределение:** $p(\mathbf{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma\mathbf{I})$.
**Вариационное распределение:** $q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q\mathbf{I})$.
Жадная оптимизация гиперпараметров:

$$\boldsymbol{\mu} = \hat{\mathsf{E}}\mathbf{w}, \quad \sigma = \hat{\mathsf{D}}\mathbf{w}.$$

Прунинг параметра $w_i$ определяется относительной плотностью:

$$\lambda = \frac{q(\mathbf{0})}{q(\boldsymbol{\mu}_{i,q})} = \exp(-\frac{\mu_i^2}{2\sigma_i^2}).$$



# ELBO: нормальное распределение

**“Обычная” функция потерь:**

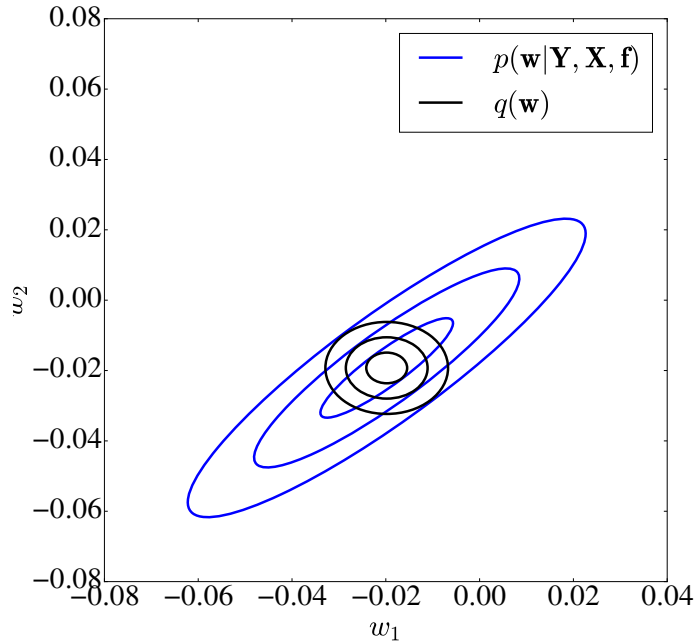
$$L = \sum_{\mathbf{x},\mathbf{y}\in\mathfrak{D}} -\log p(\mathbf{y}|\mathbf{x},\mathbf{w}) + \lambda||\mathbf{w}||_2^2.$$

Пример грубой аппроксимации нормальным диагональным распределением $q$

**Вариационный вывод при** $(p(\mathbf{w}|\mathbf{h}) \sim \mathcal{N}(\mathbf{0},\mathbf{1}))$**:**

$$L = \sum_{\mathbf{x},\mathbf{y}} \log\, p(\mathbf{y}|\mathbf{x},\hat{\mathbf{w}}) +$$

$$+\frac{1}{2}\left(\mathrm{tr}(\mathbf{A}_q) + \boldsymbol{\mu}_q^\mathsf{T}\mathbf{A}^{-1}\boldsymbol{\mu}_q - \ln\,|\mathbf{A}_q|\right).$$



# Прунинг: вероятностные предположения

Consider the following prior over a parameter $w$ where its scale $z$ is governed by a distribution $p(z)$:

$$z \sim p(z); \qquad w \sim \mathcal{N}(w; 0, z^2), \tag{3}$$

with $z^2$ serving as the variance of the zero-mean normal distribution over $w$. By treating the scales of $w$ as random variables we can recover marginal prior distributions over the parameters that have heavier tails and more mass at zero; this subsequently biases the posterior distribution over $w$ to be sparse. This family of distributions is known as scale-mixtures of normals [6, 2] and it is quite general, as a lot of well known sparsity inducing distributions are special cases.

One example of the aforementioned framework is the spike-and-slab distribution [50], the golden standard for sparse Bayesian inference. Under the spike-and-slab, the mixing density of the scales is a Bernoulli distribution, thus the marginal $p(w)$ has a delta "spike" at zero and a continuous "slab" over the real line. Unfortunately, this prior leads to a computationally expensive inference since we have to explore a space of $2^M$ models, where $M$ is the number of the model parameters. Dropout [29, 67],

# Bayesian deep compression: drpout-group sparsity

One potential choice for $p(z)$ is the improper log-uniform prior [37]: $p(z) \propto |z|^{-1}$. It turns out that we can recover the log-uniform prior over the weights $w$ if we marginalize over the scales $z$:

$$p(w) \propto \int \frac{1}{|z|} \mathcal{N}(w|0, z^2) dz = \frac{1}{|w|}. \tag{4}$$

This alternative parametrization of the log uniform prior is known in the statistics literature as the normal-Jeffreys prior and has been introduced by [16]. This formulation allows to "couple" the scales of weights that belong to the same group (e.g. neuron or feature map), by simply sharing the corresponding scale variable $z$ in the joint prior[3]:

$$p(\mathbf{W}, \mathbf{z}) \propto \prod_{i}^{A} \frac{1}{|z_i|} \prod_{ij}^{A,B} \mathcal{N}(w_{ij}|0, z_i^2), \tag{5}$$

where $\mathbf{W}$ is the weight matrix of a fully connected neural network layer with $A$ being the dimensionality of the input and $B$ the dimensionality of the output. Now consider performing variational inference with a joint approximate posterior parametrized as follows:

$$q_\phi(\mathbf{W}, \mathbf{z}) = \prod_{i=1}^{A} \mathcal{N}(z_i|\mu_{z_i}, \mu_{z_i}^2 \alpha_i) \prod_{i,j}^{A,B} \mathcal{N}(w_{ij}|z_i \mu_{ij}, z_i^2 \sigma_{ij}^2), \tag{6}$$

where $\alpha_i$ is the dropout rate [67, 37, 51] of the given group. As explained at [37, 51], the multiplicative parametrization of the approximate posterior over $\mathbf{z}$ suffers from high variance gradients; therefore we will follow [51] and re-parametrize it in terms of $\sigma_{z_i}^2 = \mu_{z_i}^2 \alpha_i$, hence optimize w.r.t. $\sigma_{z_i}^2$. The lower bound under this prior and approximate posterior becomes:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\mathbf{z}) q_\phi(\mathbf{W}|\mathbf{z})}[\log p(\mathcal{D}|\mathbf{W})] - \mathbb{E}_{q_\phi(\mathbf{z})}[KL(q_\phi(\mathbf{W}|\mathbf{z})||p(\mathbf{W}|\mathbf{z}))] - KL(q_\phi(\mathbf{z})||p(\mathbf{z})). \tag{7}$$

# Bayesian deep compression: drpout-group sparsity

At test time, in order to have a single feedforward pass we replace the distribution over $\mathbf{W}$ at each layer with a single weight matrix, the masked variational posterior mean:

$$\hat{\mathbf{W}} = \operatorname{diag}(\mathbf{m})\,\mathbb{E}_{q(\mathbf{z})q(\tilde{\mathbf{W}})}[\operatorname{diag}(\mathbf{z})\tilde{\mathbf{W}}] = \operatorname{diag}\left(\mathbf{m} \odot \boldsymbol{\mu}_z\right)\mathbf{M}_W, \tag{10}$$

where $\mathbf{m}$ is a binary mask determined according to the group variational dropout rate and $\mathbf{M}_W$ are the means of $q_\phi(\tilde{\mathbf{W}})$. We further use the variational posterior marginal variances[5] for this particular posterior approximation:

$$\mathbb{V}(w_{ij})_{NJ} = \sigma_{z_i}^2\left(\sigma_{ij}^2 + \mu_{ij}^2\right) + \sigma_{ij}^2\mu_{z_i}^2, \tag{11}$$

# Bayesian deep compression: half-cauchy

Another choice for $p(z)$ is a proper half-Cauchy distribution: $\mathcal{C}^+(0, s) = 2(s\pi(1 + (z/s)^2))^{-1}$; it induces a horseshoe prior [8] distribution over the weights, which is a well known sparsity inducing prior in the statistics literature. More formally, the prior hierarchy over the weights is expressed as (in a non-centered parametrization):

$$s \sim \mathcal{C}^+(0, \tau_0); \qquad \tilde{z}_i \sim \mathcal{C}^+(0, 1); \qquad \tilde{w}_{ij} \sim \mathcal{N}(0, 1); \qquad w_{ij} = \tilde{w}_{ij}\tilde{z}_i s, \tag{12}$$

# Bayesian deep compression: half-cauchy

$$s_b \sim \mathcal{IG}(0.5, 1); \quad s_a \sim \mathcal{G}(0.5, \tau_0^2); \quad \tilde{\beta}_i \sim \mathcal{IG}(0.5, 1); \quad \tilde{\alpha}_i \sim \mathcal{G}(0.5, 1); \quad \tilde{w}_{ij} \sim \mathcal{N}(0, 1);$$

$$w_{ij} = \tilde{w}_{ij} \sqrt{s_a s_b \tilde{\alpha}_i \tilde{\beta}_i}. \tag{14}$$

It should be mentioned that the improper log-uniform prior is the limiting case of the horseshoe prior when the shapes of the (inverse) Gamma hyperpriors on $\tilde{\alpha}_i, \tilde{\beta}_i$ go to zero [8]. In fact, several well known shrinkage priors can be expressed in this form by altering the shapes of the (inverse) Gamma hyperpriors [3]. For the variational posterior we will employ the following mean field approximation:

$$q_\phi(s_b, s_a, \tilde{\boldsymbol{\beta}}) = \mathcal{LN}(s_b|\mu_{s_b}, \sigma^2_{s_b})\mathcal{LN}(s_a|\mu_{s_a}, \sigma^2_{s_a}) \prod_i^A \mathcal{LN}(\tilde{\beta}_i|\mu_{\tilde{\beta}_i}, \sigma^2_{\tilde{\beta}_i}) \tag{15}$$

$$q_\phi(\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{W}}) = \prod_i^A \mathcal{LN}(\tilde{\alpha}_i|\mu_{\tilde{\alpha}_i}, \sigma^2_{\tilde{\alpha}_i}) \prod_{i,j}^{A,B} \mathcal{N}(\tilde{w}_{ij}|\mu_{\tilde{w}_{ij}}, \sigma^2_{\tilde{w}_{ij}}), \tag{16}$$

# Bayesian deep compression: результаты

| Model Original Error % | Method | $\frac{\lvert w \neq 0 \rvert}{\lvert w \rvert}$ % | Compression Rates (Error %) Pruning | Fast Prediction | Maximum Compression |
|---|---|---|---|---|---|
| LeNet-300-100 | DC | 8.0 | 6 (1.6) | - | 40 (1.6) |
| | DNS | 1.8 | 28* (2.0) | - | - |
| 1.6 | SWS | 4.3 | 12* (1.9) | - | 64(1.9) |
| | Sparse VD | 2.2 | 21(1.8) | 84(1.8) | 113 (1.8) |
| | BC-GNJ | 10.8 | 9(1.8) | 36(1.8) | 58(1.8) |
| | BC-GHS | 10.6 | 9(1.8) | 23(1.9) | 59(2.0) |
| LeNet-5-Caffe | DC | 8.0 | 6*(0.7) | - | 39(0.7) |
| | DNS | 0.9 | 55*(0.9) | - | 108(0.9) |
| 0.9 | SWS | 0.5 | 100*(1.0) | - | 162(1.0) |
| | Sparse VD | 0.7 | 63(1.0) | 228(1.0) | 365(1.0) |
| | BC-GNJ | 0.9 | 108(1.0) | 361(1.0) | 573(1.0) |
| | BC-GHS | 0.6 | 156(1.0) | 419(1.0) | 771(1.0) |
| VGG | BC-GNJ | 6.7 | 14(8.6) | 56(8.8) | 95(8.6) |
| 8.4 | BC-GHS | 5.5 | 18(9.0) | 59(9.0) | 116(9.2) |

# Квантизация: вероятностные предположения

The log uniform prior (Eq. (5)) can be viewed as a continuous relaxation of the spike-and-slab prior with a spike at location 0 (Louizos et al., 2017). We use this insight to formulate a quantizing prior, a continuous relaxation of a "multi-spike-and-slab" prior which has multiple spikes at locations $c_k,\ k \in \{1, \ldots, K\}$. Each spike location corresponds to one target value for subsequent quantiza-

# Квантизация: вероятностные предположения

$$p(w_{ij}) = \int \mathcal{N}(w_{ij}|m_{ij}, z_{ij}) p_z(z_{ij}) p_m(m_{ij}) \, \mathrm{d}z_{ij} \mathrm{d}m_{ij} \qquad p_m(m_{ij}) = \sum_k a_k \delta(m_{ij} - c_k), \quad (11)$$

with $p(z_{ij}) \propto |z_{ij}|^{-1}$. The location prior $p_m(m_{ij})$ is a mixture of weighted delta distributions located at the quantization values $c_k$. Marginalizing over $m$ yields the quantizing prior

$$p(w_{ij}) \propto \sum_k a_k \int \frac{1}{|z_{ij}|} \mathcal{N}(w_{ij}|c_k, z_{ij}) \, \mathrm{d}z_{ij} = \sum_k a_k \frac{1}{|w_{ij} - c_k|}. \quad (12)$$

In our experiments, we use $K = 3$, $a_k = 1/K \ \ \forall k$ and $c_2 = 0$ unless indicated otherwise.

# Квантизация: назначение значений параметров

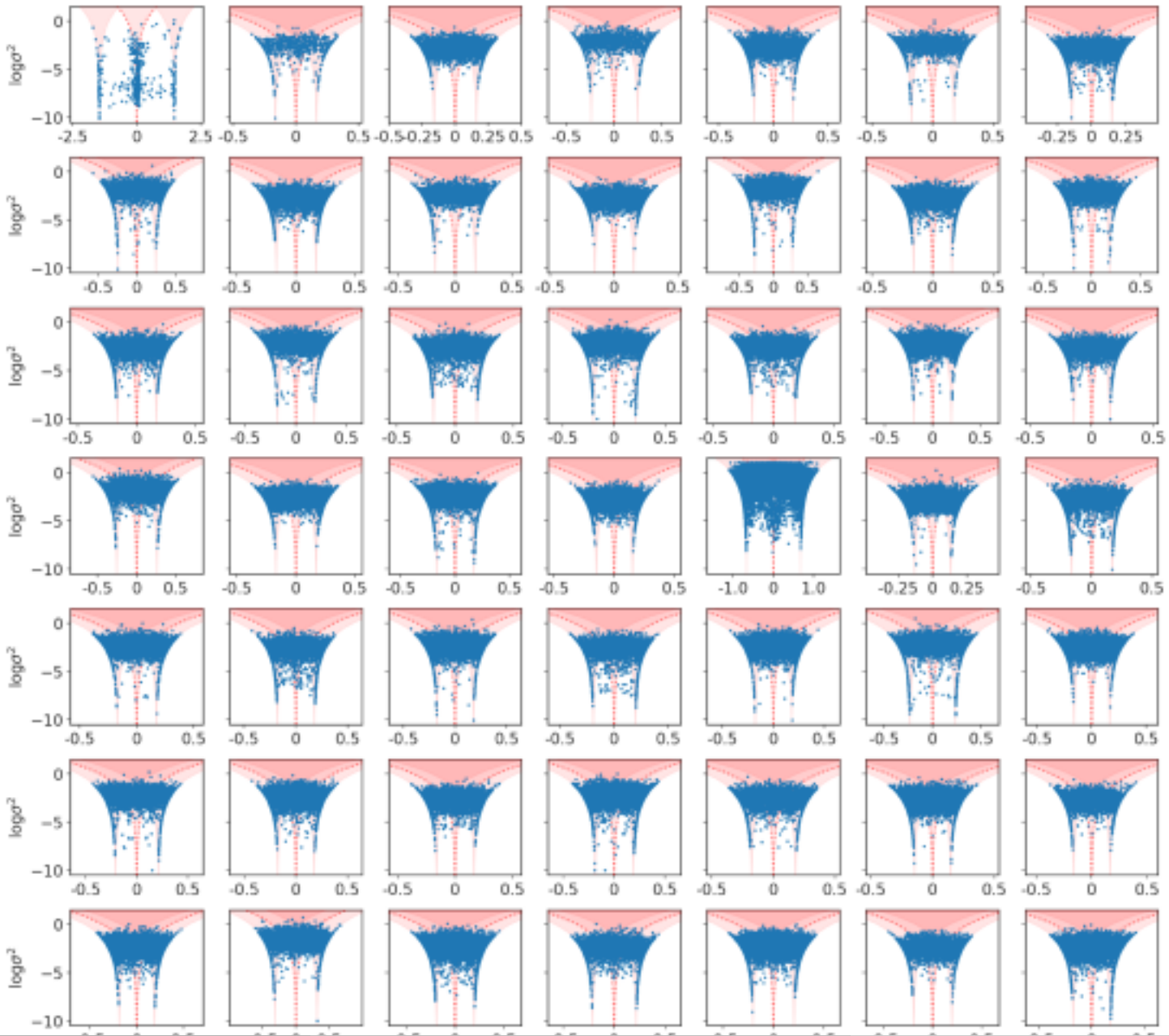
An equivalent interpretation is that a weight is pruned if the likelihood for the value 0 under the approximate posterior exceeds the threshold given by the standard-deviation band (Eq. (13)):

$$\mathcal{N}(0|\theta_{ij}, \sigma_{ij}^2) \geq \mathcal{N}(\theta_{ij} \pm \frac{\sigma_{ij}}{\sqrt{T_\alpha}}|\theta_{ij}, \sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{1}{2T_\alpha}}. \tag{14}$$

Extending this argument for pruning weights to a quantization setting, we design a post-training quantization scheme that assigns each weight the quantized value $c_k$ with the highest likelihood under the approximate posterior. Since variational posteriors over weights are Gaussian, this translates into minimizing the squared distance between the mean $\theta_{ij}$ and the quantized values $c_k$:

$$\arg\max_k \ \mathcal{N}(c_k|\theta_{ij}, \sigma_{ij}^2) = \arg\max_k \ e^{-\frac{(c_k-\theta_{ij})^2}{2\sigma_{ij}^2}} = \arg\min_k \ (c_k - \theta_{ij})^2. \tag{15}$$

# Квантизация: результаты для MC



# Квантизация: вычисление KL

with $\epsilon_{ij} \sim \mathcal{N}(0, 1)$. In standard (Gaussian) dropout training, the dropout rates $\alpha$ (or $p$ to be precise) are fixed and the expected log likelihood $L_{\mathcal{D}}(\phi)$ (first term in Eq. (1)) is maximized with respect to the means $\theta$. Kingma et al. (2015) show that Gaussian dropout training is mathematically equivalent to maximizing the ELBO (both terms in Eq. (1)), under a prior $p(w)$ and fixed $\alpha$ where the KL term does not depend on $\theta$:

$$\mathcal{L}(\alpha, \theta) = \mathbb{E}_{q_\alpha}[L_{\mathcal{D}}(\theta)] - D_{\mathrm{KL}}(q_\alpha(w)||p(w)), \tag{4}$$

where the dependencies on $\alpha$ and $\theta$ of the terms in Eq. (1) have been made explicit. The only prior that meets this requirement is the scale invariant log-uniform prior:

$$p(\log|w_{ij}|) = \text{ const.} \quad \Leftrightarrow \quad p(|w_{ij}|) \propto \frac{1}{|w_{ij}|}. \tag{5}$$

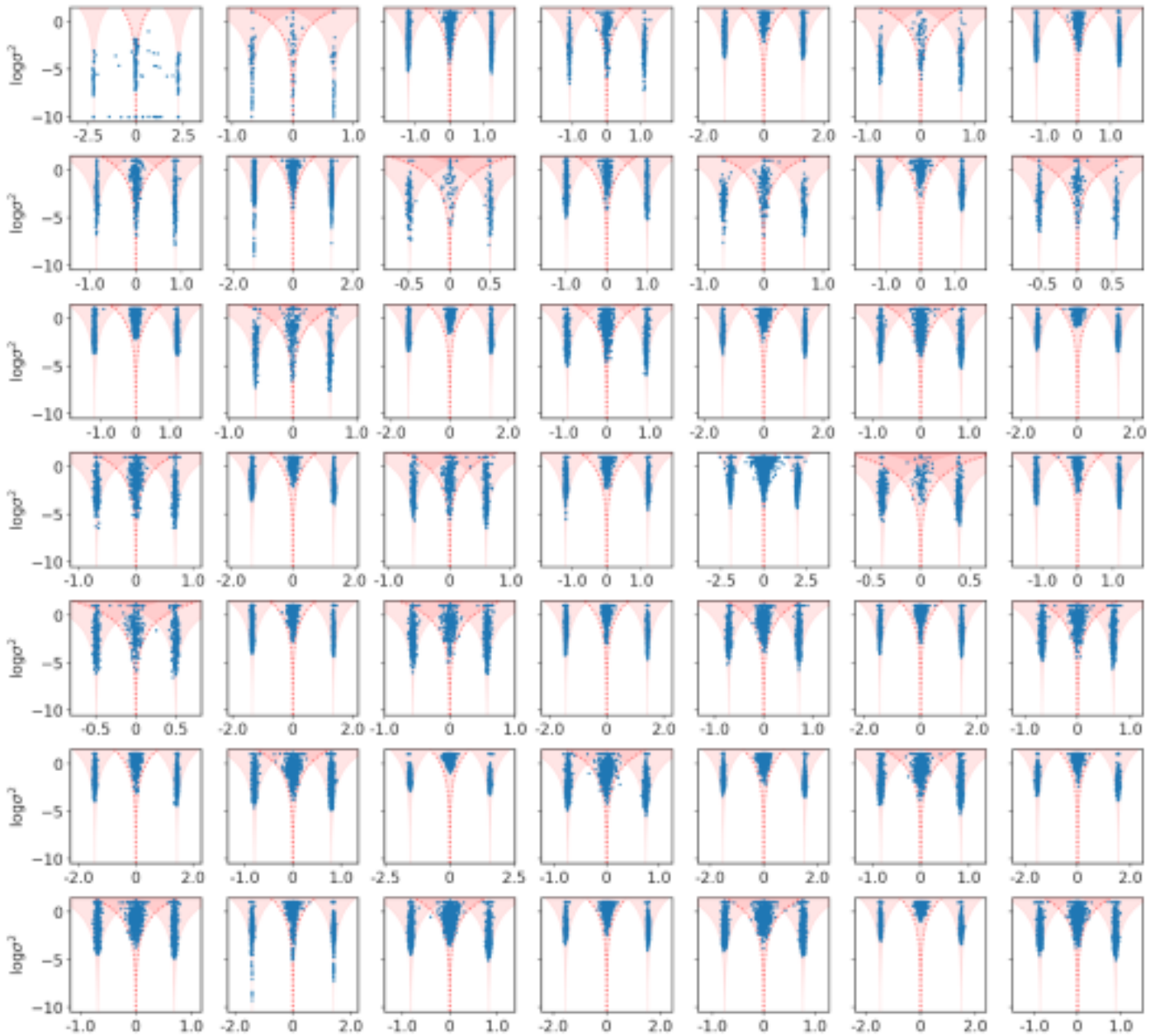
$$-D_{\mathrm{KL}}(q_\phi(w_{ij})||p(w_{ij})) \approx \text{const.} + k_1 S(k_2 + k_3 \log \alpha_{ij}) - 0.5 \log(1 + \alpha_{ij}^{-1}) = F_{\mathrm{KL,LU}}(\theta_{ij}, \sigma_{ij}),$$

$$F_{\mathrm{KL}}(\theta, \sigma, c) = \underbrace{\sum_{k:c_k \neq 0} \Omega(\theta - c_k) \mathrm{F}_{\mathrm{KL,LU}}(\theta - c_k, \sigma)}_{\text{local behavior}} + \underbrace{\Omega_0(\theta) \mathrm{F}_{\mathrm{KL,LU}}(\theta, \sigma)}_{\text{global behavior}} \tag{16}$$

with

$$\Omega(\theta) = \exp(-\frac{1}{2}\frac{\theta^2}{\tau^2}) \qquad \Omega_0(\theta) = 1 - \sum_{k:c_k \neq 0} \Omega(\theta - c_k). \tag{17}$$

# Квантизация: результаты



# Источники


- Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V., 2018. Learning transferable architectures for scalable image recognition
- David J. C. MacKay, Information Theory, Inference & Learning Algorithms
- Peter Grunwald, A tutorial introduction to the minimum description length principle
- Christopher Bishop, Pattern Recognition and Machine Learning
- A. Graves, Practical Variational Inference for Neural Networks
- Louizos C., Ullrich K., Welling M. Bayesian compression for deep learning
- Achterhold, J., Koehler, J. M., Schmeink, A., & Genewein, T. Variational network quantization.