# Bayesian multimodeling: Variational inference-2

MIPT

2022

# Model selection: coherent Bayesian inference

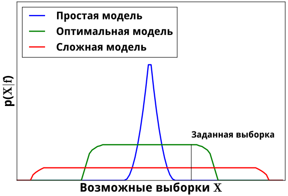
*First level:* find optimal parameters:

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

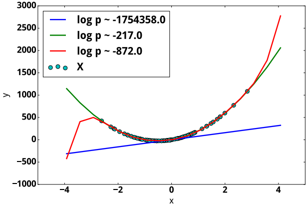
*Second level:* find optimal model:

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



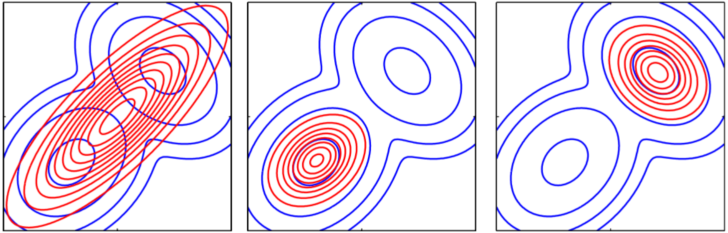

Polynomial regression example

# Evidence lower bound, ELBO

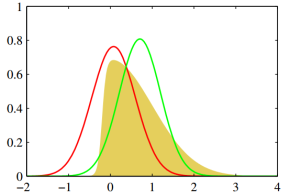
**Evidence lower bound** is a method of approximation of intractable distribution $p(\mathrm{w}|\mathfrak{D}, \mathrm{h})$ with a distribution $q(\mathrm{w}) \in \mathfrak{Q}$.

Evidence lower bound estimation often reduces to optimization problem

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



Variational inference vs. expectation propogation (Bishop)




Laplace Approximation vs Variational inference

# ELBO estimation

ELBO maximization

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

is equivalent to KL-divergence minimization between $q(\mathrm{w}) \in \mathfrak{Q}$ and posteriod distribution $p(\mathrm{w}|\mathrm{y}, \mathrm{X}, \mathrm{h})$:

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

$$\hat{q} = \arg\min_{q \in \mathfrak{Q}} \mathrm{D}_{\mathrm{KL}}\big(q(\mathrm{w})||p(\mathrm{w}|\mathrm{y}, \mathrm{X}, \mathrm{h})\big),$$

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

# MCMC and variational inference

**MCMC idea:** Sample from the simple distribution and accpet them, if the ratio is greater than some threshold:

$$\min\left(1, \frac{p(\mathrm{w}^\tau|\mathrm{y},\mathrm{X},\mathrm{h})}{p(\mathrm{w}^{\tau-1}|\mathrm{y},\mathrm{X},\mathrm{h})}\right),$$

where $\mathrm{w}^\tau$ is set based on the previous sample:

$$\mathrm{w}^\tau = T(\mathrm{w}^{\tau-1}).$$

**Salimans et al., 2014:** let's interperete the sequence of some operator $T$ application as a variational optimization:

$$T^1 \circ \dots T^\eta(\mathrm{w}) \to p(\mathrm{w}^\tau|\mathrm{y},\mathrm{X},\mathrm{h}).$$

**Maclaurin et. al, 2015:** use gradient descent as such operator. Do not reject samples at all.

# Optimization operator, Maclaurin et. al, 2015

**Definition**

Let $T$ be an algorithm of changing model parameters $\mathrm{w}'$ using previous parameter values $\mathrm{w}$:

$$\mathrm{w}' = T(\mathrm{w}).$$

**Definition**

Let $L$ be a continuos loss function.
Define a gradient descent operator in the following way:

$$T(\mathrm{w}) = \mathrm{w} - \beta \nabla L(\mathrm{w}, \mathrm{y}, \mathfrak{D}).$$

# Gradient descent for evidence estimation

Consider posterior probability maximization:

$$L = -\log p(\mathfrak{D}, \mathrm{w}|\mathrm{h}) = -\sum_{\mathfrak{D} \in \mathfrak{D}} \log p(\mathfrak{D}|\mathrm{w}, \mathrm{h}) p(\mathrm{w}|\mathrm{h})$$

Optimize neural network in a multi-start regime with $r$ initial parameter values $\mathrm{w}_1, \dots, \mathrm{w}_r$ using (stochastic) gradient descent:

$$\mathrm{w}' = T(\mathrm{w}).$$

The parameter vectors $\mathrm{w}_1, \dots, \mathrm{w}_r$ are from some latent distribution $q(\mathrm{w})$.

# Entropy

We can rewrite variational inference using differential entropy term:

$$\log p(\mathfrak{D}|\mathrm{f}) \geq \int_{\mathrm{w}} q(\mathrm{w}) \log \frac{p(\mathfrak{D}, \mathrm{w}|\mathrm{h})}{q(\mathrm{w})} d\mathrm{w} =$$

$$\mathsf{E}_{q(\mathrm{w})}[\log p(\mathfrak{D}, \mathrm{w}|\mathrm{h})] + \mathsf{S}(q(\mathrm{w})),$$

where $\mathsf{S}(q(\mathrm{w}))$ is a differential entropy:

$$\mathsf{S}(q(\mathrm{w})) = -\int_{\mathrm{w}} q(\mathrm{w}) \log q(\mathrm{w}) d\mathrm{w}.$$

# Gradient descent for evidence estimation

**Statement**

Let $L$ be a Lipschitz function, and optimization operator be a bijection. Then entropy difference for two steps is:

$$\mathsf{S}(q'(\mathsf{w})) - \mathsf{S}(q(\mathsf{w})) \simeq \frac{1}{r}\sum_{g=1}^{r}\bigl(-\beta\mathrm{Tr}[\mathsf{H}(\mathsf{w}'^{g})] - \beta^2\mathrm{Tr}[\mathsf{H}(\mathsf{w}'^{g})\mathsf{H}(\mathsf{w}'^{g})]\bigr).$$

Final estimation for the $\tau$ optimization step:

$$\log\hat{p}(\mathsf{Y}|\mathfrak{D},\mathsf{h}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\mathsf{w}_{\tau}^{g},\mathfrak{D},\mathsf{Y}) + \mathsf{S}(q^0(\mathsf{w})) +$$

$$+\frac{1}{r}\sum_{b=1}^{\tau}\sum_{g=1}^{r}\bigl(-\beta\mathrm{Tr}[\mathsf{H}(\mathsf{w}_b^{g})] - \beta^2\mathrm{Tr}[\mathsf{H}(\mathsf{w}_b^{g})\mathsf{H}(\mathsf{w}_b^{g})]\bigr),$$

$\mathsf{w}_b^g$ is a parameter vector for optimization $g$ on the step $b$, $\mathsf{S}(q^0(\mathsf{w}))$ is an initial entropy.

# How to calculate Hessian trace?

**Problem**

$$\mathrm{Tr}[\mathrm{H}(\mathrm{w}_b^g)]$$

**Statement**

Let U be a symmetric matrix and v be the random vector with the following properties:

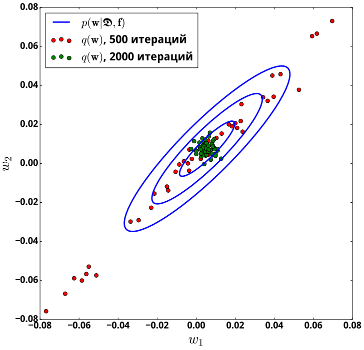
1. $\mathrm{E}v_i = 0$;
2. $Var(v_i) = 1$.

Then

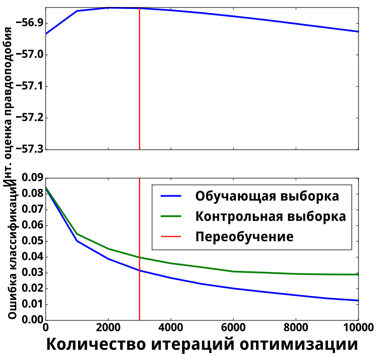
$$\mathrm{E}\mathrm{v}^{\mathrm{T}}\mathrm{U}\mathrm{v} = Tr[\mathrm{U}].$$

# Overfitting, Maclaurin et. al, 2015

Gradient descent does not optimize KL-divergence $\mathrm{KL}(q(\mathrm{w})||p(\mathrm{w}|\mathfrak{D}, \mathrm{h}))$. Evidence estimation gets worse while optimization tends to the optimal parameter values. This can be considered as a overfitting start.




Convergence




Overfitting start

# Stochastic gradient Langevin dynamics

A modification of SGD:

$$T = \mathrm{w} - \beta \nabla L + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\beta}{2})$$

where $\beta$ changes with a number of iterations:

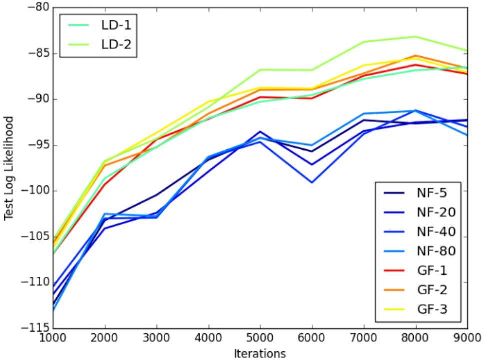
$$\sum_{\tau=1}^{\infty} \beta_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \beta_\tau^2 < \infty.$$

**Statement [Welling, 2011].** Distribution $q^\tau(\mathrm{w})$ converges to posterior distribution $p(\mathrm{w}|\mathrm{X}, \mathrm{f})$.
Entropy adjustment:

$$\hat{\mathrm{S}}(q^\tau(\mathrm{w})) \geq \frac{1}{2}|\mathrm{w}|\log(\exp(\frac{2\mathrm{S}(q^\tau(\mathrm{w}))}{|\mathrm{w}|}) + \exp(\frac{2\mathrm{S}(\epsilon)}{|\mathrm{w}|})).$$

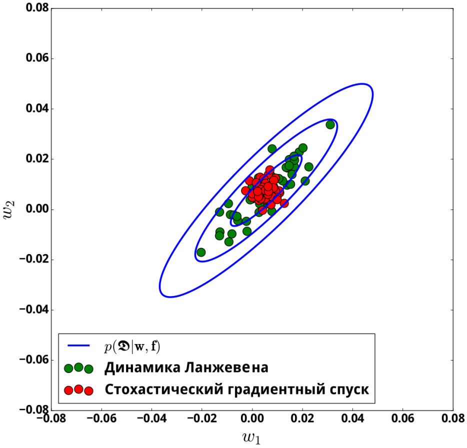
# Stochastic gradient Langevin dynamics for generative models

**Altieri et al., 2015:** sample latent variable z and use SGLD as a normalizing flow.



# SGLD vs SGD

Parameter distribution after 2000 iterations:



# Reparametrization trick: problems

Reparamterization idea:

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

Then:

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

**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.

# Implicit reparametrization trick

$$\nabla_{\boldsymbol{\theta}} E_q f(\mathrm{w}) = \mathrm{E}_q \nabla_{\mathrm{w}} f(\mathrm{w}) \nabla_{\boldsymbol{\theta}} \mathrm{w}.$$

Use a total gradient formula for $\varepsilon = S_{\boldsymbol{\theta}}(\mathrm{w})$:

$$\nabla_{\mathrm{w}} S_{\boldsymbol{\theta}}(\mathrm{w}) \nabla_{\boldsymbol{\theta}} \mathrm{w} + \nabla_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}(\mathrm{w}) = 0 \to$$

$$\to \nabla_{\boldsymbol{\theta}} \mathrm{w} = -(\nabla_{\mathrm{w}} S_{\boldsymbol{\theta}}(\mathrm{w}))^{-1} \nabla_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}.$$

Obtain an expression without inverse function for $S$.
For 1d samples we can use, for example:

$$S(\mathrm{w}) = F(\mathrm{w}|\boldsymbol{\theta}) \sim \mathcal{U}(0, 1).$$

Table 4: Test negative log-likelihood (lower is better) for VAE on MNIST. Mean $\pm$ standard deviation over 5 runs. The von Mises-Fisher results are from [9].

| Prior | Variational posterior | $D=2$ | $D=5$ | $D=10$ | $D=20$ | $D=40$ |
|---|---|---|---|---|---|---|
| $\mathcal{N}(0,1)$ | $\mathcal{N}(\mu,\sigma^2)$ | $131.1 \pm 0.6$ | $107.9 \pm 0.4$ | $92.5 \pm 0.2$ | $88.1 \pm 0.2$ | $88.1 \pm 0.0$ |
| $\mathrm{Gamma}(0.3, 0.3)$ | $\mathrm{Gamma}(\alpha,\beta)$ | $132.4 \pm 0.3$ | $108.0 \pm 0.3$ | $94.0 \pm 0.3$ | $90.3 \pm 0.2$ | $90.6 \pm 0.2$ |
| $\mathrm{Gamma}(10, 10)$ | $\mathrm{Gamma}(\alpha,\beta)$ | $135.0 \pm 0.2$ | $107.0 \pm 0.2$ | $92.3 \pm 0.2$ | $88.3 \pm 0.2$ | $88.3 \pm 0.1$ |
| $\mathrm{Uniform}(0,1)$ | $\mathrm{Beta}(\alpha,\beta)$ | $128.3 \pm 0.2$ | $107.4 \pm 0.2$ | $94.1 \pm 0.1$ | $88.9 \pm 0.1$ | $88.6 \pm 0.1$ |
| $\mathrm{Beta}(10,10)$ | $\mathrm{Beta}(\alpha,\beta)$ | $131.1 \pm 0.4$ | $\mathbf{106.7} \pm 0.1$ | $\mathbf{92.1} \pm 0.2$ | $\mathbf{87.8} \pm 0.1$ | $\mathbf{87.7} \pm 0.1$ |
| $\mathrm{Uniform}(-\pi,\pi)$ | $\mathrm{vonMises}(\mu,\kappa)$ | $\mathbf{127.6} \pm 0.4$ | $107.5 \pm 0.4$ | $94.4 \pm 0.5$ | $90.9 \pm 0.1$ | $91.5 \pm 0.4$ |
| $\mathrm{vonMises}(0,10)$ | $\mathrm{vonMises}(\mu,\kappa)$ | $130.7 \pm 0.8$ | $107.5 \pm 0.5$ | $92.3 \pm 0.2$ | $\mathbf{87.8} \pm 0.2$ | $87.9 \pm 0.3$ |
| $\mathrm{Uniform}(S^D)$ | $\mathrm{vonMisesFisher}(\boldsymbol{\mu},\kappa)$ | $132.5 \pm 0.7$ | $108.4 \pm 0.1$ | $93.2 \pm 0.1$ | $89.0 \pm 0.3$ | $90.9 \pm 0.3$ |

# Model selection: coherent Bayesian inference

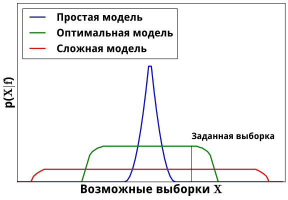
*First level:* find optimal parameters:

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

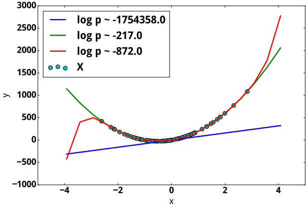
*Second level:* find optimal model:

Evidence:

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

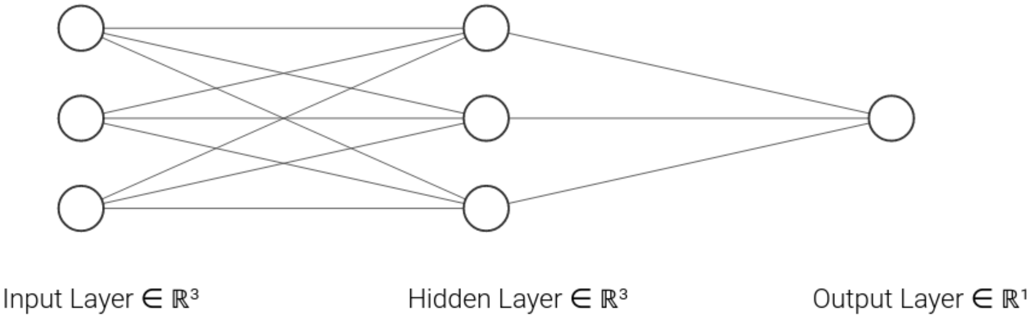



Model selection scheme

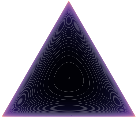



Polynomial regression example

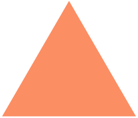
# Discrete variational optimization



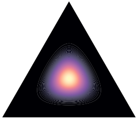
# Discrete distibution: relaxation



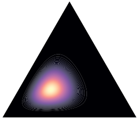
$\bar{\alpha} = [1, 1, 1], t = 0.9$



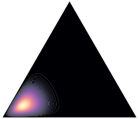
$t = 1.0$



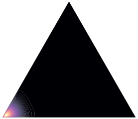
$t = 10.0$



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



$[0.75, 0.125, 0.125]$



$[0.9, 0.05, 0.05]$

# Discrete distribution in variational inference

Relaxation:

- Dirichlet distribution (+ Implicit reparametrization trick)
- Gumbel-softmax:

$$p(\mathrm{w}) = \Gamma(k)\tau^{k-1}(\sum_{i=1}^{k} \alpha_i / w_i)^{-k} \prod_{i=1}^{k} (\alpha_i / w_i \tau + 1)$$

  - Reparameterization works well
  - KL divergence is intractable
- Invertible Gaussian reparametrization:

$$p(\mathrm{w}) = \mathrm{soft\bar{m}ax}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \sim \mathcal{N},$$

(there should be a $\epsilon$ in the denominator for invertibility of the function)

  - Reparameterization works well
  - $KL(\mathrm{w}_1|\mathrm{w}_2) = KL(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2)$
  - Poor interpretation

# Local reparametrization

Let $y = \text{ReLU}(\text{XW})$ and parameter matrix W be distributed normally: $w_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$.
Then XW is a Gaussian matrix:

$$\text{G} = \text{XW}, \quad G_{i,j} \sim \mathcal{N}(\sum_k x_{i,k}\mu_{k,j}, \sum_k x^2_{i,k}\sigma^2_{k,j}).$$

Instead of sampling parameters, sample elements from G (units after activation).
Using CLT:

$$\sum_k x_{i,k} w_{k,j} \sim \mathcal{N}(\cdot, \cdot).$$

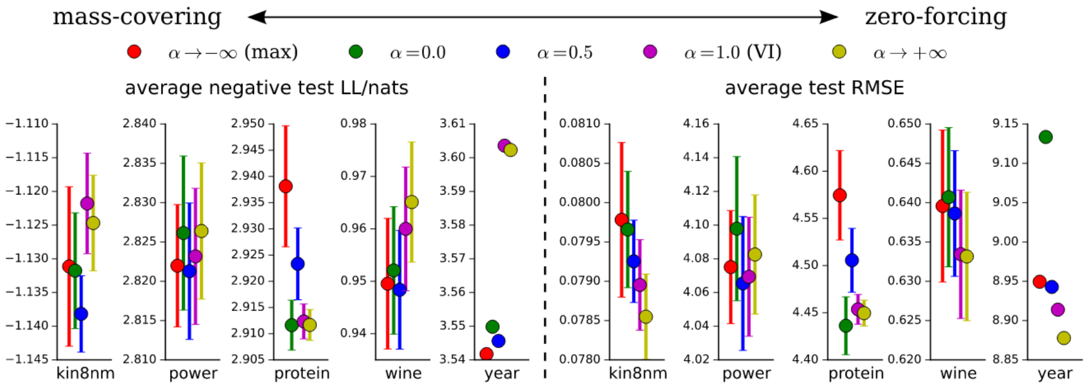
Conclusion: we can use the local reparametrization for discrete parameters too.

# Rényi divergence

$$D_\alpha(p(\mathrm{w})|q(\mathrm{w})) = \frac{1}{\alpha - 1} \log \int p(\mathrm{w})^\alpha q(\mathrm{w})^{1-\alpha} d\mathrm{w}.$$

Table 1: Special cases in the Rényi divergence family.

| $\alpha$ | Definition | Notes |
|---|---|---|
| $\alpha \to 1$ | $\int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$ | *Kullback-Leibler (KL) divergence*, used in VI ($\mathrm{KL}[q\|\|p]$) and EP ($\mathrm{KL}[p\|\|q]$) |
| $\alpha = 0.5$ | $-2\log(1 - \mathrm{Hel}^2[p\|\|q])$ | function of the square *Hellinger distance* |
| $\alpha \to 0$ | $-\log \int_{p(\boldsymbol{\theta})>0} q(\boldsymbol{\theta}) d\boldsymbol{\theta}$ | zero when $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$ (not a divergence) |
| $\alpha = 2$ | $-\log(1 - \chi^2[p\|\|q])$ | proportional to the $\chi^2$-divergence |
| $\alpha \to +\infty$ | $\log \max_{\boldsymbol{\theta}\in\Theta} \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$ | *worst-case regret* in *minimum description length principle* [24] |



mass-covering
zero-forcing
$\alpha \to -\infty$ (max)
$\alpha = 0.0$
$\alpha = 0.5$
$\alpha = 1.0$ (VI)
$\alpha \to +\infty$
average negative test LL/nats
average test RMSE
kin8nm
power
protein
wine
year
kin8nm
power
protein
wine
year

# 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.
- Salimans, Tim, Diederik Kingma, and Max Welling, 2015. Markov chain monte carlo and variational inference: Bridging the gap
- Altieri: http://approximateinference.org/accepted/AltieriDuvenaud2015.pdf
- Stephan Mandt, Matthew D. Hoffman, David M. Blei, 2017. Stochastic Gradient Descent as Approximate Bayesian Inference
- Бахтеев О. Ю., Стрижов В. В. Выбор моделей глубокого обучения субоптимальной сложности //Автоматика и телемеханика. – 2018. – №. 8. – С. 129-147.
- Figurnov M., Mohamed S., Mnih A. Implicit reparameterization gradients //arXiv preprint arXiv:1805.08498. – 2018.
- Jang E., Gu S., Poole B. Categorical reparameterization with gumbel-softmax //arXiv preprint arXiv:1611.01144. – 2016.
- Potapczynski A., Loaiza-Ganem G., Cunningham J. P. Invertible gaussian reparameterization: Revisiting the gumbel-softmax //arXiv preprint arXiv:1912.09588. – 2019.
- Maddison C. J., Mnih A., Teh Y. W. The concrete distribution: A continuous relaxation of discrete random variables //arXiv preprint arXiv:1611.00712. – 2016.
- Shayer O., Levi D., Fetaya E. Learning discrete weights using the local reparameterization trick //arXiv preprint arXiv:1710.07739. – 2017.
- Li Y., Turner R. E. Rényi Divergence Variational Inference //arXiv preprint arXiv:1602.02311. – 2016.
- Hutchinson, M. F. (1990). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. Communications in Statistics - Simulation and Computation, 19(2), 433–450.