Bayesian multimodeling: Distributions, expectation, likelihood

2023

Random variable

Given:

- A set of elementary events Ω
- A sigma-algebra $\mathfrak{F}$ over Ω
- A system $\mathfrak{B}(\mathbb{R})$ of Borel sets over $\mathbb{R}$

A real-valued function $w(\omega) : \Omega \to \mathbb{R}$ is called random variable, whenever for each $B \in \mathfrak{B}(\mathbb{R})$:

$$\{\omega : w(\omega) \in B\} \in \mathfrak{F}.$$

Random variable characteristics

A discrete random variable w takes a countable set of values $A = \{a_1, a_2, \dots\}$ with probabilities $p_1, p_2, \dots, \sum_i p_1 = 1$.

$f(a_i) = p(w = a_i) = p_i$ is a **probability function**.

A continious random variable is set using **cumulative distribution function (CDF)**:

$$F_w(t) = p(w \leq t)$$

or **probability density functon (PDF)**:

$$f(w) : \int_a^b f(w)\, dw = p(a \leq w \leq b).$$

Maximum likelihood

$$\mathrm{X} \sim f(\mathrm{x}, w),$$

$$L(\mathrm{X}, w) = \prod_{x \in \mathrm{X}} f(x, w),$$

$$\hat{w} \equiv \arg\max_{w} L(\mathrm{X}, w).$$

We can also use a logartihmic form:

$$\log L(\mathrm{X}, w) = \sum_{x \in \mathrm{X}} \log f(x, w),$$

$$\hat{w} \equiv \arg\max_{w} \log L(\mathrm{X}, w).$$

Maximum likelihood variations

Score function:

$$S(w) \equiv \frac{\partial}{\partial w} \log L(w)$$

Maximum likelihood estimator is a solution of the score equation:

$$S(w) = 0$$

Fisher information:

$$I(w) \equiv -\frac{\partial^2}{\partial w^2} \log L(w)$$

MLE dispersion:

$$\mathbb{D}\hat{\theta} \approx I^{-1}(\hat{w})$$

MLE properties

- Consistency:

$$\hat{w}_n \to^p w$$

- Asymptotic normality: $n \to \infty$

$$\hat{w} \sim \mathcal{N}\left(w, I^{-1}(w)\right)$$

- Efficiency: MLE's dispersion is minimal over all the unbiased estimations
- Invariance: $g\left(\hat{w}\right)$ is MLE over $g\left(w\right)$

Likelihood maximization

Likelihood maximization is a KL divergence mimimization:

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

Proof sketch

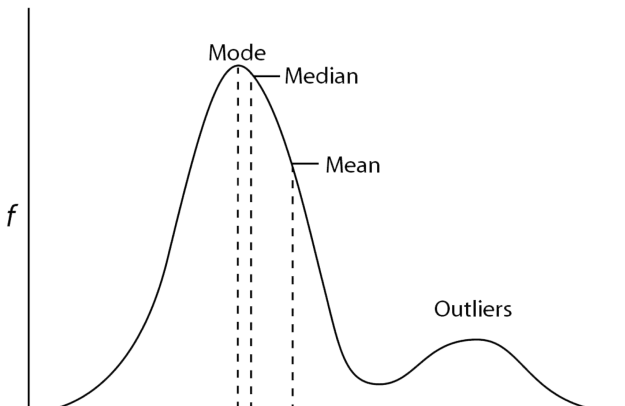
$$KL\left(p^*(\mathrm{X}) \,|\, p(\mathrm{X}|w)\right) = \mathrm{E}_{x\sim p^*(\mathrm{X})} \log\left(\frac{p^*(\mathrm{X})}{p(\mathrm{X}|w))}\right) =$$

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

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

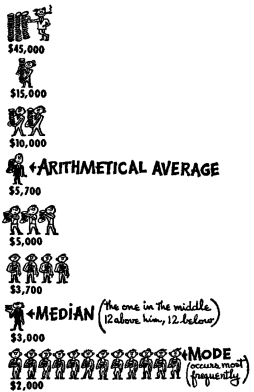
Central tendency

Empirical mean is an arithmetic mean for the given data.
Empirical median is a central element in the variational series.
Empirical mode is the most popular element in the data



Central tendency

(Huff, 1954):



Median

- α-quantile:

$$w_\alpha: \quad p(w \le w_\alpha) \ge \alpha, \quad p(w \ge w_\alpha) \ge 1 - \alpha$$

  or, equivalently:

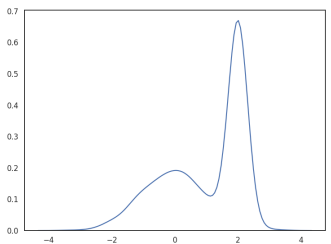
$$w_\alpha = F^{-1}(\alpha) = \inf\{w: F(w) \ge \alpha\}$$

- **median** is a 0.5-quantile, the central value of the distribution:

Mode

A mode is a maximum of the CDF or PDF:

$$\text{mode}(w) = \arg\max_{w} f(w)$$



Distribution can have multiple modes:

$$w \sim \alpha_1 \mathcal{N}(\mu_1, \sigma_1^2) + \alpha_1 \mathcal{N}(\mu_2, \sigma_2^2).$$

Expectation

Average value of the random variable w:

$$\mathrm{E}w = \int w \, dF(w).$$

- can be undetermined;
- linear;
- does not depend on the 0-measure values;
- Law of large numbers:
$$\bar{w}_n \to_{n\to\infty} \mathrm{E}w;$$
- Central Limit Theorem:
$$\sqrt{n}\frac{\bar{w}_n - \mathrm{E}w}{\sqrt{\mathrm{D}w}} \to \mathcal{N}(0,1).$$

Expectation: convergence

Often we need to investigate the convergence of the random variable, which is constructed as a composition of functions:

$$w \to^{\text{d/p/a.s.}} w^*,$$

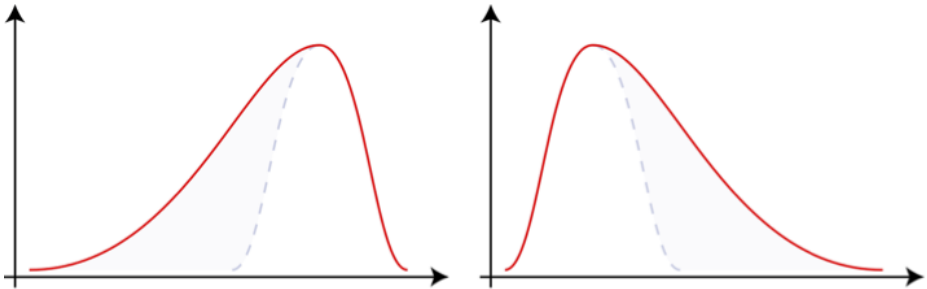
$$\mathrm{E}f(w) \to \mathrm{E}f(w^*)?$$

- f is continuous and limited, then $\mathrm{E}f(w) \to^{d} \mathrm{E}f(w^*)$.
- f is continuous almost everywhere, then $f(w) \to^{\text{d/p/п.н.}} f(w^*)$ (Mann–Wald theorem).
- A.s. convergence: Lebesgue theorem (swap limit and expectation).

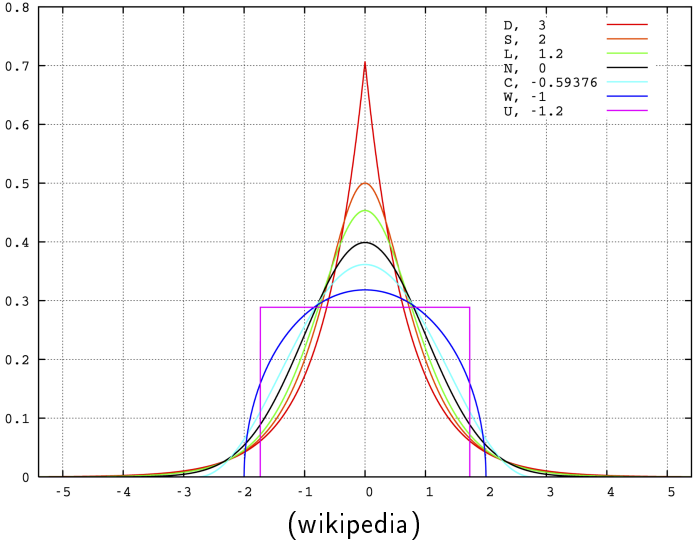
Other moments

- Second moment $\mathrm{D}w = \mathrm{E}(w - \mathrm{E}w)^2$: dispersion.
- Third moment $\frac{\mathrm{E}(w-\mathrm{E}w)^3}{\mathrm{D}w^{3/2}}$: asymmetry coefficient.
- Forth moment $\frac{\mathrm{E}(w-\mathrm{E}w)^4}{\mathrm{D}w^2} - 3$: Excess coefficient, pointedness of the PDF.

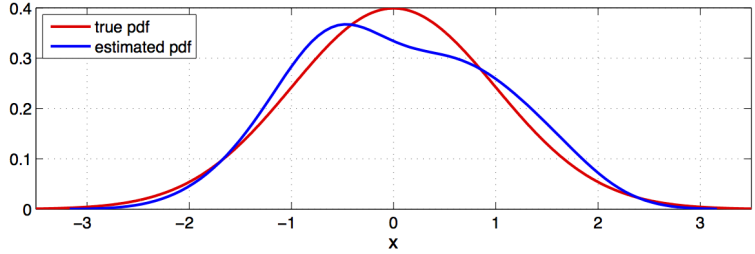
Other moments

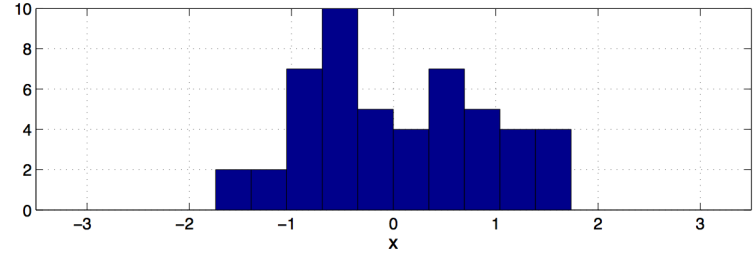


Other moments



PDF estimation

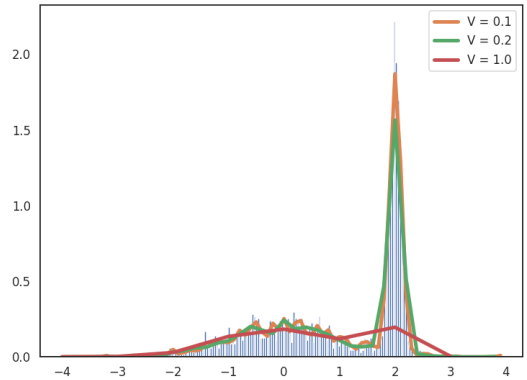




KDE

Estimation using kernel function K:

$$p(x) = \frac{1}{N} \sum_{x_i \in \mathrm{X}} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

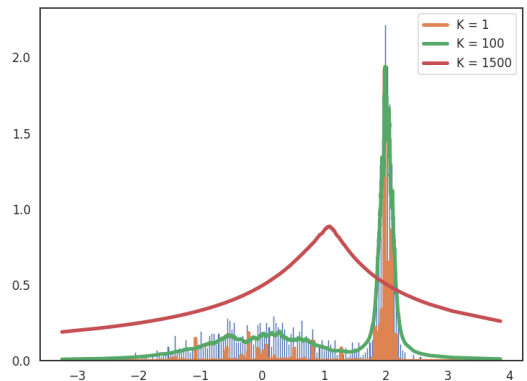


KDE

KNN can be considered as a generalization of KDE:

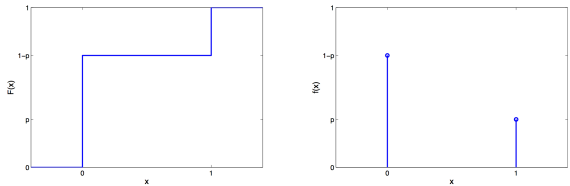
$$p(x) = \frac{1}{N} \sum_{x_i \in X} \frac{1}{h} K\left(\frac{x - x_i}{h_i}\right),$$

$$h_i = \text{dist}\left(x_i, \text{neighbour}\left(x_i, K\right)\right)$$



Bernoulli distribution

$w \in \{0, 1\} \sim Ber(p), \quad p \in (0, 1)$



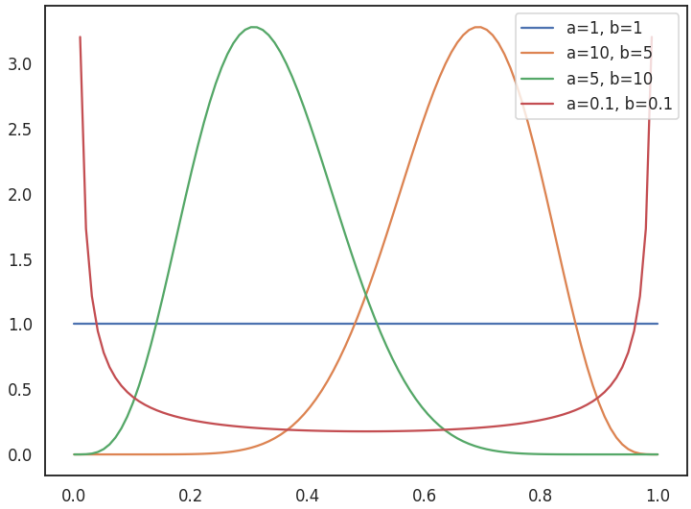
$$F(w) = \begin{cases} 0, & w < 0, \\ 1 - p, & 0 \le w < 1, \\ 1, & w \ge 1. \end{cases}$$

$$f(w) = \begin{cases} 1 - p, & w = 0, \\ p, & w = 1. \end{cases}$$

- example: coin flipping

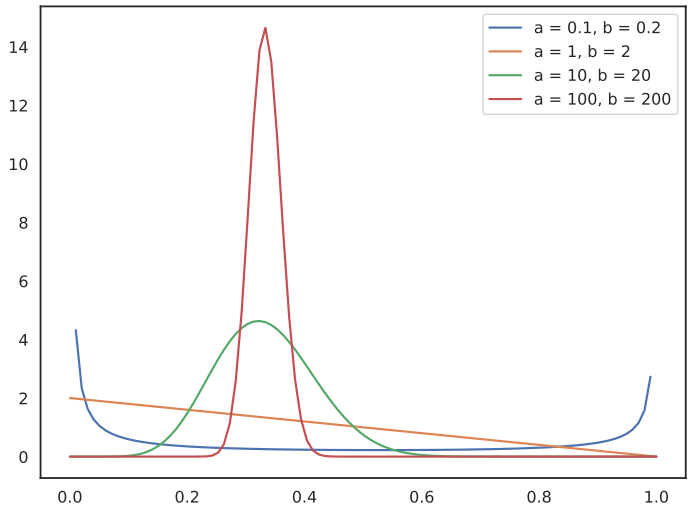
Beta-distribution

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}$$



Beta-distribution

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}$$



Beta-distribution

distribution

- corresponds to the *prior* beliefs about Bernoulli distribution
- interpretation: “effective number of events $w = 1, w = 0$”
- With $n \to \infty$ converges toδ-distribtuion with PDF concentration at MLE for Bernoulli.

Multinomial distribution

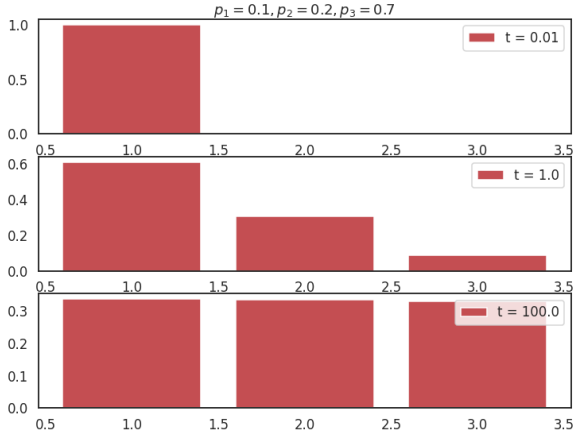
- Generalization of Bernoulli distribution for a larger number of events:

$$p(w = w_i | p_1, \ldots, p_n) = p_i.$$

- Can be parametrized using softmax:

$$\mathbf{softmax}(\log \mathrm{p}, t) = \frac{\mathbf{exp}(-\mathbf{log}\mathrm{p} \cdot t^{-1})}{\sum_{i=1}^{n} \exp(-logp_i \cdot t^{-1})}$$

Multinomial distribution



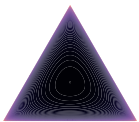
Dirichlet distribution

$$p(w_1, \ldots, w_K, \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} w_i^{\alpha_i - 1}$$

- Generalization of Beta-distribution for a larger number of events
- Support: $K-1$-simplex:
$$w_i \geq 0, \sum w_i = 1.$$
- Interpretation: a probability of each of K mutually exclusive events equals to w_i given that each event was observed $\alpha_i - 1$ times
- Can be parametrized using the following expression:

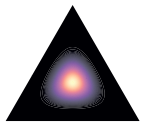
$$\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}} \cdot t, ||\boldsymbol{\alpha}|| = 1.$$

Dirichlet distribution

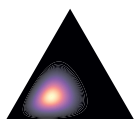


$\bar{\boldsymbol{\alpha}} = [1, 1, 1], t = 0.9$

$t = 1.0$



$t = 10.0$



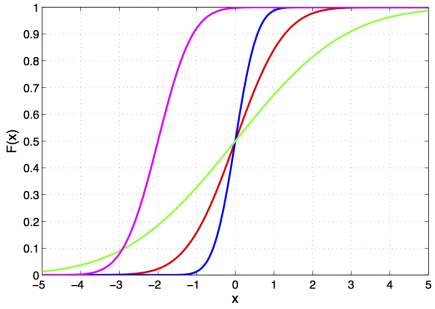
$\bar{\boldsymbol{\alpha}} = [0.5, 0.25, 0.25], t = 30.0$

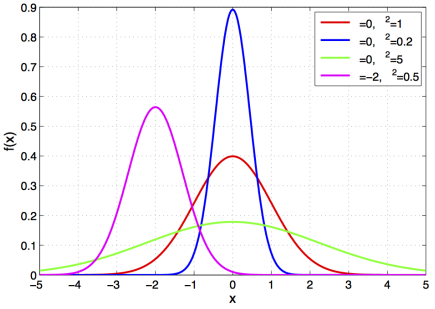
$[0.75, 0.125, 0.125]$

$[0.9, 0.05, 0.05]$

Normal distribution

$w \in \mathbb{R} \sim \mathcal{N}\left(\mu, \sigma^2\right),\ \sigma^2 > 0$





$$F(w) = \Phi\left(\frac{w-\mu}{\sigma}\right)$$

$$f(w) = \frac{1}{\sigma}\phi\left(\frac{w-\mu}{\sigma}\right)$$

$$\Phi(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w} e^{-\frac{t^2}{2}}\,dt$$

$$\phi(w) = \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}$$

Normal distribution

- limit of the sum of the weakly dependent random variables.
- $\mathrm{E}w = \mathrm{median}(w) = \mathrm{mode}(w) = \mu$, $\mathrm{D}w = \sigma^2$, the higher moments equal to zero
- if $w_1, \ldots, w_n$ are independent, $w_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$, then $\forall a_1, \ldots, a_n$

$$\sum_{i=1}^{n} a_i w_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

- Normal distribution has the maximal differential entropy among all the continuous distribtuions with limited dispersion.
- example: estimation error

Student distribution

- $Ew = 0$ при $\nu > 1$, $\text{median}(w) = \text{mode}(w) = 0$ always
- let $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$ independent, then

$$T = \frac{Z}{\sqrt{V/\nu}} \sim St(\nu)$$

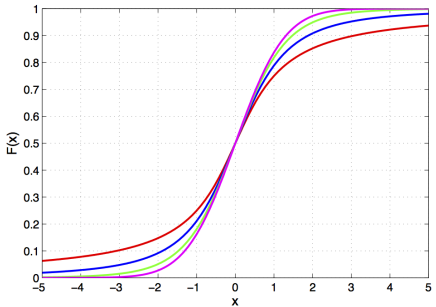
- if $w \sim St(\nu)$, then

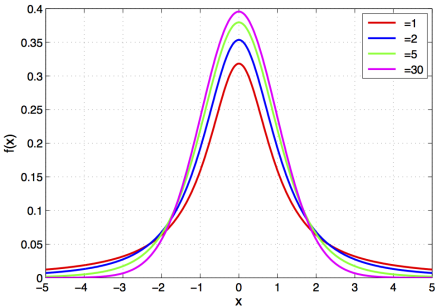
$$Y = \lim_{\nu \to \infty} w \sim \mathcal{N}(0,1)$$

- can be met during empirical mean estimation

Student distribution

$X \in \mathbb{R} \sim St(\nu),\ \nu > 0$





$$F(x) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right)$$

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Periodical distribution

?

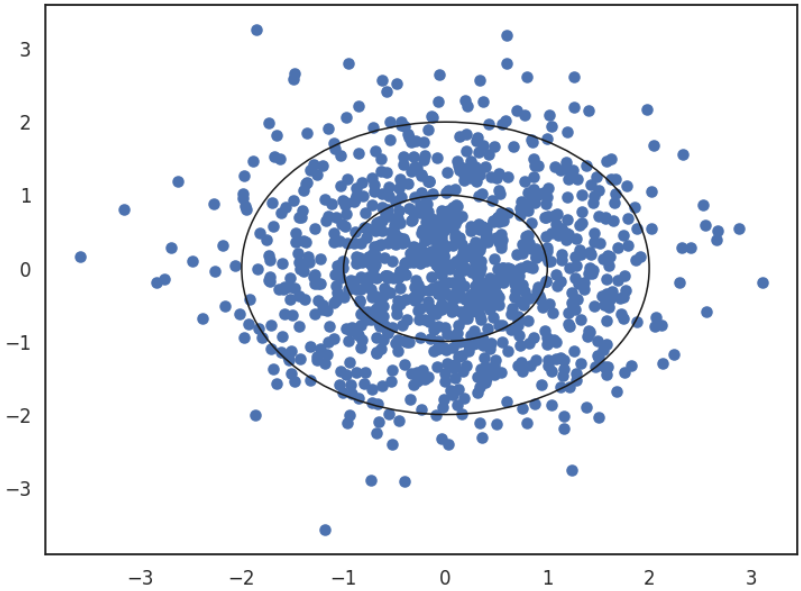
Periodical distribution

Naive approach: use a non-periodical distribution (let's say, Normal?)

Periodical distribution

- What if our dataset contains two objects on the circle: (1 degree and 359 degree)?
- $Ew = 180$, $D = 179$

Periodical distribution



Periodical distribution

- von Mises distribution:

$$p = \frac{\exp(k\cos(x-\mu))}{2\pi I_0(k)}, \quad I_0 \text{ is a Bessel function..}$$

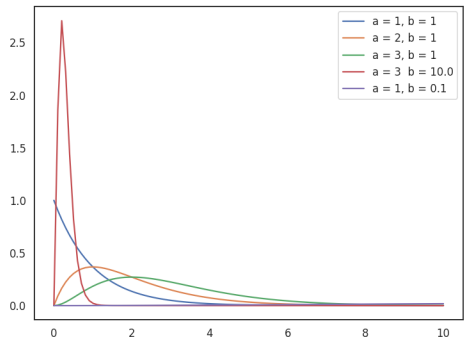
- Distribution idea:
  - $[w_1, w_2] \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
  - Parameterize using polar coordinate system, limit the radius:

$$w_1 = r\cos\phi, w_2 = r\sin\phi, r = \text{Const.}$$

- See also Anton Bishuk's 2021 task 2 about Kent distribution.

Gamma distribution

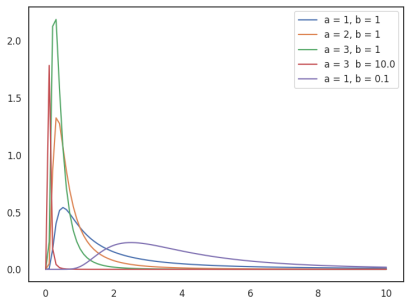
$$p(w, \alpha, \beta) = \frac{\beta^{\alpha} w^{\alpha-1} e^{-\beta}}{\Gamma(\alpha)}$$



- Can be used for prior distribution modeling, for the values inverse to standard deviation.

Inverse Gamma-distribution

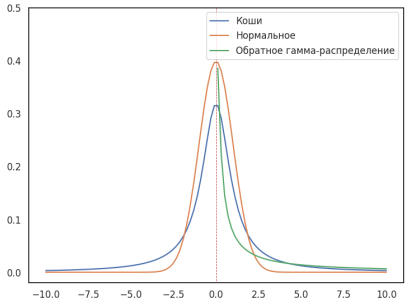
$$p(w, \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} w^{-\alpha-1} \exp(-\beta \cdot w^{-1}).$$



- $w \sim \Gamma(\alpha, \beta) \rightarrow w^{-1} \sim \text{Inv} - \Gamma(\alpha, \beta^{-1})$
- Can be used for prior distribution modeling, for the standard deviation for example.

Cauchy distribution

$$p(w, w_0, \gamma) = \frac{\gamma}{\pi\left((w - w_0)^2 + \gamma^2\right)}.$$



- Alternative to Normal distribution and Inverse Gamma-distribution
- Heavy tails
- The moments are undetermined

References

- Ширяев А. Н. Вероятность-1. – МЦНМО, 2007. – С. 552-552.
- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- KL minimization: https://wiseodd.github.io/techblog/2017/01/26/kl-mle/

Project overivew

Activities

Each team must assign roles for all teammates. Each activity is evaluated independently, thus number of activites per each teammate must be ≈equal.

- Project planning
- Basic code writing
- Algorithm implementation **(1 activity per teammate)**
- Project wrapping
- Tests writing
- Documentation writing
- Cross-review **(1 activity per teammate)**
- Final demo
- Blog post

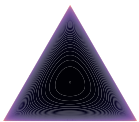
Dirichlet distribution

$$p(w_1, \ldots, w_K, \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} w_i^{\alpha_i - 1}$$

- Generalization of Beta-distribution for a larger number of events
- Support: $K-1$-simplex:
$$w_i \geq 0, \sum w_i = 1.$$
- Interpretation: a probability of each of K mutually exclusive events equals to w_i given that each event was observed $\alpha_i - 1$ times
- Can be parametrized using the following expression:

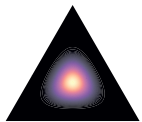
$$\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}} \cdot t, ||\boldsymbol{\alpha}|| = 1.$$

Dirichlet distribution

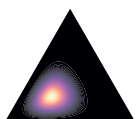


$\bar{\boldsymbol{\alpha}} = [1, 1, 1], t = 0.9$

$t = 1.0$



$t = 10.0$



$\bar{\boldsymbol{\alpha}} = [0.5, 0.25, 0.25], t = 30.0$

$[0.75, 0.125, 0.125]$

$[0.9, 0.05, 0.05]$

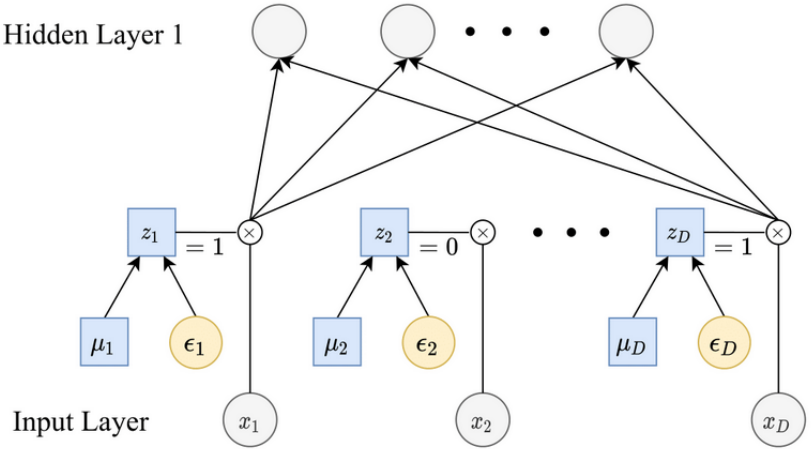
Project: Discrete variables relaxation

In this project the authors need to implement different alternatives to it. **Since the implementation is quite simple, each teammate must implement 2 algorithms, not 1.**

- Relaxed Bernoulli
- Correlated relaxed Bernoulli
- Gumbel-softmax TOP-K
- Straight-Trhrough Bernoulli, distrubtion (don't mix with Relaxed distribution from pyro)
- Invertible Gaussian reparametrization with KL implemented
- Hard concrete
- REINFORCE (not a distribution actually, think how to integrate it with other distributions)
- Logit-normal distribution with KL implemented and Laplace-form approximation of Dirichlet

Think about integration with pyro, torch.distributions or distrax.

Stochastic gating



Yamada et al.

Project: Stochastic gating

- Feature selection with L2 regularization and straight-through estimation see here for example , see also basic paper for ST-estimator
- Gumbel-softmax for gating
- Original stochastic gating
- Correlated features in stochastic gating

The team must decide the structure of the project: are the modules they propose are just layer classes (like dropout layers) or they propose some "solvers"? What's more preferable for reproducibility and furhter usage?

Model selection: coherent Bayesian inference

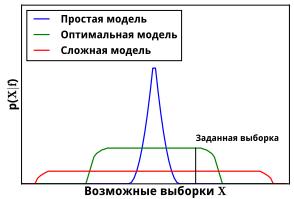
First level: find optimal parameters:

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

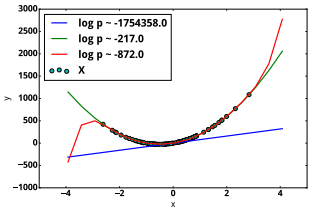
Second level: find optimal model:

Evidence:

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



Model selection scheme



Polynomial regression example

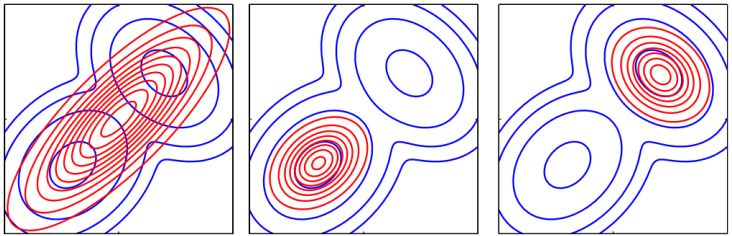
Evidence lower bound, ELBO

Evidence lower bound is a method of approximation of intractable distribution $p(\mathrm{w}|\mathfrak{D},\mathrm{h})$ with a distribution $q(\mathrm{w}) \in \mathfrak{Q}$.

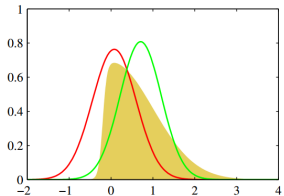
Evidence lower bound estimation often reduces to optimization problem

$$\log\ p(\mathfrak{D}|\mathrm{h}) \geq$$

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



Variational inference vs. expectation propogation (Bishop)



Laplace Approximation vs Variational inference

Project: Bayesian deep compression [1]

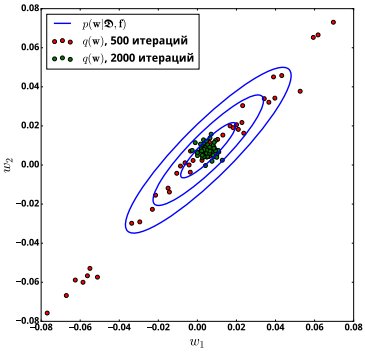
Each algorithm of evidence approximation proposes different strategies to prune/remove uninformative parameters.

- A baseline ELBO: proposed by Graves in 2011. Must be implemented with local reparameterization trick , hyperparameter optimization and pruning.
- Bayesian deep comrpession method: uses similar framework, but allows to effectively prune parameters due to the more sophisticated prior.
- Alternative method: scalable Laplace approximation
- Renyi divergence: a generalization of the ELBO which can potentially plugged into the algorithms 1 and 2 and make models more extedable

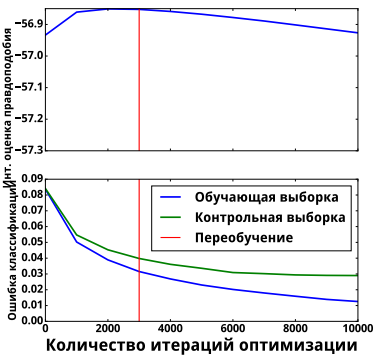
[1] The details will be discussed during classes on Bayesian inference and variational inference, see previous years

Overfitting, Maclaurin et. al, 2015

Instead of Monte-Carlo sampling, we can say that the model parameters are samples from distribution q_0. After a SGD step, they are transformed to samples fro mdiustribution of q_1, etc.



Convergence



Overfitting start

Project: Optimization operator as evidence estimators [2]

From the researcher perspective, the operator-based ELBO estimation is useufl because doesn't need to change the optimization at all.

- Current standard for these algorithms: SWA-G and Stein variational gradient (you can just call if from the library)
- Evidence lower bound using SGD
- SGLD-based operator
- Another approach based on approximation of SGD with Gaussian distribution

[2] The details will be discussed during classes on variational inference, see previous years

Reparametrization trick

Reparamterization idea:

$$\varepsilon = S_\theta(\mathrm{w}), \quad \mathrm{w} = S_\theta^{-1}(\varepsilon).$$

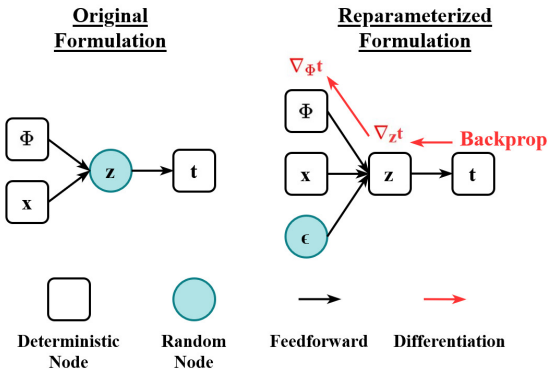
Then:

$$\nabla_\theta E_q f(\mathrm{w}) = \mathrm{E}_q \nabla_\theta f(S_\theta^{-1}(\varepsilon)).$$

Example:

$$w \sim \mathcal{N}(\mu, \sigma^2) \to S(w) = \frac{w-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

Challenge: calculation of S^{-1} is an expensive operation.



Source: wikipedia

Project: Implicit reparametrization trick [3]

The problem reparametrization trick is that it's available only for the limited number of distributions. There is a paper which proposes reformulation of sampling allowing to sample variables from any continuos distribution.

- Sample from Gaussian distribution (for comparison with reparametrization trick)
- Sample from Dirichlet distribution
- Sample from the mixture of distribution of the same family
- Sample from any arbitary factorized distribution

[3] The details will be discussed during classes on variational inference, see previous years

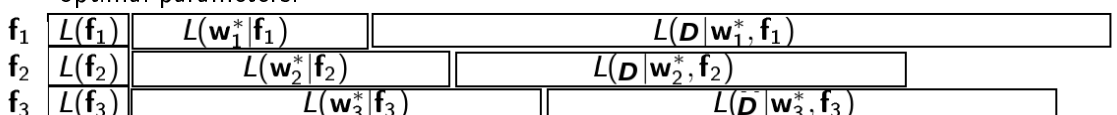
Minimum description length principle

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

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

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

w^* — optimal parameters.



Project: MDL for language model evaluation [4]

Probing: we train a classifier for some specific layer of language model to extract some information about the analyzed words.
For example, for PoS or syntactic properties of the words.
Main idea: accuracy of the model is not enough to evaluate the quality of probing.

- MDL approach, online coding
- MDL approach, variational coding
- Bayesian approach

[4] The details will be discussed during classes on model complexity, see previous years

Graphical models

GM is a broad class of probabilistic models: from linear regression to LDA and VAE.

Elements:

- White circles (random variables);
- Grey circels(observed variables);
- Small circles (deterministic values);
- Plates (batching).

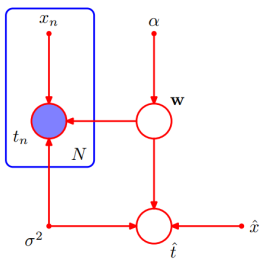


Plate notation for linear regression (Bishop)

Project: Graphical model generation [5]

The question of automatic model construction is still open, since it involves different model design decision and interpretation from user. The project proposes to use LLM to generate such models by the description in natural langauage.
Among all the projects this looks most simple, but in reality this project **is more a research project** than a project on implementation of existing methods. There is a belief it will work (since it's very similar to code generation task), but there is no proof. So, please **think twice before involving in this project**.

[5] The details will be discussed during classes on graphical models, see previous years

Hometask

1. The table for project assignment (at yandex disk) will be available on 21:00 Msk.
   - You should write you name and acitvites/roles.
2. The talks for the next classes are available on the “talk.md” page.
   - If you want to make a talk, write your name
   - You can propose your topic for the talk, it must be aligned with classes topic.
   - Reminder: in the talk there must be two questions for the questionnaire.
   - After the talk you will be asked to upload your slides and source files.
3. Kind reminder that after each classes we will have a questionnaire with very simple and trivial questions.
4. Score formulae a bit changed, sorry for that.