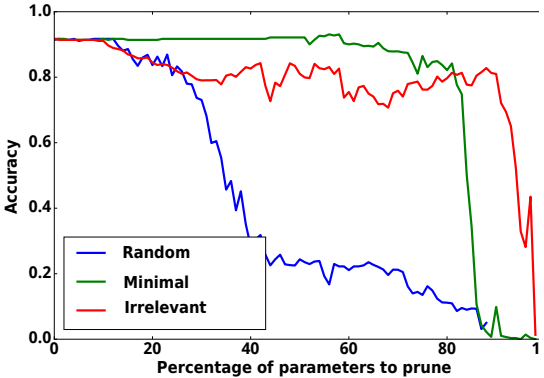
Bayesian multimodeling: model complexity

MIPT

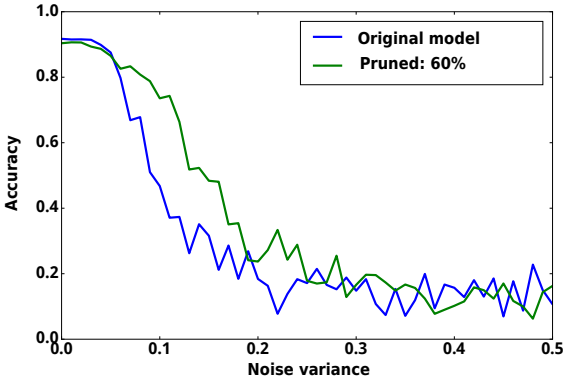
2023

Model structure selection challenge

Data likelihood does not change with removing redundant parameters.



Redundancy of model parameters



Model robustness

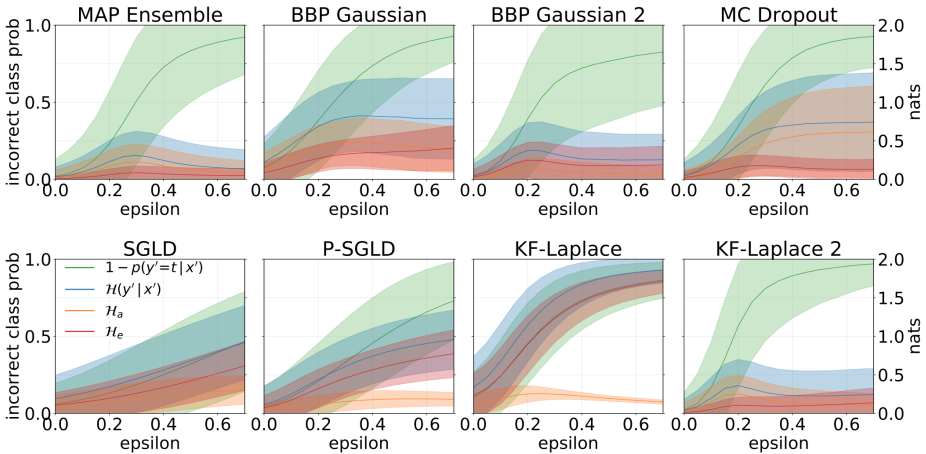
Deep learning models have implicitly redundant complexity.

Model complexity: application

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**

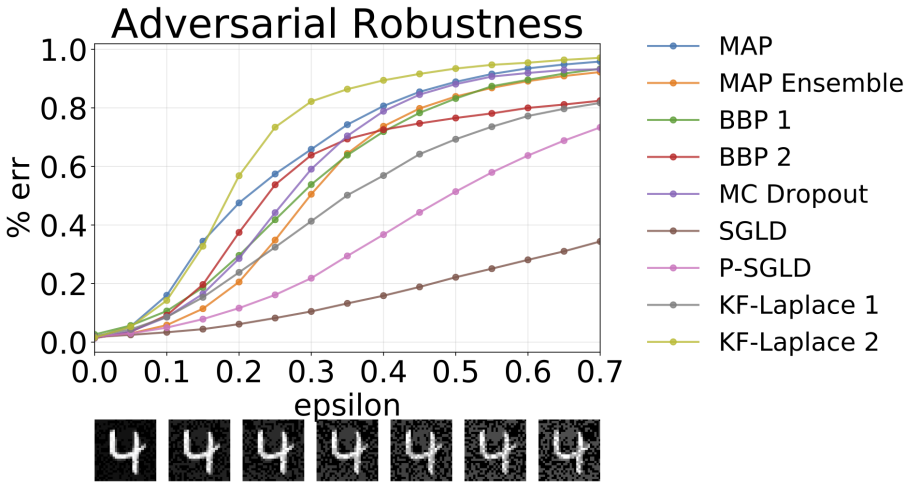
Zoph et al., 2017. Model parameter number differ twice, the performance is similar.

Robustness



https://github.com/JavierAntoran/Bayesian-Neural-Networks

Robustness



https://github.com/JavierAntoran/Bayesian-Neural-Networks

Naive model complexity

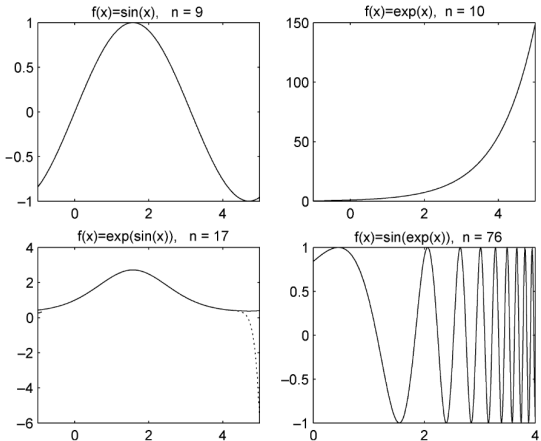
$C(\mathbf{f})$ is a number of parameters in the model.

Naive model complexity

$C(\mathbf{f})$ is a number of parameters in the model.

- Actually, a generalization of feature selection.
- Not differentiable.
- Does not respect the model structure.

Geometric model complexity



Complexity as a regularizer

Complexity doesn't come alone: usually it's a regularizer for a main term.

- L2-regularization: $||\mathbf{W}||^2 + \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{W})$.
- Evidence: $\int_{\mathbf{W}} p(\mathbf{W}) \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{W}) d\mathbf{W}$.

Corollary: complexity is seamless without main term, it's just a penalty for unnecessary/uninformative parameters/structure/models.

Occam's razor

- William of Ockham: «Plurality must never be posited without necessity».

Occam's razor

- William of Ockham: «Plurality must never be posited without necessity».
- Modern interpretation: entities should not be multiplied beyond necessity.
- Paul Dirac: «A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data.»
- Albert Einstein: «Everything Should Be Made as Simple as Possible, But Not Simpler»

When Occam's razor does not work

Occam's razor is an empricial rule for sorting hypothesis during research.
It can be wrong:

- Ernst Mach: molecules are a fictitious construct, as they are not observable.

Minimum description length

Task

Given a string: 001011001011001011... 001011, wheere the pattern 001011 repeats 100500 times.
How can we describe this string?

- `s == "001011...0010110010110010011")`
- `s == (''.join('001011' for _ in range(100500))`

Kolmogorov complexity

Definition

Given a computable partially defined mapping from a set of binary words into itself:

$$T : \{0,1\}^* \to \{0,1\}^*.$$

Kolmogorov complexity of the binary string x is a minimal descriptiong length w.r.t. T:

$$K_T(x) = \min_{f \in \{0,1\}^*} \{|f| : T(f) = x\},$$

Kolmogorov complexity

Generally, Kolmogorov complexity is incomputable.

Definition

Given a computable partially defined mapping from a set of binary words into itself:

$$T : \{0,1\}^* \to \{0,1\}^*.$$

Kolmogorov complexity of the binary string x is a minimal descriptiong length w.r.t. T:

$$K_T(x) = \min_{f \in \{0,1\}^*} \{|f| : T(f) = x\},$$

Discrete distribution entropy

Definition

Given a discrete variable x with probability p and values $x_1, \ldots, x_n$, An entropy of x is:

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log \; p(x = x_i).$$

Discrete distribution entropy

Definition

Given a discrete variable x with probability p and values $x_1, \ldots, x_n$, An entropy of x is:

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log \; p(x = x_i).$$

- interpretation: measure of disorder in the distribution;
- maximal at uniform distribution;
- minimal at a disribution with concentration only at one event ($x_i = 1, x_j = 0, i \neq j$).

Discrete distribution entropy

Definition

Given a discrete variable x with probability p and values $x_1, \ldots, x_n$, An entropy of x is:

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log \; p(x = x_i).$$

- interpretation: measure of disorder in the distribution;
- maximal at uniform distribution;
- minimal at a disribution with concentration only at one event $(x_i = 1, x_j = 0, i \neq j)$.
- **relation to Kolmogorov complexity:**

$$K(x) \leq H(x) + O(\log n)$$

for binary string with length n.

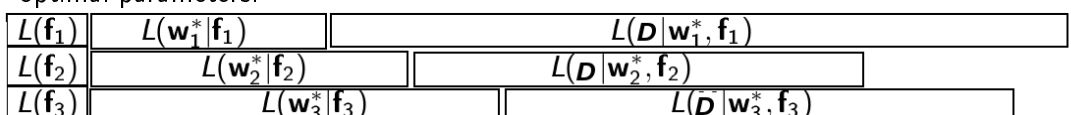
Minimum description length principle

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

where $\mathbf{f}$ is a model, $\mathfrak{D}$ is a dataset, L is a description length in bits.

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

$\mathbf{w}^*$ — optimal parameters.



MDL: example

Задача

Given a string: 001011001011001011... 001011, wheere the pattern 001011 repeats 100500 times.
How can we describe this string?

- `s == "001011...0010110010110010011")`
- `s == (''.join('001011' for _ in range(100500))`
- `import re; re.match('(001011){100500}')`
- $L(\mathbf{f}_1) = 0,\ L(\mathfrak{D}|\mathbf{f}_1) = 100505;$
- $L(\mathbf{f}_2) = 0,\ L(\mathfrak{D}|\mathbf{f}_2) = 45;$
- $L(\mathbf{f}_3) \gg 0,\ L(\mathfrak{D}|\mathbf{f}_3) = 38;$

MDL and Kolmogorov complexity

Kolmogorov complexity is a minimum descriptiong length for the dataset, described by a given programming language.

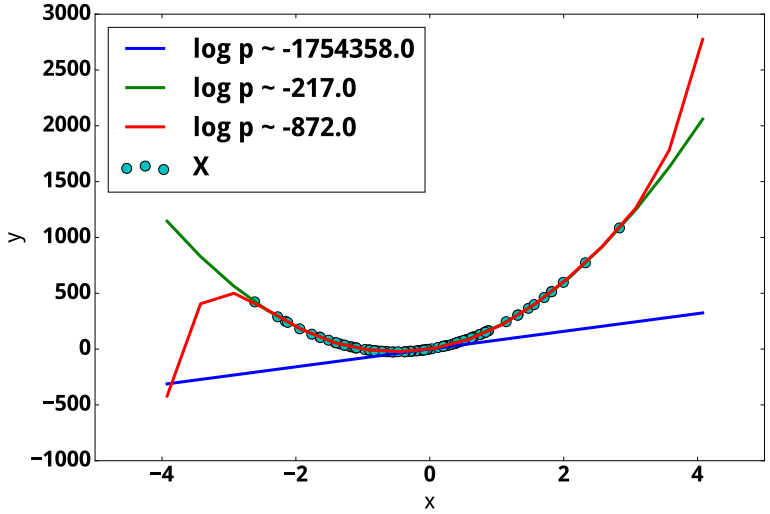
Invariance theorem

Given any description language L, the optimal description language is at least as efficient as L, with some constant overhead.

Difference from MDL:

- Incomputable.
- Code length depends on the language. For small sample size the invariance theorem can give poor results.

Probability coding



MDL + probability

Model selection problem can be viewed as a problem of information transmission, from encoder to decoder.

Given a dataset $\mathbf{X}, x \in \mathbf{X}$.

- Encoder encodes information about the dataset $\mathbf{X}$ with some code $\mathbf{f}$ and transmits it to the decoder.
- Decoder decodes the infomration $\mathbf{f}(\mathbf{X})$ ans restores the data $\mathbf{X}$ (with loss).
- Problem is to find optimal coding method for $\mathbf{x}$
- Code length: $-\log\ p(x)$

Quality criterion is *Regret*:

$$R(x) = -\log P(x) + \min_{\mathbf{f} \in \mathfrak{F}}(\log P(x|\mathbf{f})).$$

It shows the difference between the length of the real code $\log P(x)$ for x and the best code from the set of codes $\mathfrak{F}$.

Regret for datasets with parameterized distribution:

$$R(\mathbf{X}) = \max_{x \in \mathbf{X}}(-\log P(x) + \min_{\mathbf{w}}(\log P(x|\mathbf{w})).$$

MDL and Evidence

Statement

Let the likelihood function $p(\mathbf{X}|\mathbf{w},\mathbf{f})$ be from an exponential family of distributions:

$$p(x|\mathbf{w},\mathbf{f}) = h(x)g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}\cdot\mathbf{T}(x)\right),$$

where $h, g, \mathbf{T}$ are some functions, $\boldsymbol{\eta}$ is a distribution parameter.
Let prior be Jeffreys prior:

$$p(\mathbf{w}|\mathbf{f}) = \frac{\sqrt{I(w)}}{\int_w \sqrt{I(w)}},$$

Then regret and evidence differ only by some constant when $n \to \infty$:

$$\lim_{n\to\infty}\left(R(\mathbf{X}) - \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w},\mathbf{f})p(\mathbf{w}|\mathbf{f})d\mathbf{w}\right) = \text{Const}.$$

MDL vs. Evidence

Evidence	MDL
Uses prior knowledge	Independent from prior
Uses data generation hypothesis	Minimizes description length

Recap: Likelihood maximization

Likelihood maximization is a KL divergence mimimization:

$$\max_w L(\mathbf{X}, w) \iff \min KL\left(p^*\left(\mathbf{X}\right)|p\left(\mathbf{X}|w\right)\right).$$

Proof sketch

$$KL\left(p^*\left(\mathbf{X}\right)|p\left(\mathbf{X}|w\right)\right) = \mathrm{E}_{x\sim p^*(\mathbf{X})} \log\left(\frac{p^*\left(\mathbf{X}\right)}{p\left(\mathbf{X}|w\right))}\right) =$$

$$= \mathrm{Const} - \mathrm{E}_{x\sim p^*(\mathbf{X})} \log p\left(\mathbf{X}|w\right) \approx^{\text{Law of large numbers}}$$

$$\approx \mathrm{Const} - L(\mathbf{X}, w).$$

Akaike information criterion (AIC)

Main idea

- Function to consider: $KL\left(p\left(\mathbf{X}\right)|p\left(\mathbf{X}|\mathbf{w}\right)\right)$.
- Removing parts not depending on the model parameters $\mathbf{w}$, we get: $\mathrm{E}\log\ p\left(\mathbf{X}|\mathbf{w}\right)$.
- A naive estimation $\mathrm{E}\log\ p\left(\mathbf{X}|\mathbf{w}\right) \approx \log\ p\left(\mathbf{X}|\mathbf{w}\right)$ is biased, make an adjustement:

$$AIC = -2\log\ p\left(\mathbf{X}|\mathbf{w}\right) + 2|\mathbf{w}|.$$

Bayesian information criterion (BIC)

Main idea

- Consider evidence with flat prior: $p(\mathbf{w}) \propto \text{Const}$.
- Use Laplace approximation: $\log\ p(\mathbf{X}) \approx \log\ p(\mathbf{X}|\hat{\mathbf{w}}) + n\log\ |\mathbf{w}| + \mathit{Const}$.
- If $n >> |\mathbf{w}|$, ignore const term:

$$BIC = \log\ p(\mathbf{X}|\mathbf{w}) - \frac{1}{2} n \log\ |\mathbf{w}|.$$

For details see the talk of Konstantin Yakovlev, 2022

AIC vs BIC

- $BIC > AIC$ when $n \geq 8$
- Model selection with BIC leads to consistent estimation: when n grows, a probability of correct model selection converges to 1
- AIC minimzation asymptotically gives the model with maximal likelihood
- Heuristic: models, which criteria values differ by 1 or 2 from best values are distinguishable.
- Both criteria work poorly for large models.

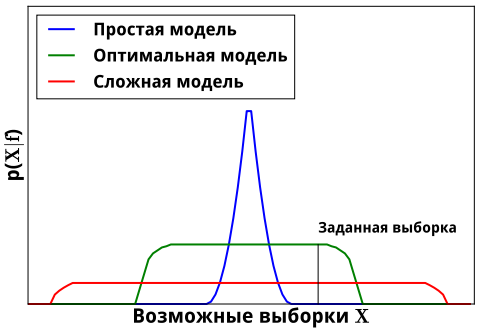
Cross-validation as a model selection method

Pros

- Easy to implement
- Straightforward implementation is simple for understanding. No approximations, no tricks.

Cons

- Either we lose data part, or we run performance.
- No explicit analysis of the extremum region.



Evidence vs Cross-validation

Evidence:

$$\log\, p(\mathbf{X}|\mathbf{f}) = \log\, p(\mathbf{x}_1|\mathbf{f}) + \log\, p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{f}) + \cdots + \log\, p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{f}).$$

Leave-one-out:

$$\mathrm{LOU} = \mathrm{E}\log\, p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{f}).$$

Cross-validation uses average value of the last term $p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{f})$.
Evidence considers **full** description length.

References

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Vladislavleva, E. J., Smits, G. F., & Den Hertog, D. (2008). Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. IEEE Transactions on Evolutionary Computation, 13(2), 333-349.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Grunwald P. A tutorial introduction to the minimum description length principle //arXiv preprint math/0406077. – 2004.
- Успенский В., Шень А., Верещагин Н. Колмогоровская сложность и алгоритмическая случайность. – Litres, 2017
- Grunwald P., Vitányi P. Shannon information and Kolmogorov complexity //arXiv preprint cs/0410002. – 2004.
- Vereshchagin N. K., Vitányi P. M. B. Kolmogorov's structure functions and model selection //IEEE Transactions on Information Theory. – 2004. – Т. 50. – №. 12. – С. 3265-3290.
- Штарьков Ю. М. Универсальное последовательное кодирование отдельных сообщений //Проблемы передачи информации. – 1987. – Т. 23. – №. 3. – С. 3-17.
- When Occam's razor does not work: https://hsm.stackexchange.com/questions/26/was-occam-s-razor-ever-wrong