Bayesian selection of deep learning model structure

Oleg Bakhteev

Supervisor: Prof. Vadim Strijov

Moscow Institute of Physics and Technology

November 21, 2019

Selection of deep learning model structure

Goal : to propose a method of selection of deep learning model structure.
Objectives:

1. Proposal of suboptimal and optimal complexity criteria for deep learning models.
2. Proposal of an algorithm suboptimal deep learning model selection and optimization of model parameters.

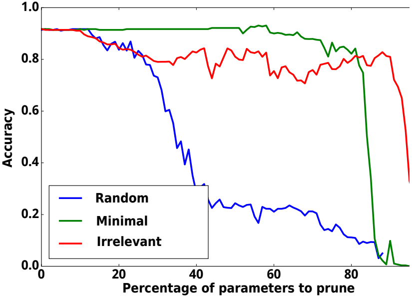
Investigated problems

1. Large number of parameters and hyperparameters, high computational complexity.
2. Multiextremality and non-convexity of optimization.

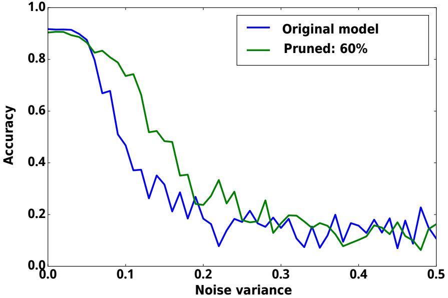
Methods
A deep learning model is considered as a multigraph. For the suboptimal model selection we use a composition of methods of automatic relevance determination and hyperparameter gradient optimization methods.

Model structure selection challenge

Data likelihood does not change with removing redundant parameters.



Redundancy of model parameters



Model robustness

Deep learning models have implicitly redundant complexity.

Deep learning model

Definition

Model $\mathbf{f}(\mathbf{w}, \mathbf{x})$ is a differentiable function with respect to parameters $\mathbf{w}$ from the set of object descriptions into the set of labels:

$$\mathbf{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{Y},$$

where $\mathbb{W}$ is a space of parameters of model $\mathbf{f}$.

Main challenge of deep learning model selection is in large number of parameters of models. This disallows to use many classical approached for the model and structure selection (AIC, BIC, cross-validation).

A model is defined by its parameters $\mathbf{W}$ and structure $\boldsymbol{\Gamma}$.
A **structure** defines a set of functional superpositions in the model. It is selected using statistical complexity criteria.
Empirical model complexity estimations:

1. number of parameters;
2. number of superpositions in the model.

Structure selection: one-layer network

The model $\mathbf{f}$ is defined by the **structure** $\mathbf{\Gamma} = [\gamma^{0,1}, \gamma^{1,2}]$.

$$\text{Model: } \mathbf{f}(\mathbf{x}) = \mathbf{softmax}\left((\mathbf{w}_0^{1,2})^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})\right), \quad \mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

$$\mathbf{f}_1(\mathbf{x}) = \gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) + \gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}),$$

where $\mathbf{w} = [\mathbf{w}_0^{0,1}, \mathbf{w}_1^{0,1}, \mathbf{w}_0^{1,2}]^{\mathsf{T}}$ — parameter matrices, $\{\mathbf{g}_{0,1}^0, \mathbf{g}_{0,1}^1, \mathbf{g}_{1,2}^0\}$ — generalized-linear functions, alternatives of layers of the network.

$$\begin{tikzcd}[column sep=huge]
\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left=60, "\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_0^{0,1})^{\mathsf{T}}\mathbf{x}\right)"] \arrow[r, bend right=60, "\gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}) = \gamma_1^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_1^{0,1})^{\mathsf{T}}\mathbf{x}\right)"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\mathbf{x}) = \gamma_0^{1,2}\mathbf{softmax}\left((\mathbf{w}_0^{1,2})^{\mathsf{T}}\mathbf{x}\right)"] & \mathbf{f}_2(\mathbf{x})
\end{tikzcd}$$

Deep learning model structure as a graph

Define:

1. acyclic graph (V, E);
2. for each edge $(j, k) \in E$: a vector primitive differentiable functions $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$ with length of $K^{j,k}$;
3. for each vertex $v \in V$: a differentiable aggregation function $\mathbf{agg}_v$.
4. a function $\mathbf{f} = \mathbf{f}_{|V|-1}$:

$$\mathbf{f}_v(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_v \left(\{ \langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k} \rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \mathrm{Adj}(v_k) \} \right), v \in \{1, \dots, |V|-1\}, \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \tag{1}$$

that is a function from $\mathbb{X}$ into a set of labels $\mathbb{Y}$ for any value of $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$.

Definition

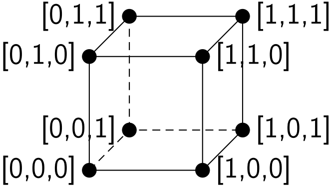
A *parametric set of models* $\mathfrak{F}$ is a graph (V, E) with a set of primitive functions $\{\mathbf{g}^{j,k}, (j,k) \in E\}$ and aggregation functions $\{\mathbf{agg}_v, v \in V\}$.

Statement

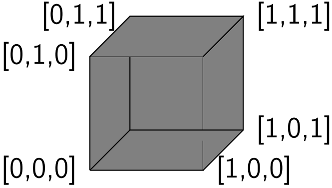
A function $\mathbf{f} \in \mathfrak{F}$ is a model for each $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$.

Structure restrictions

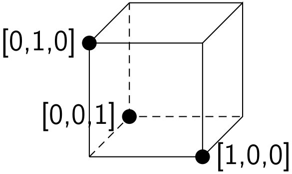
An example of restrictions for structure parameter γ, $|\gamma| = 3$.



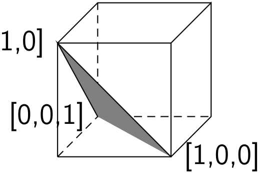
Cube vertices



Cube interior

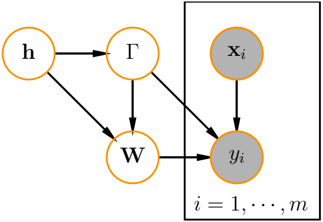


Simplex vertices



Simplex interior

Prior distribution



Definition

Prior distribution for parameters $\mathbf{w}$ and structure $\mathbf{\Gamma}$ of model $\mathbf{f}$ is a distribution $p(\mathbf{W}, \mathbf{\Gamma} | \mathbf{h}, \boldsymbol{\lambda}) : \mathbb{W} \times \mathbb{\Gamma} \times \mathbb{H} \to \mathbb{R}^+$, where $\mathbb{W}$ is a parameter space, $\mathbb{\Gamma}$ is a structure space, $\boldsymbol{\lambda}$ is a vector of metaparameters.

Definition

Hyperparameters $\mathbf{h} \in \mathbb{H}$ are the parameters of prior distribution $p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{h}, \mathbf{f})$ (parameters of the distribution of the parameters and structure of model $\mathbf{f}$).

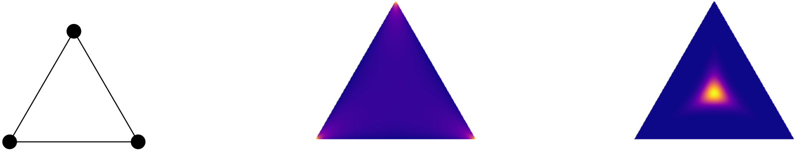
A model $\mathbf{f}$ is defined by:

- **Parameters** $\mathbf{w} \in \mathbb{W}$ that define superpositions $\mathbf{f}_v$ in the model $\mathbf{f}$.
- **Structure** $\mathbf{\Gamma} = \{\gamma^{j,k}\}_{(j,k) \in E} \in \mathbb{\Gamma}$ that define the contribution of all the superpositions $\mathbf{f}_v$ into $\mathbf{f}$.
- **Hyperparameters** $\mathbf{h} \in \mathbb{H}$ that define the prior distribution.
- **Metaparameters** $\boldsymbol{\lambda} \in \mathbb{\Lambda}$ that define the optimization function.

Prior distribution for the model structure

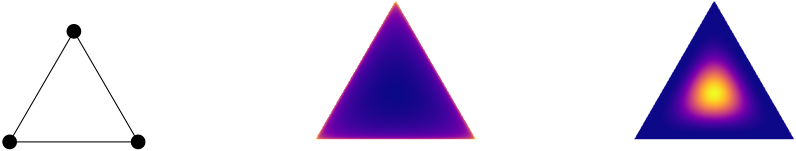
Every point in a simplex defines a model.

Gumbel-Softmax distribution: $\boldsymbol{\Gamma} \sim \mathrm{GS}(\mathbf{s}, \lambda_{\mathrm{temp}})$



$\lambda_{\mathrm{temp}} \to 0$ $\lambda_{\mathrm{temp}} = 0.995$ $\lambda_{\mathrm{temp}} = 5.0$

Dirichlet distribution: $\boldsymbol{\Gamma} \sim \mathrm{Dir}(\mathbf{s}, \lambda_{\mathrm{temp}})$

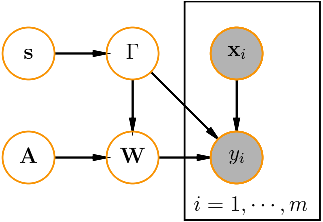


$\lambda_{\mathrm{temp}} \to 0$ $\lambda_{\mathrm{temp}} = 0.995$ $\lambda_{\mathrm{temp}} = 5.0$

Bayesian model selection

Base model:

- **parameters**
  $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1})$,
- **hyperparameters**
  $\mathbf{h} = [\alpha]$.



Proposed model:

- **parameters**
  $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, (\gamma_r^{j,k})^2(\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ is a diagonal matrix for the parameters of the primitive function $\mathbf{g}_r^{j,k}$,
  $(\mathbf{A}_r^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$,
- **structure**
  $\mathbf{\Gamma} = \{\gamma^{j,k}, (j, k) \in E\}$,
  $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- **hyperparameters** $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$,
- **metaparameters** $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Evidence as a statistical complexity

Minimum description length for the model $\mathbf{f}$:

$$\text{MDL}(\mathbf{y}, \mathbf{f}) = -\log\, p(\mathbf{h}|\mathbf{f}) - \log\, p(\hat{\mathbf{w}}|\mathbf{h}, \mathbf{f}) - \log\, \left(p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}, \mathbf{f})\delta\mathfrak{D}\right),$$

where $\delta\mathfrak{D}$ is an information transmission precision.

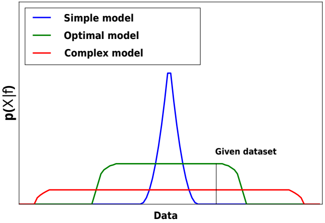
Bayesian approach:

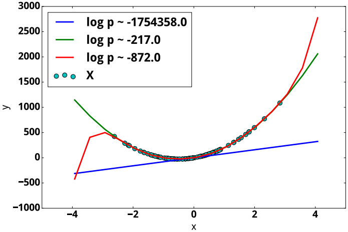
Obtain values of parameters $\mathbf{w}$ with respect to **posterior distribution of parameters**:

$$L = \log p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{h}, \boldsymbol{\lambda}) \propto \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}) + \log p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda}).$$

Hyperparameters are optimized using **posterior distribution of hyperparameters**:

$$Q = \log p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \propto \log p(\mathbf{h}|\mathbf{f}) + \log \int_{\mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\lambda}) p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda}) d\mathbf{w}.$$





Evidence lower bound

The evidence is analytically intractable.

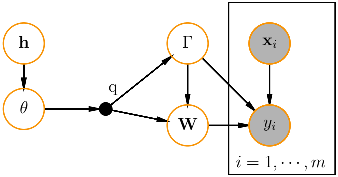
Model evidence:

$$p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \iint\limits_{\mathbf{w},\boldsymbol{\Gamma}} p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})d\mathbf{w}d\boldsymbol{\Gamma}.$$

Definition

Variational parameters of the model $\boldsymbol{\theta} \in \Theta$ are the parameters of the distribution q that approximates posterior distribution $p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{X},\mathbf{y},\mathbf{h},\boldsymbol{\lambda})$:

$$q \approx \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})}{\iint\limits_{\mathbf{w}',\boldsymbol{\Gamma}'} p(\mathbf{y}|\mathbf{X},\mathbf{w}',\boldsymbol{\Gamma}')p(\mathbf{w}',\boldsymbol{\Gamma}'|\mathbf{h},\boldsymbol{\lambda})d\mathbf{w}'d\boldsymbol{\Gamma}'}.$$



Lower bound of $\log p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$:

$$\log\ p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) \geq \mathsf{E}_q \log\ p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - \mathsf{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big).$$

The lower bound equals to evidence when

$$D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma})|p\left(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}\right)\big) = 0.$$

Model selection problem

Define a variational distribution $q = q_{\mathbf{w}} q_{\mathbf{\Gamma}}$ with parameters $\boldsymbol{\theta}$ that approximates posterior distribution $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{h}, \mathbf{f})$.

Definition

Loss function $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ is a differentiable function interpreted as a performance of the model on the train dataset.

Validation function $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ is a differentiable function interpreted as a general performance of the model.

The *model selection problem* $\mathbf{f}$ is a level optimization:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*, \boldsymbol{\lambda}),$$

where $\boldsymbol{\theta}^*$ is a solution for the following optimization:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathbb{U}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}).$$

Generalizing optimization problem

The model selection problem $\mathbf{h}^*, \boldsymbol{\theta}^*$ is a generalizing problem on the compact $U_\theta \times U_{\boldsymbol{h}} \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{\Lambda}$, if the following conditions are met:

1. For each parameter, hyperparameter and metaparameters its domain is not empty and not a point.
2. For each $\mathbf{h} \in U_{\boldsymbol{h}}$ and each $\boldsymbol{\lambda} \in U_\lambda$ the solution $\boldsymbol{\theta}^*$ is uniquely defined.
3. **Continuance:** L, Q are continuous with respect to metaparameters.
4. **optimal structure exhaustive search:** there is a constant $K_3 > 0$ and a value for the metaparameters $\boldsymbol{\lambda}$ such that for all pairs of local optima $\mathbf{h}_1, \mathbf{h}_2$ of Q with metaparameters $\boldsymbol{\lambda}$ such that

$$D_{\mathrm{KL}}\left(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\right) > K_3, D_{\mathrm{KL}}\left(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\right) > K_3,$$

$$Q(\mathbf{h}_1|\boldsymbol{\lambda}) > Q(\mathbf{h}_2|\boldsymbol{\lambda}),$$

   there exists another value of metaparameters $\boldsymbol{\lambda}' \neq \boldsymbol{\lambda}$ that

   1. the correspondence between optimal variational parameters and hyperparameters $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$ remains for $\boldsymbol{\lambda}'$,
   2. the following inequality is satisfied: $Q(\mathbf{h}_1|\boldsymbol{\lambda}') < Q(\mathbf{h}_2|\boldsymbol{\lambda}')$.

Generalizing optimization problem

The model selection problem $\mathbf{h}^*, \boldsymbol{\theta}^*$ is generalizing on the compact $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{\Lambda}$, if the following conditions are met:

5. **Likelihood maximization:** there is a metaparameter value $\boldsymbol{\lambda} \in U_\lambda$ and $K_1 \in \mathbb{R}_+$ such that for each pair of hyperparameter vectors $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$ the following inequality is satisfied : $\mathsf{E}_q \log\ p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f}) > \log \mathsf{E}_q\ p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \lambda_{\text{temp}}, \mathbf{f})$.
6. **Complexity minimization:** there is a metaparameter value $\boldsymbol{\lambda} \in U_\lambda$ and $K_2 \in \mathbb{R}_+$ such that for each pair of hyperparameter vectors $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, $\mathsf{E}_q \log\ p(\mathbf{y}|\boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) = \log \mathsf{E}_q\ p(\mathbf{y}|\boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$, the complexity of the first model is less than the second one.
7. **Evidence lower bound optimization:** there is a metaparameter value $\boldsymbol{\lambda}$, such that the optimization is equivalent to the evidence lower bound optimization:
$$\mathbf{h}^* \propto \arg\max \log \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}\left(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\right) + \log p(\mathbf{h}|\boldsymbol{\lambda}),$$
$$\boldsymbol{\theta}^* = \arg\min D_{\text{KL}}(q|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})).$$

Model selection problem analysis

Theorem [Bakhteev, 2019]

The following problems are not generalizing:

1. maximum likelihood criterion: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{h}, \boldsymbol{\lambda})$;
2. maximum posterior probability criterion: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h}, \boldsymbol{\lambda})$;
3. evidence lower bound maximization:
   $\max_{\mathbf{h}} \max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\left(p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})||q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})\right) + \log p(\mathbf{h}|\mathbf{f})$;
4. cross-validation: $\max_{\mathbf{h}} \mathsf{E}_q \log p(\mathbf{y}_{\mathsf{valid}}|\mathbf{X}_{\mathsf{valid}}, \boldsymbol{\theta}^*)$,
   $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}_{\mathsf{train}}|\mathbf{X}_{\mathsf{train}}, \mathbf{h}, \boldsymbol{\lambda}) p(\boldsymbol{\theta}|\mathbf{h})$.
5. AIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\mathsf{temp}}, \mathbf{f}) - |\theta_i : \mathrm{D}_{\mathrm{KL}}\left(q(w_i)|p(w_i|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}\right) < \lambda|$;
6. BIC:
   $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\mathsf{temp}}, \mathbf{f}) - \frac{1}{2}\log(|\mathbb{W}||\theta_i : \mathrm{D}_{\mathrm{KL}}\left(q(w_i)|p(w_i|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}\right) < \lambda|$;
7. structure exhaustive search:
   $\max \boldsymbol{\Gamma}' \max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\mathsf{temp}}, \mathbf{f}) \mathbb{I}(q(\boldsymbol{\Gamma}\Gamma = p')$, where p' is a distribution on a structure (metaparameter).

Proposed optimization problem

Theorem [Bakhtreev, 2019]

The following problem is generalizing:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q =$$

$$= \lambda^{\text{Q}}_{\text{likelihood}} \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta}^*)} \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda}) -$$

$$-{}^{\text{prior}}_{\text{Q}}\mathsf{D}_{KL}\big(q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta}^*)||p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) -$$

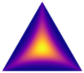
$$- \sum_{p' \in \mathfrak{P}, \lambda \in \boldsymbol{\lambda}^{\text{struct}}_{\mathbf{Q}}} \lambda \mathsf{D}_{KL}(p(\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})|p') + \log p(\mathbf{h}|\boldsymbol{\lambda}),$$

where

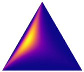
$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L = \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})$$

$$-{}^{\text{prior}}_{\text{L}}\mathsf{D}_{KL}\big(q^*(\mathbf{w},\mathbf{\Gamma})||p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big).$$

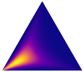
The proposed optimization generalized different optimization problems: maximum likelihood and evidence lower bound optimization, model complexity increase and decrease, exhaustive structure search.



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [0;0;0].$



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [1;0;0].$



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [1;1;0].$

Bayesian interpretation of the proposed optimization

Theorem, [Bakhteev, 2018]

Define a set of variational distribution $q(\boldsymbol{\theta})$.
Let $\lambda^{L}_{\text{likelihood}} = \lambda^{L}_{\text{prior}} = \lambda^{Q}_{\text{prior}} = 1, \boldsymbol{\lambda}^{Q}_{\text{struct}} = \mathbf{0}$. Then:

1. Solution of the proposed optimization problem obtains a maximum posterior distribution for the hyperparameters with evidence lower bound approximation:
$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) + \log p(\mathbf{h}|\mathbf{f}) \to \max_{\mathbf{h}}.$$
2. Variational distribution q for the solution approximates posterior distribution $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})$ in the best way:
$$D_{\text{KL}}(q||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})) \to \min_{\boldsymbol{\theta}}.$$

Let q be decomposed into two distributions for parameters $\mathbf{w}$ and structure $\boldsymbol{\Gamma}$ of the model $\mathbf{f}$:
$$q = q_{\mathbf{w}} q_{\boldsymbol{\Gamma}}, q_{\boldsymbol{\Gamma}} \approx p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \mathbf{f}), q_{\mathbf{w}} \approx p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{y}, \mathbf{X}, \mathbf{h}, \mathbf{f}).$$
If there are values for the variational parameters such that $q(\mathbf{w}) = p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$, $q(\boldsymbol{\Gamma}) = p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$, then the solution of optimization of L is equal to these values.

Optimization operator

Definition

An *optimization operator* T is an estimation of the new vector of parameters $\boldsymbol{\theta}'$ using the previous one $\boldsymbol{\theta}$.

Stochastic gradient descent operator :

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \dots \circ T(\boldsymbol{\theta}_0, \mathbf{h}) = T^\eta(\boldsymbol{\theta}_0, \mathbf{h}), \quad \text{где} T(\boldsymbol{\theta}, \mathbf{h}) =$$

$$= \boldsymbol{\theta} - \lambda_{\text{lr}} \nabla \left(-L(\boldsymbol{\theta}, \mathbf{h})|_{\hat{\mathfrak{D}}} \right),$$

λ_{lr} is a learning rate, $\boldsymbol{\theta}_0$ is an initial state for $\boldsymbol{\theta}$, $\hat{\mathfrak{D}}$ is a random subsample of the dataset $\mathfrak{D}$.
Reformulate the optimization problem:

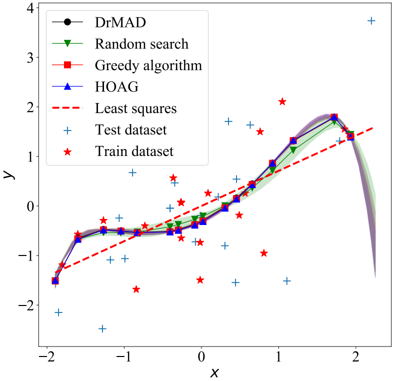
$$\mathbf{h}' = T^\eta\big(Q, \mathbf{h}, T^\eta(L, \boldsymbol{\theta}_0, \mathbf{h})\big).$$

Theorem, [Bakhteev, 2019]

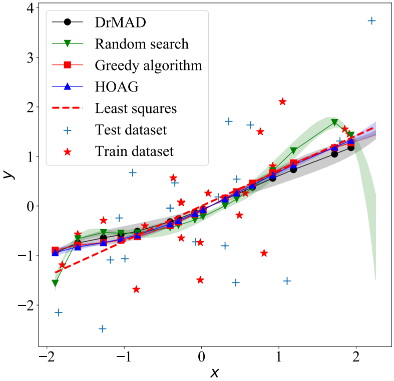
Let $\frac{\lambda^{\boldsymbol{Q}}_{\textbf{prior}}}{\lambda^{\boldsymbol{Q}}_{\textbf{likelihood}}} = \lambda^{L}_{\text{prior}}$. Then the proposed optimization is an one-level optimization.

Hyperparameter optimization: example

The hyperparameter gradient-based optimization methods were investigated.

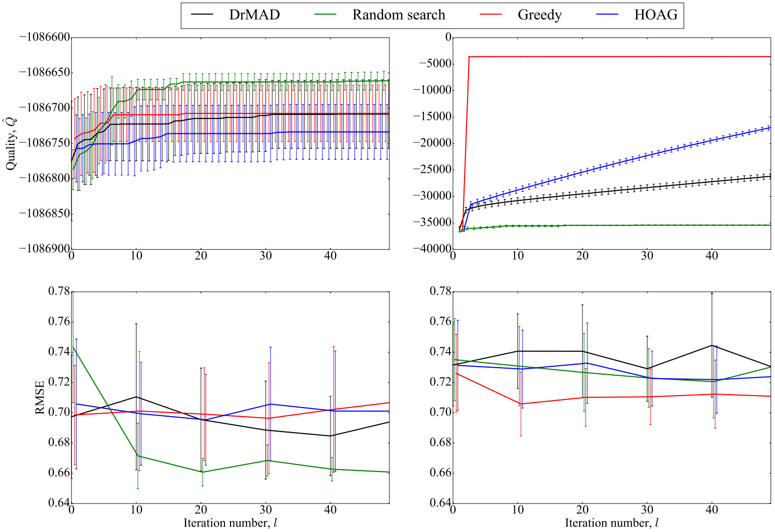


Cross-validation

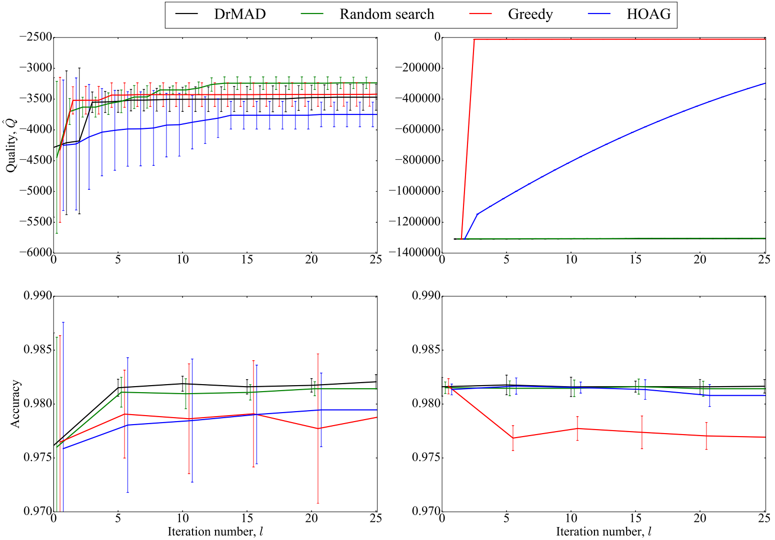


Evidence lower bound

Experiments: WISDM



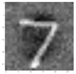
Experiments: MNIST



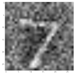
Experiments: MNIST

Noise adjusment $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$:

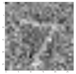
Original images



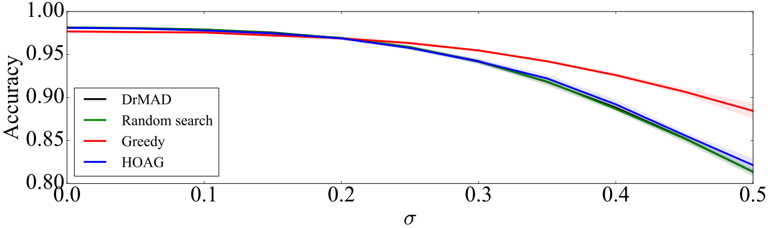
$\sigma = 0.1$



$\sigma = 0.25$



$\sigma = 0.5$



Evidence lower bound using multi-start

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \mathbf{f}) \geq \mathsf{E}_{q(\mathbf{W})} \log \, p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h}, \mathbf{f}) - \mathsf{E}_{q_{\mathbf{w}}}(-\log(q_{\mathbf{w}})).$$

Theorem [Bakhteev, 2016]

Let L be a loss function with continuously-differentiable gradient with Lipshitz constant C.
Let $\boldsymbol{\theta} = [\mathbf{w}^1, \dots, \mathbf{w}^k]$ be a vector of initial states of multiple model optimizations, λ_{lr} is a learning rate.
Then the difference of differentiable entropies for the optimization step can be estimated:

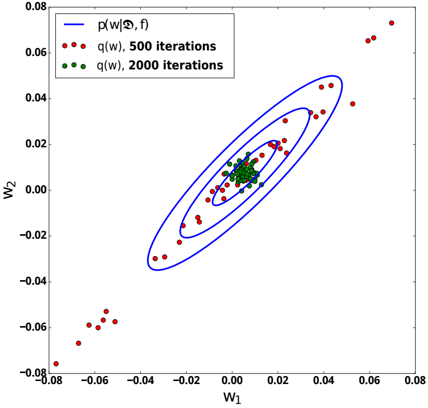
$$\mathsf{E}_{q^{\tau}_{\mathbf{w}}}(-\log(q^{\tau}_{\mathbf{w}})) - \mathsf{E}_{q^{\tau-1}_{\mathbf{w}}}(-\log(q^{\tau-1}_{\mathbf{w}})) \approx \frac{1}{k}\sum_{r=1}^{k}\bigl(\lambda_{\mathrm{lr}} Tr[\mathbf{H}(\mathbf{w}^r)] - \lambda^2_{\mathrm{lr}} Tr[\mathbf{H}(\mathbf{w}^r)\mathbf{H}(\mathbf{w}^r)]\bigr),$$

where $\mathbf{H}$ is a Hessian of the negative loss function $-L$, $q^{\tau}_{\mathbf{w}}$ is a distribution q at the iteration τ.

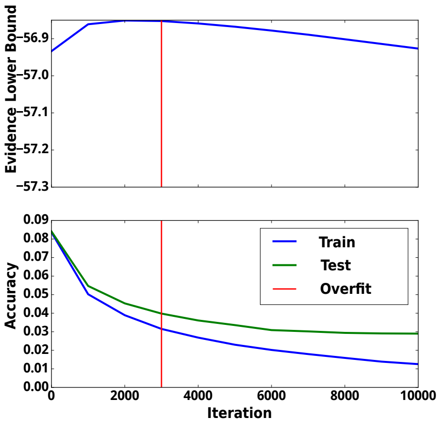
Gradient descent as an evidence lower bound

Empirical distribtuion of the optimized model parameters is a variational distribution.

Gradient descent does not optimize evidence lower bound.



Evidence lower bound decrease is a signal of overfitting.



Proposed optimization analysis

Theorem, [Bakhteev, 2018]

Let $\lambda_{\text{prior}}^{L} > 0, m \gg 0, \frac{m}{\lambda_{\text{prior}}^{L}} \in \mathbb{N}$. Then optimization of

$$L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) - \lambda_{\text{prior}}^{L} D_{KL}(q||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp,f}})))$$

is a minimization of $\mathsf{E}_{\hat{\mathbf{X}}, \hat{\mathbf{y}} \sim p(\mathbf{X}, \mathbf{y})} D_{KL}(q||p(\mathbf{w}, \mathbf{\Gamma}|\hat{\mathbf{X}}, \hat{\mathbf{y}}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}))$, where $\hat{\mathbf{X}}, \hat{\mathbf{y}}$ is a random sample of size $\frac{m}{\lambda_{\text{prior}}^{L}}$.

Definition

Parametric complexity of the model is a minimal divergence:

$$C_p = \min_{\mathbf{h}} D_{\text{KL}}(q||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})).$$

Theorem, [Bakhteev, 2018]

Let $\boldsymbol{\lambda}_{\text{struct}}^{Q} = \mathbf{0}$. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{h}_1, \mathbf{h}_2$ are the optimization solutions for different metaparameter values $\lambda_{\text{prior}_1}^{\text{Q}}, \lambda_{\text{prior}_2}^{\text{Q}}, \lambda_{\text{prior}_1}^{\text{Q}} > \lambda_{\text{prior}_2}^{\text{Q}}$ on a compact U. Let function $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ be concave on U for $\lambda_{\text{prior}_2}^{\text{Q}}$. Then:

$$C_p(\boldsymbol{\theta}_1|U_{\mathbf{h}}, \boldsymbol{\lambda}_1) - C_p(\boldsymbol{\theta}_2|U_{\mathbf{h}}, \boldsymbol{\lambda}_2) < \frac{\lambda_{\text{prior}}^{\text{L}}}{\lambda_{\text{prior}_2}^{\text{Q}}}(\lambda_{\text{prior}_2}^{\text{Q}} - \lambda_{\text{prior}}^{\text{L}})C,$$

where C is a constant.

Proposed optimization analysis

Definition

Relative variational density is a ratio:

$$\rho(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{q_\mathbf{w}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}))}{q_\mathbf{w}(\text{mode } q_\mathbf{w})}.$$

Theorem, [Bakhteev, 2018]

Given $U_\mathbf{h} \subset \mathbb{H}, U_{\boldsymbol{\theta}_\mathbf{w}} \subset \mathbb{\Theta}_\mathbf{w}, U_{\boldsymbol{\theta}_\mathbf{\Gamma}} \subset \mathbb{\Theta}_\mathbf{\Gamma}$, variational and prior distributions $q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})$, $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ are absolutely continuous and unimodal $U_{\boldsymbol{\theta}}$ with equality of mode and mean. Let mode and mean of prior distribution be independent on the hyperparameters $\mathbf{h}$ and the structure $\mathbf{\Gamma}$.
Given a infinite sequence $\boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \dots, \boldsymbol{\theta}[i], \dots \in U_{\boldsymbol{\theta}}$ such that $\lim_{i\to\infty} C_p(\boldsymbol{\theta}[i]|U_\mathbf{h}, \boldsymbol{\lambda}) = 0$. Then

$$\lim_{i\to\infty} \mathsf{E}_{q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}[i])}\rho(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}[i], \mathbf{h}[i], \boldsymbol{\lambda})^{-1} = 1, \mathbf{h}[i] = \arg\min D_{\mathrm{KL}}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}_i)||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

Main results

The following results were proposed:

1. method of Bayesian selection of suboptimal structure;
2. optimal and suboptimal complexity criteria;
3. deep learning model graph description;
4. generalizing function that includes other methods of model selection:
   - evidence lower bound;
   - sequential complexity increase;
   - sequential complexity decrease;
   - structure exhaustive search;
5. method of evidence lower bound optimization based on mutlistart model optimization;
6. algorithm of optimization hyperparameters, structure and parameters for deep learning model.
7. The properties of the proposed optimization were investigated and comprehensively analyzed.

Publications

Main publications

1. Bakhteev, O., Kuznetsova, R., Romanov, A. and Khritankov, A. A monolingual approach to detection of text reuse in Russian-English collection // In 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) (pp. 3-10). IEEE.
2. Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации. // Системы и средства информатики. 2016. № 26.2. С. 4-22.
3. Romanov, A., Kuznetsova, R., Bakhteev, O. and Khritankov, A. Machine-Translated Text Detection in a Collection of Russian Scientific Papers. // Computational Linguistics and Intellectual Technologies. 2016.
4. Bakhteev, O. and Khazov, A., 2017. Author Masking using Sequence-to-Sequence Models // In CLEF (Working Notes). 2017.
5. Бахтеев О.Ю., Стрижов В.В. Выбор моделей глубокого обучения субоптимальной сложности. // Автоматика и телемеханика. 2018. №8. С. 129-147.
6. Огальцов А.В., Бахтеев О.Ю. Автоматическое извлечение метаданных из научных PDF-документов. // Информатика и её применения. 2018.
7. Смердов А.Н., Бахтеев О.Ю., Стрижов В.В. Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза. // Информатика и ее применения. 2019.
8. Грабовой А.В., Бахтеев О.Ю., Стрижов В.В. Определение релевантности параметров нейросети. // Информатика и её применения. 2019.
9. Bakhteev O., Strijov V. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research. 2019.

Conference talks

1. "Восстановление панельной матрицы и ранжирующей модели в разнородных шкалах", Всероссийская конеренция «57-я научная конеренция МФТИ», 2014.
2. "Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия", Международная конференция «Интеллектуализация обработки информации», 2016.
3. "Градиентные методы оптимизации гиперпараметров моделей глубокого обучения", Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
4. "Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ", Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
5. "Байесовский выбор наиболее правдоподобной структуры модели глубокого обучения", Международная конференция «Интеллектуализация обработки информации», 2018.