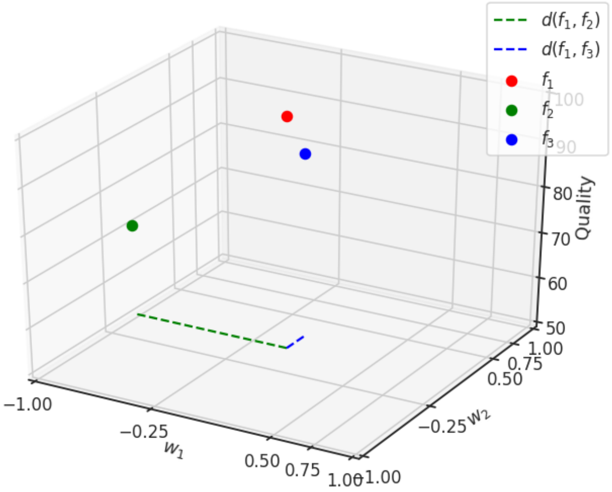
Probabilistic metric spaces

MIPT

2024

Motivation

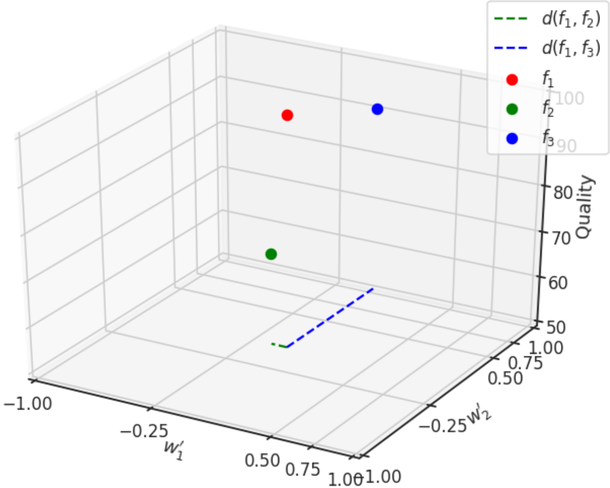
Which model is closer to f_1?



Motivation

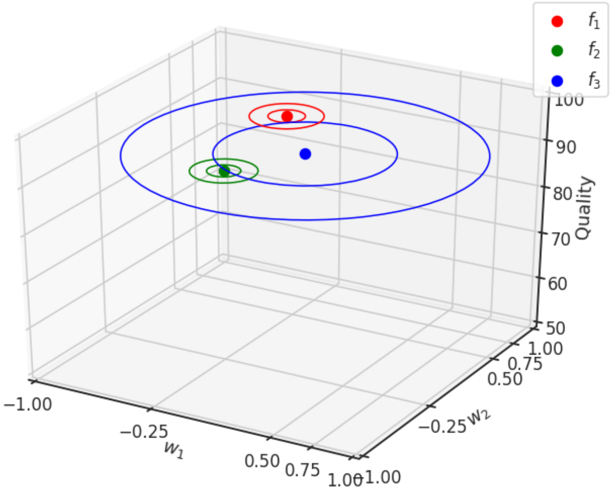
Which model is closer to f_1?

Metric change≈coordinate change. Different metrics represent different model space properties.



Motivation

Which model is closer to f_1?



Definition and properties

Given a parameter space $\mathbf{w}$.
A distance function d is a function, defined on the pair of distributions $p_1, p_2 \to \mathbb{R}_+$.
Probable Properties

- Metric axioms
  - $d(p_1, p_1) = 0$
  - $d(p_1, p_2) = d(p_2, p_1)$
  - $d(p_1, p_2) \leq d(p_1, p_3) + d(p_3, p_2)$
- (Aduenko, 2017)
  - $d \in [0, 1]$
  - d is defined in case of different support for p_1, p_2
  - d is nearly zero, if p_2 is a low-informative distribution
- Performance criteria
  - Tractable
  - Easy to compute

Total variation

For two probability measures P_1, P_2 on the set $\mathfrak{A}$

$$TV = \sup_{\mathfrak{a}\in\mathfrak{A}}|P_1(\mathfrak{a}) - P_2(\mathfrak{a})|$$

Properties:

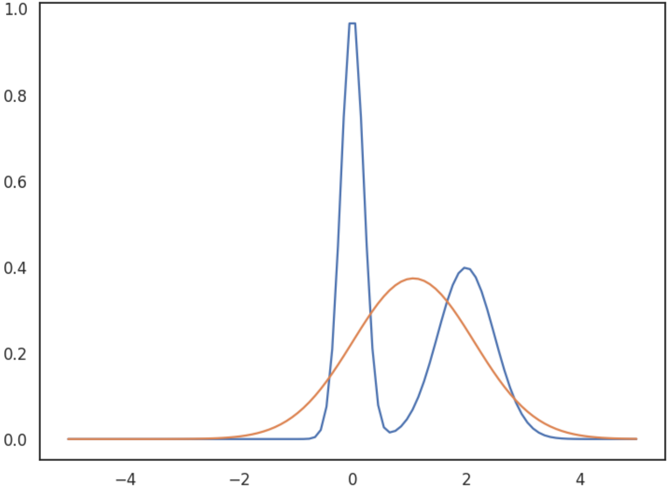
- $0 \leq TV \leq 1$
- TV is a metric
- $TV = 0 \iff P_1 = P_2$
- Scheffe lemma: for differentiable distributions with PDF f_i defined on$\mathbb{R}^d$:

$$TV = \frac{1}{2}\int |f_1(\mathbf{x}) - f_2(\mathbf{x})|d\mathbf{x} = \frac{1}{2}||f_1 - f_2||_1.$$

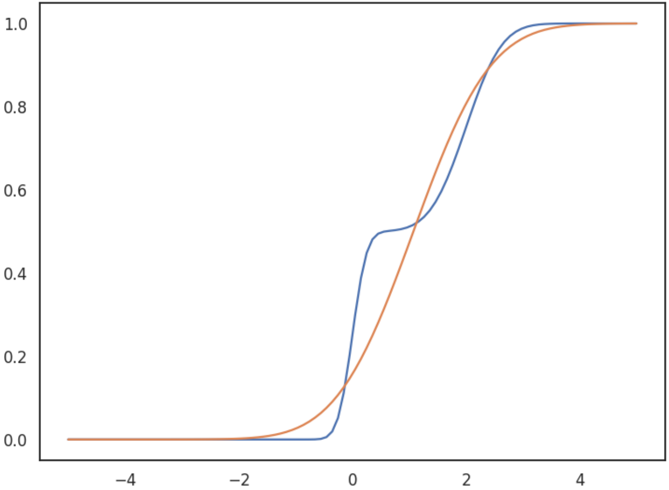
- $TV(\prod_i P_1^i, \prod_i P_2^i) \leq \sum_i TV(P_1^i, P_2^i)$
- Corresponds to statistics in KS-test

Total variation: example

Approximation of Gaussian mixture by Gaussian distribution.

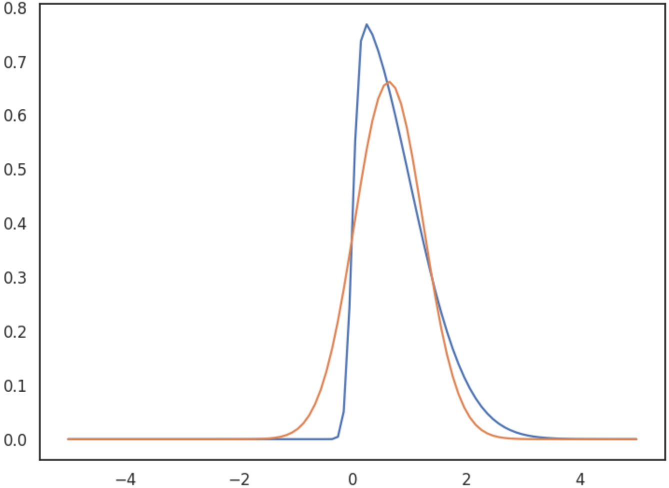


Total variation: example

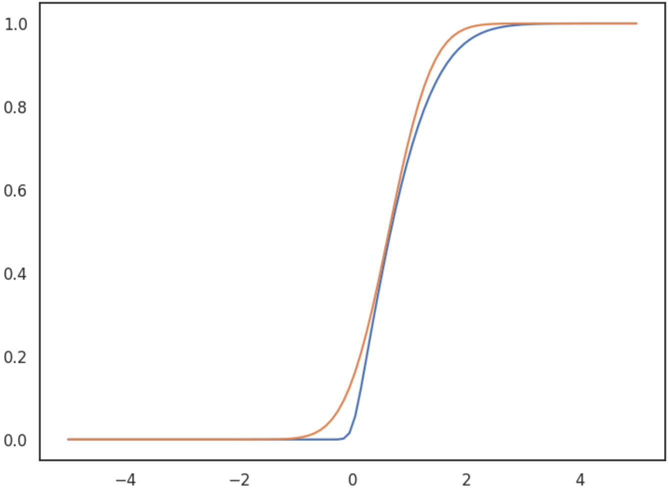


Total variation: example

Approximation of skewed distribution by Gaussian.



Total variation: example

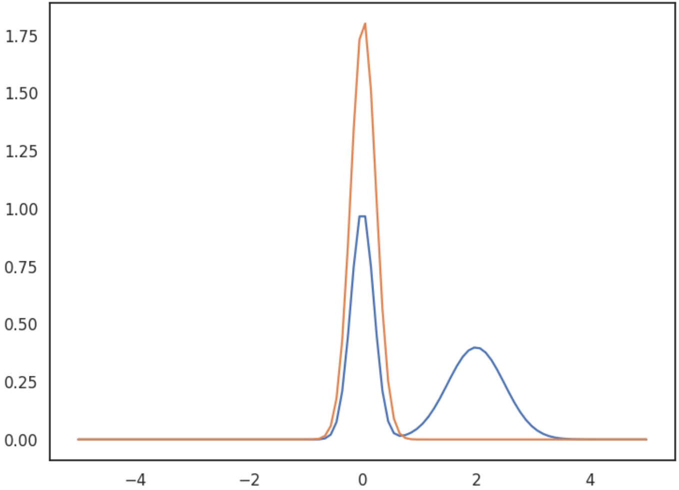


Hellinger distance

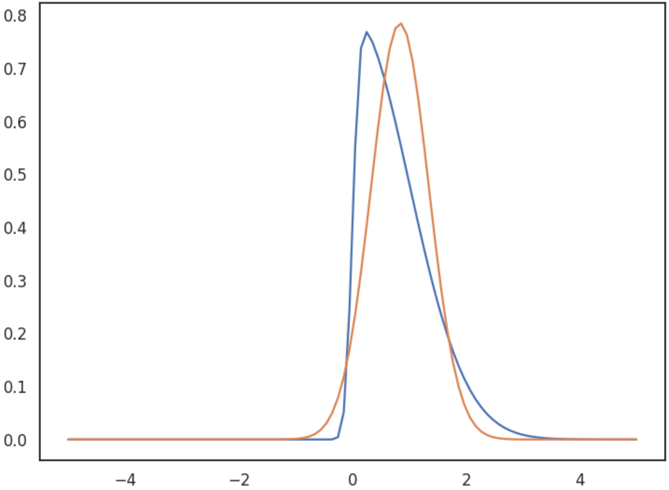
$$H = \sqrt{\int (f_1(\mathbf{x}) - f_2(\mathbf{x}))^2 \, d\mathbf{x}} = ||\sqrt{f_1} - \sqrt{f_2}||_2$$

- $0 \leq H \leq 2$
- H is metric
- $H = 0 \iff P_1 = P_2$
- $H^2(\prod_i P_1^i, \prod_i P_2^i) \leq \sum_i H^2(P_1^i, P_2^i)$
- $1 - H^2 = 1 - \int \sqrt{f_1(\mathbf{x}) f_2(\mathbf{x})} d\mathbf{x}$

Hellinger distance: example



Hellinger distance: example



KL divergence

$$KL(P_1, P_2) = \int \log \frac{f_1(x)}{f_2(x)} f_1(x) dx$$

- $KL \geq 0$
- KL is not a metric: not a symmetric
- KL is not a metric: does not respect triangle inequality
- $KL = 0 \iff P_1 = P_2$
- $KL(\prod_i P_1^i, \prod_i P_2^i) = \sum_i KL(P_1^i, P_2^i)$
- If we have a dependence between 2 random values $\mathbf{w}, \gamma$, then

$$KL\left(p_1(\mathbf{w}, \gamma), p_2(\mathbf{w}, \gamma)\right) = KL(p_1(\mathbf{w}), p_2(\mathbf{w})) + \int_{\mathbf{w}} p_1(\mathbf{w}) \int_{\gamma} \log \frac{p_1(\gamma|\mathbf{w})}{p_2(\gamma|\mathbf{w})} p_1(\gamma|\mathbf{w}) d\gamma d\mathbf{w}$$

Entropy

Differential entropy is a generalization of Shannon entropy:

$$h(\mathbf{w}) = -\int_{\mathbf{w}} \log f(\mathbf{w}) f(\mathbf{w}) d\mathbf{w}$$

- Non-invariant under change of variables
  - $h(\mathbf{F}(\mathbf{w})) \leq h(\mathbf{w}) + \int f(\mathbf{w}) \log \left| \frac{\partial \mathbf{F}}{\partial \mathbf{w}} \right| d\mathbf{w}$
  - If $\mathbf{F}$ is a bijection, inequality turns into equality
- Can be negative

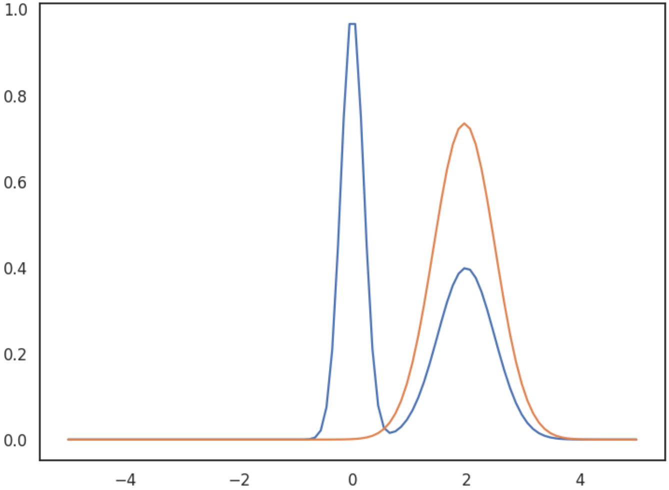
KL is a special case of entropy that

- Invariant under change of variables
- Always positive

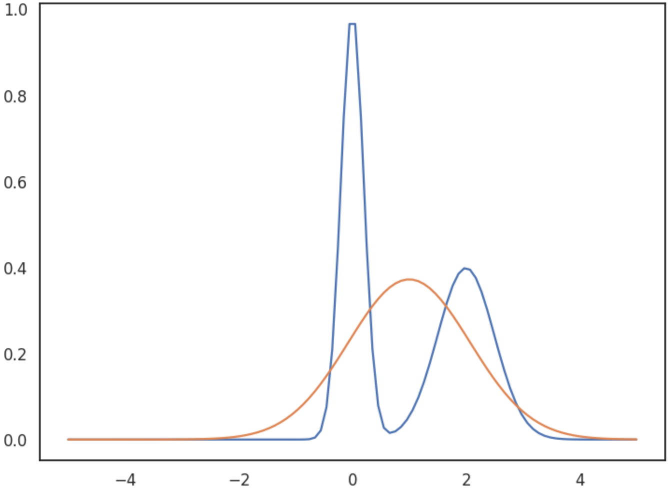
Interpretation of $KL(P_1, P_2)$:

- Amount of information that we can get if use P_1 instead of P_2
- Amount of information that we need to use for coding of data distributed by P_1, if the decoder uses P_2.

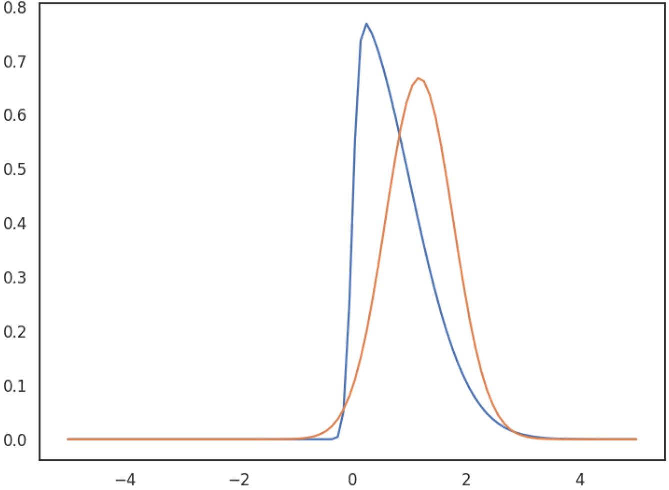
KL: example



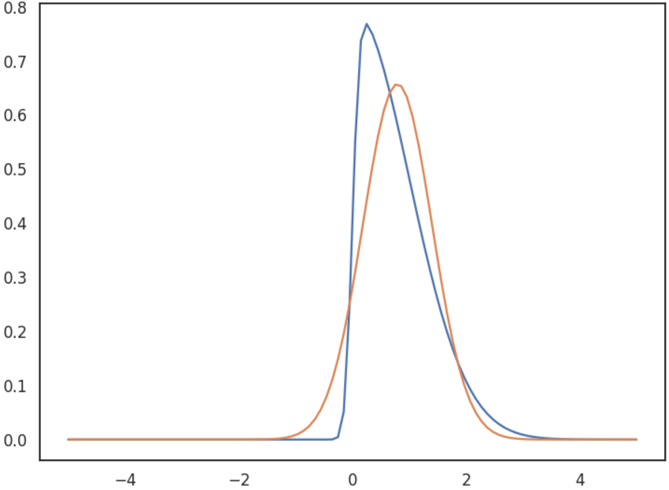
KL: example



KL: example



KL: example

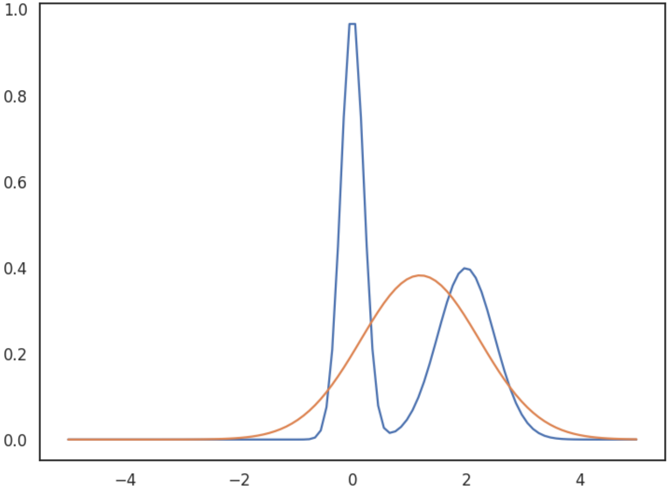


JS

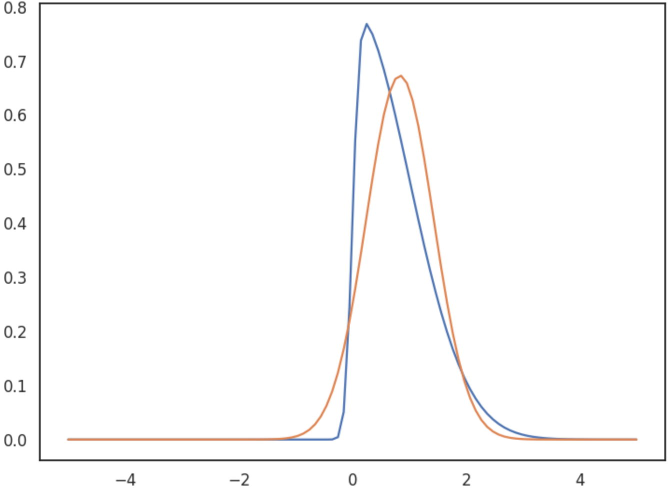
$$JS(P_1, P_2) = \frac{1}{2}KL\left(P_1 \middle| \frac{1}{2}P_1 + \frac{1}{2}P_2\right) + \frac{1}{2}KL\left(P_2 \middle| \frac{1}{2}P_1 + \frac{1}{2}P_2\right)$$

- $0 \leq JS \leq 1$
- $\sqrt{JS}$ is a metric
- $JS = 0 \iff P_1 = P_2$

JS: example

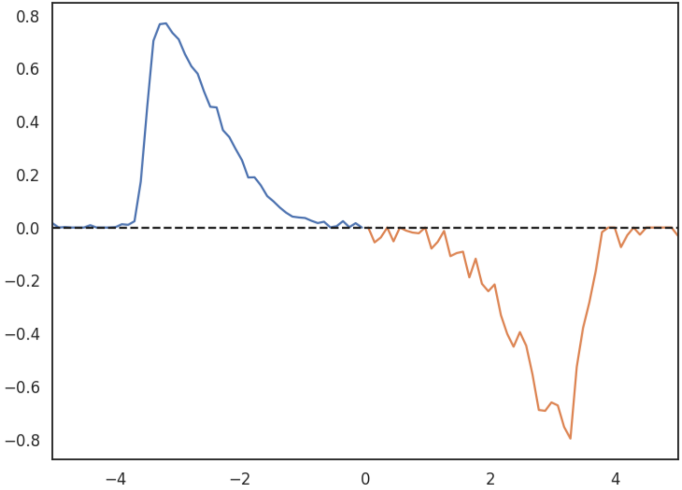


JS: example



Wasserstein distance: motivation

Gaspard Monge: how to move sand into hole in a cheapest way?



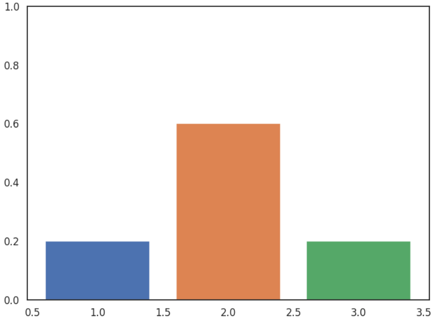
Wasserstein distance: discrete problem

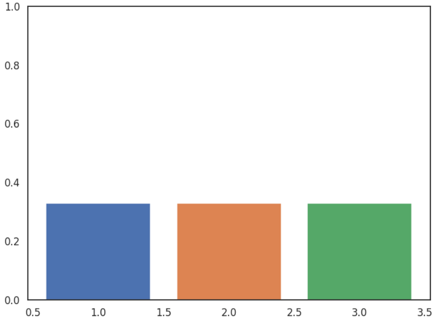
Given two discrete probability measures $p_1(\mathbf{w}_i^1), i \in \{1, \dots n_1\}$, $p_2(\mathbf{w}_j^2), j \in \{1, \dots n_2\}$.
Given a cost matrix $\mathbf{C}$: $c_{ij} \in \mathbb{R}_+$.

We need to find a mapping induced my matrix t_{ij} that:

- $\sum_i t_{ij} = p_2(\mathbf{w}_j^2), \sum_j t_{ij} p_2(\mathbf{w}_i^1)$
- $\sum_i \sum_j c_{ij} t_{ij} \to \min.$

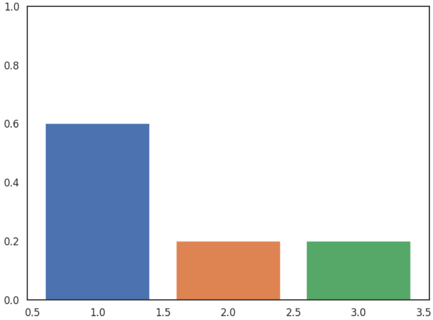
Discrete problem: example

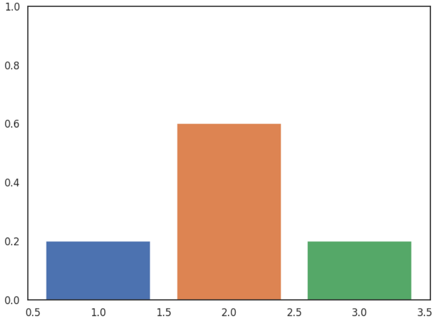




Cost: 0.4

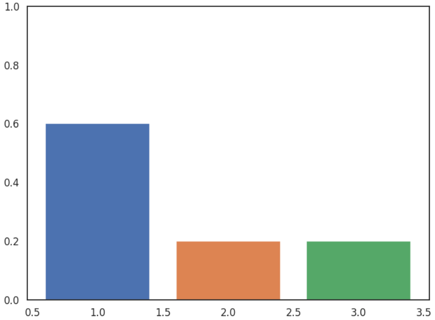
Discrete problem: example

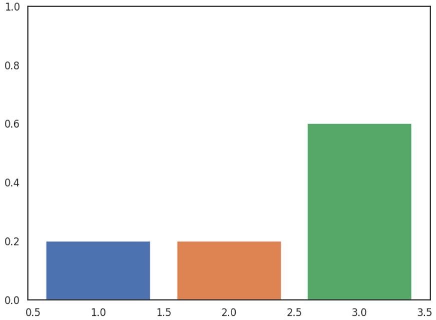




Cost: 0.4

Discrete problem: example





Cost: 0.8

Continuos problem

Given 2 continuos measures $P_1(\mathbf{w}^1), \mathbf{w}^1 \in \mathbb{W}_1,\ P_2(\mathbf{w}^2), \mathbf{w}^2 \in \mathbb{W}_2$.
Given a cost function $C : \mathbb{W}_1 \times \mathbb{W}_2 \to \mathbb{R}_+$.

We need to find a join distribution T on $\mathbb{W}_1 \times \mathbb{W}_2$ that:

- $\int_{\mathbb{W}_1} dT(\mathbf{w}_1, \mathbf{w}_2) = P_1, \quad \int_{\mathbb{W}_2} dT(\mathbf{w}_1, \mathbf{w}_2) = P_2$
- $\int_{\mathbb{W}_1 \times \mathbb{W}_2} C(\mathbf{w}_1, \mathbf{w}_2) dT(\mathbf{w}_1, \mathbf{w}_2) \to \min.$

Dual problem

$$\max_{\hat{T}_1, \hat{T}_2} \int_{\mathbb{W}_1} \hat{T}_1(\mathbf{w}_1) f_1(\mathbf{w}_1) d\mathbf{w}_1 + \int_{\mathbb{W}_2} \hat{T}_2(\mathbf{w}_2) f_2(\mathbf{w}_2) d\mathbf{w}_2$$

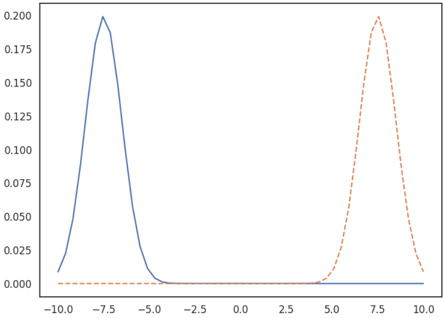
when $\hat{T}_1(\mathbf{w}_1) + \hat{T}_2(\mathbf{w}_2) \leq C(\mathbf{w}_1, \mathbf{w}_2)$

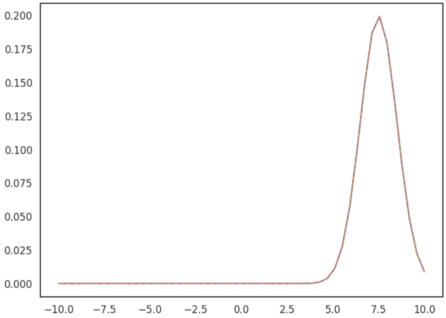
Kantorovich–Rubinstein theorem

Let $\mathbb{W}_1 = \mathbb{W}_2$ and $C = ||\cdot||_1$. Then:

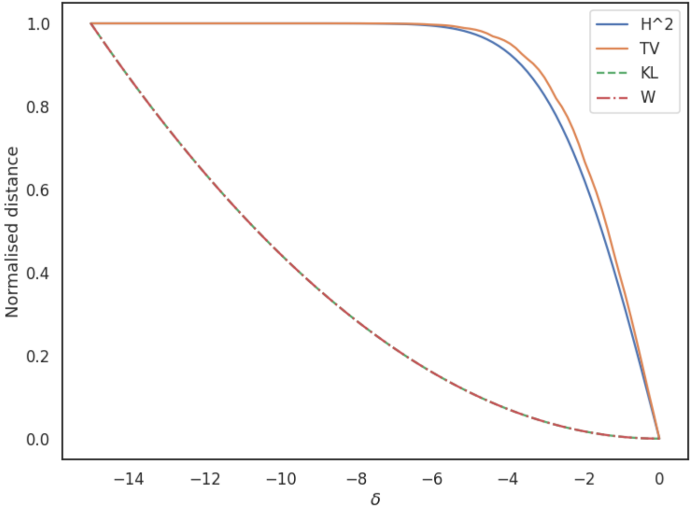
$$\max \hat{T} \in \mathrm{Lip}_1 \int_{\mathbb{W}} \hat{T}(\mathbf{w}) f_1(\mathbf{w}) d\mathbf{w} - \int_{\mathbb{W}} \hat{T}(\mathbf{w}) f_2(\mathbf{w}) d\mathbf{w}$$

Distance between peaks: example

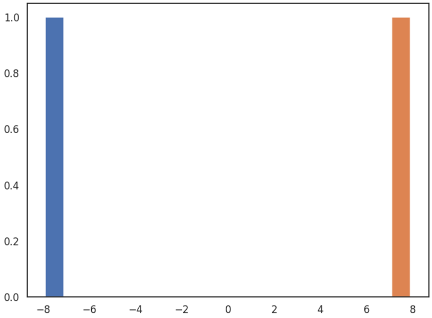


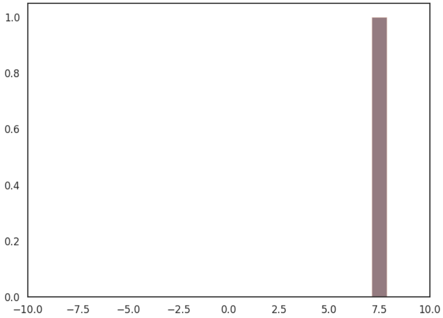


Distance between peaks: example



Distance between peaks: example





Distance between peaks: example

$$TV = 0$$

$$H = 0$$

$$KL = \begin{cases} 0, & \delta = 0 \\ \infty, & \text{otherwise} \end{cases}$$

$$JS = \begin{cases} 0, & \delta = 0 \\ \log 2, & \text{otherwise} \end{cases}$$

$$W = |\delta|.$$

Conclusion: W-distance has good properties to work with different support sets.

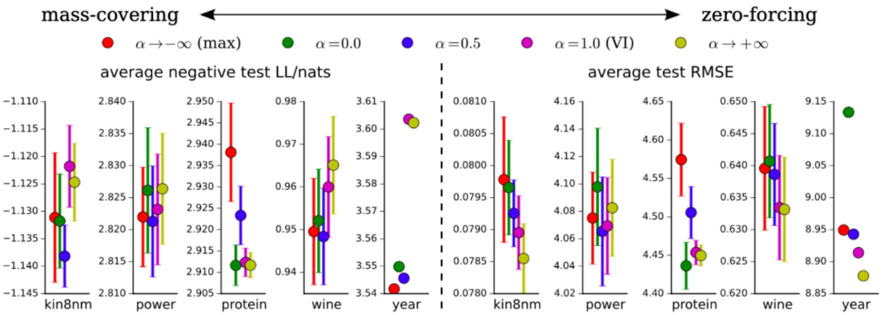


Figure: Test LL and RMSE results for Bayesian neural network regression.

See talk by Kseniia Petrushina, 2023

References

- Адуенко А. А. Выбор мультимоделей в задачах классификации : дис. – Федер. исслед. центр"Информатика и управление"РАН, 2017.
- Andrew Nobel: Distances and Divergences for Probability Distributions, https://nobel.web.unc.edu/wp-content/uploads/sites/13591/2020/11/Distance-Divergence.pdf
- Про KL с условной вероятонстью: http://akosiorek.github.io/ml/2017/09/10/kl-hierarchical-vae.html
- Kolouri, Cattell, Rohde: Optimal Transport: A Crash Course, http://imagedatascience.com/transport/OTCrashCourse.pdf
- Computational Optimal Transport - https://arxiv.org/pdf/1803.00567.pdf
- Про GAN и WGAN: https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html#kullbackleibler-and-jensenshannon-divergence
- Wasserstein GAN - https://arxiv.org/abs/1701.07875
- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – No. 9.
- Bishop C. Bayesian pca //Advances in neural information processing systems. – 1998. – Т. 11.
- Sola J., Deray J., Atchuthan D. A micro Lie theory for state estimation in robotics //arXiv preprint arXiv:1812.01537. – 2018.
- Brehmer J., Cranmer K. Flows for simultaneous manifold learning and density estimation //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 442-453.
- Alain G., Bengio Y. What regularized auto-encoders learn from the data-generating distribution //The Journal of Machine Learning Research. – 2014. – Т. 15. – No. 1. – С. 3563-3593.
- Kingma D. P., Welling M. Auto-encoding variational bayes //arXiv preprint arXiv:1312.6114. – 2013.
- Ranasinghe T., Orˇasan C., Mitkov R. Semantic textual similarity with siamese neural networks //Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). – 2019. – С. 1004-1011.
- https://russianblogs.com/article/5172713037/
- Karaletsos T., Belongie S., Ratsch G. Bayesian representation learning with oracle constraints //arXiv preprint arXiv:1506.05011. – 2015.
- Kuznetsova R., Bakhteev O., Ogaltsov A. Variational learning across domains with triplet information //arXiv preprint arXiv:1806.08672. – 2018.
- Zhang M. et al. Differentiable neural architecture search in equivalent space with exploration enhancement //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 13341-13351.
- Luo R. et al. Neural architecture optimization //Advances in neural information processing systems. – 2018. – Т. 31.
- Yan S. et al. Does unsupervised architecture representation learning help neural architecture search? //Advances in Neural InformationProcessing Systems. – 2020. – Т. 33. – С. 12486-12498