ON THE DISCRIMINABILITY OF SAMPLES IN THE HAMMING SPACE OF BINARIZED RELU ACTIVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Binarized ReLU activations are considered as a metric space equipped with the Hamming distance. While for two layer ReLU networks with random Gaussian weights it can be shown theoretically that local metric properties are approximately preserved, we experimentally study the discrimination capability in this Hamming space for deeper ReLU networks and look also at the non-local behavior. It turns out that the discrimination capability is approximately preserved as expected, but showing small saturation effects that differ from standard metrics based on full activation information. These effects are explained based on the fact that the binarized activation states induce a tessellation of polyhedral cells in the input space.

1 INTRODUCTION

In this paper we concentrate on deep ReLU networks with binarized activations and outputs. ReLU networks perform particularly well in many practical tasks, such as generative adversarial networks (Heusel et al. (2017); Salimans et al. (2016)), domain adaptation methods (Ganin et al. (2016); Long et al. (2017); Zellinger et al. (2021)), and two-sample tests based on neural networks (Lopez-Paz & Oquab (2017); Kirchler et al. (2020)). Particularly for embedded systems, the reduction of precision in the inference is interesting from the point of view of keeping computational efforts and power consumption low, see e.g. Conti et al. (2018); Meloni et al. (2019). But there is also a more theoretical motivation that comes from metric embedding by looking at neural networks as metric preserving mappings in some appropriate spaces (Indyk & Matousek (2004); van der Maaten & Hinton (2008); Suárez-Díaz et al. (2018); Xiao et al. (2018); Courty et al. (2018); Giryes et al. (2016)). For example, in Giryes et al. (2016) two layer ReLU networks with random Gaussian weights and binarized activations and outputs, respectively, are considered. Interestingly, this setting guarantees approximately isometric embedding into the Hamming space. This analysis shows that each standard DNN layer (with random Gaussian weights) performs a stable embedding of the data from one layer to the next by preserving local structures in the manifold. For deeper networks, the analysis becomes much more complicated due to the nested composition of non-linear functions. In contrast to shallow networks deep networks allow representing restricted Boltzmann machines with a number of parameters exponentially greater than the number of the network parameters, as shown by Montúfar & Morton (2015). Looking at the preimage in the layer below induced by the layer above for some specific output, for example 0, shows that a ReLU network leads to a tree of nested polyhedral cells which become smaller in size the deeper the network. By constructing specific ReLU networks Montufar et al. (2014) show exemplarily that deep networks divide the input space into an exponential number of (polyhedral) sets, which is not possible with a single layer with the same number of parameters. This way, deep neural networks are more expressive than shallow ones. Serra et al. (2018) provide an upper and lower bound on the number of polyhedral cells, and consider the influence of width versus depth of ReLU network on the number of created linear regions, finding out that wider ReLU networks result in finer a tessellation then deeper ones. Though such partial results it is not yet fully understood how the network architecture influences the geometry and distribution of the induced cells, Shepeleva et al. (2020) points out that these induced polyhedral cells C_i are actually equivalence classes $[x]$ (up to the border) resulting from the equivalence relation, $x \sim y$, that two sample points in the input space are considered equivalent if they show the same binarized activation profile. We take up this view by introducing a metric in the tessellation space of cells, $\mathcal{T} = \{\overline{[x]} \mid, x \in \mathbb{R}^d\}$, as editing distance

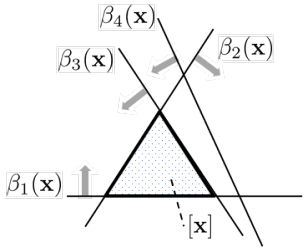


Figure 1: Illustration that d_H does not establish an isometry if hyperplanes in the tessellation do not intersect with the cell. For example, flipping the activation state β_4 determines a cell not adjacent with $[x]$.

of adjacent cells. This means, the number of minimal adjacent cells sharing a common face connecting two cells in this tessellation introduces a notion of distance, $d_{\mathcal{T}}$. In our analysis we exploit the relation that the Hamming distance, d_H, on the binary activation states and the distance in the tessellation have the same distinguishability behavior. This means, the Hamming distance is able to distinguish between points in different cells, meaning $d_H(a(x_1), a(x_2)) > 0 \iff d_{\mathcal{T}}(x_1, x_2) > 0$, where $a(x) = (\beta_{1,1}(x), \ldots, \beta_{1,n_1}(x); \ldots; \beta_{L,1}(x), \ldots, \beta_{L,n_L}(x))$ denotes the vector of binarized activations β_{k,i_k} in the k-th layer of neuron $i_k \in \{1, \ldots, n_k\}$: $\beta_{k,i_k}(x) = 1$ if $a_{k,i_k}(x) > 0$ and $\beta_{k,i_k}(x) = 0$ else with a total number of L layers, where a_{k,i_k} denotes the activation in node i_k in layer k. Note that d_H does not establish an isometry in general as illustrated by the example in figure 1.

Our contribution: While for two layer ReLU networks with random Gaussian weights it can be shown theoretically that local metric properties are approximately preserved, we experimentally study the discrimination capability in this Hamming space for deeper ReLU networks and look also at the non-local behavior. We show experimentally that the discrimination capability is approximately preserved locally also for deeper networks. In this context we give synthetic examples which indicate that binarized activation values contain enough of information to distinguish between points localized differently in the data space.

Section 2 discusses the used distances for the experiments in section 3, followed by final conclusions and an outlook in section 4. Appendix A supplements some convergence results.

2 Metrics Used in the Experimental Setup

We study the behaviour of the following distances:

(a) **Hamming distance.** For samples $S_1 = \{x_1\}$ and $S_2 = \{y_1\},\ x_1, y_1 \in \mathbb{R}^d$, we study the behaviour of the Hamming distance on the binarized activation values,

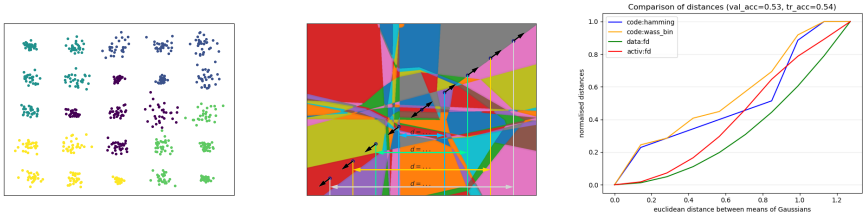
(b) **Wasserstein.** a Wasserstein-p distance with a Hamming base distance and $p = 1$ (see e.g. Peyré et al. (2019)),

(c) **Maximum Mean Discrepancy (MMD).** (Gretton et al. (2012)) with exponentiated Hamming kernel $k(x, y) = \exp\{-d_H(x, y)\}$ (see Yang et al. (2018)).

Maximum Mean Discrepancy and Wasserstein-p distance belong to the family of integral probability measures. Moreover, Wasserstein distance enjoys a geometrical interpretation, what makes it particularly well-suited to work with once geometry of data is involved. In appendix A we complement our findings by convergence results of type

$$\mathbb{P}\left(|W_p(S_1, S_2) - W_p(\{\beta(x) \mid x \in S_1\}, \{\beta(x) \mid x \in S_2\})| > \epsilon\right) < \delta.$$

In our experiments, we estimate Wasserstein distance based on the Sinkhorn algorithm as proposed in Cuturi (2013). Figure 2 illustrates the three steps of our experimental setup for distance analysis based on discriminating two points.



(a) Training data. (b) Tessellation and test data. (c) Metric analysis.

Figure 2: A "visual guide" to our paper: figure 2a shows the training set-up for a ReLU with 3 layers and 15 neurons, figure 2b the testing set-up, and figure 2c the result in terms of metrics measured in the activation space versus the Euclidean distance in the input space. The y axis is a normalized value of proposed distances. For detailed description please see section 3.

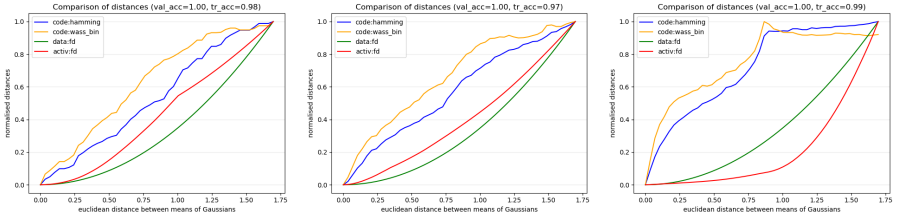
3 Experiments

We remain in the controlled environment: we create a number of multivariate normal distributions, with means spaced evenly on n-dim space, and covariance matrices such that observations sampled from differently centered distributions do not overlap significantly when visually inspected (as in figure 2a). To verify discriminability, we check the alignment of distances listed in section 2 with a chosen benchmark, here Wasserstein-2 distance, which in case of normal distributions can be estimated using first and second moments (see Fréchet (1957), abbreviated fd in the plots), computed on (1) data space, (2) space of activations. We follow the following procedure:

1. train a ReLU network to distinguish an indicated number of groups (for example, in figure 2a we distinguish 5 groups among 25 differently centered probability distributions, as indicated by colours; grouping is done with k-means algorithm: we group each Gaussian in the grid based on proximity to the nearest cluster center), and store the network's parameters,
2. create sample consisting of one point (subsection 3.1) or more points (subsection 3.2) on the diagonal of the cube of our training set-up (as in figure 2b)
3. propagate such a "diagonal data" through pretrained ReLU network, working henceforth with its binarized activation values,
4. compare the behaviour of distances listed in section 2 with the benchmark distance described above.

3.1 Influence of Number of Layers

We present partly results in figure 3. Note that the saturation region tends to be larger for lower numbers (e.g., 3) of layers compared to higher numbers (e.g., 10). This effect reflects the higher concentration of cells in the tessellation for smaller distances versus decreasing concentration of cells for larger distances and will be amplified by a higher number of cells caused by a higher number of layers in consistency with the analysis of (Montufar et al. (2014)).

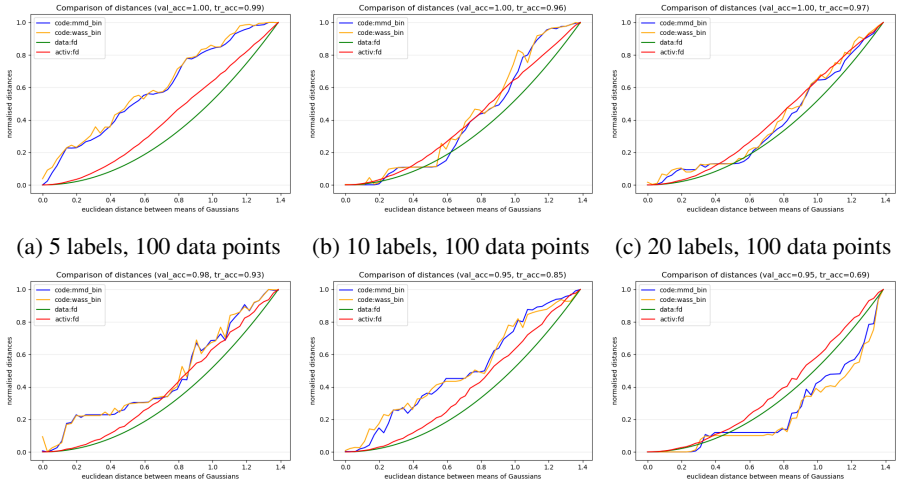


(a) 3 layers, 50 neurons each (b) 5 layers, 128 neurons each (c) 10 layers, 256 neurons each

Figure 3: Behaviour of metrics with varying number of layers. Prefixes indicate how we calculate the distances: "code" - using binarized activation values, "activ" - using activation values on neurons after applying ReLU, "data" - using pure data samples.

3.2 Influence of Sample Size and Number of Training Labels

In the following we show that our results extend to larger sample sizes, different number of labels and also consider the Maximum Mean Discrepancy.



(a) 5 labels, 100 data points (b) 10 labels, 100 data points (c) 20 labels, 100 data points

(d) 5 labels, 10 data points (e) 10 labels, 10 data points (f) 20 labels, 10 data points

Figure 4: Comparison of distance behaviour with varying number of data points and number of labels we assign to points. We used ReLU network with 3 hidden layers 50 neurons each.

4 Conclusion and Future Work

Though we definitely lose the isometry property compared to the setting of (Giryes et al. (2016)) due to the effect pointed out in figure 1, binarized activated states preserve astonishing distinguishability capabilities. The analysis could be refined by taking the geometry of the cells into account. This means, by checking which of the activation states β_{k,i_k} refer to hyperplanes that touch the cell. This way we expect to establish an isometry embedding, what is left for an upcoming paper. In this context we will also check applications, e.g., by constructing a two-sample test statistics using binarized activation values of some sample. Moreover, we will check the effects where data is sparse compared to the number of dimensions.

Acknowledgements

REFERENCES

François Bolley and Cédric Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 14(3):331–352, 2005. URL http://www.numdam.org/item/AFST_2005_6_14_3_331_0.

Francesco Conti, Pasquale D. Schiavone, and Luca Benini. Xnor neural engine: a hardware accelerator ip for 21.6 fj/op binary neural network inference. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2940 – 2951, 2018. ISSN 0278-0070. doi: 10.3929/ethz-b-000279119.

Nicolas Courty, Rémi Flamary, and Mélanie Ducoffe. Learning Wasserstein Embeddings. In *ICLR 2018 - 6th International Conference on Learning Representations*, pp. 1–13, Vancouver, Canada, April 2018. URL https://hal.inria.fr/hal-01956306.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*, volume 2, pp. 4, 2013.

Amir Dembo and Ofer Zeitouni. *Applications-The Finite Dimensional Case*, pp. 71–114. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-03311-7. doi: 10.1007/978-3-642-03311-7_3. URL https://doi.org/10.1007/978-3-642-03311-7_3.

M. Maurice Fréchet. Sur la distance de deux lois de probabilité. *C.R. Acad. Sci. Paris*, 244:689–692, 1957.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

Raja Giryes, G. Sapiro, and A. Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, 64:3444–3457, 2016.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

Piotr Indyk and Jiri Matousek. Low-distortion embeddings of finite metric spaces. In *in Handbook of Discrete and Computational Geometry*, pp. 177–196. CRC Press, 2004.

Matthias Kirchler, Shahryar Khorasani, Marius Kloft, and Christoph Lippert. Two-sample testing using deep learning. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1387–1398. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/kirchler20a.html.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2208–2217, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/long17a.html.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, Toulon, France, April 2017. URL https://hal.inria.fr/hal-01862834.

Paolo Meloni, Daniela Loi, Paola Busia, Gianfranco Deriu, Andy D. Pimentel, Dolly Sapra, Todor Stefanov, Svetlana Minakova, Francesco Conti, Luca Benini, Maura Pintor, Battista Biggio, Bernhard Moser, Natalia Shepeleva, Nikos Fragoulis, Ilias Theodorakopoulos, Michael Masin, and Francesca Palumbo. Optimization and deployment of cnns at the edge: The aloha experience. In *Proceedings of the 16th ACM International Conference on Computing Frontiers*, CF

'19, pp. 326–332, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366854. doi: 10.1145/3310273.3323435. URL https://doi.org/10.1145/3310273.3323435.

Guido F. Montúfar and Jason Morton. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29(1):321–347, jan 2015. doi: 10.1137/140957081.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/109d2dd3608f669ca17920c511c2a41e-Paper.pdf.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11:355–607, 2019.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *International Conference on Machine Learning*, pp. 4558–4566, 2018.

Natalia Shepeleva, Werner Zellinger, Michal Lewandowski, and Bernhard Moser. Relu code space: A basis for rating network quality besides accuracy. Arxiv pre-print, 2020. URL https://arxiv.org/pdf/2005.09903.pdf.

Juan Luis Suárez-Díaz, Salvador García, and Francisco Herrera. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms, Experimental Analysis, Prospects and Challenges (with Appendices on Mathematical Background and Detailed Algorithms Explanation). *arXiv e-prints*, art. arXiv:1812.05944, December 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Chang Xiao, Peilin Zhong, and Changxi Zheng. Bourgan: Generative networks with metric embeddings. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 (NIPS 2018)*, pp. 2269–2280, 2018.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5561–5570, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/yang18c.html.

Werner Zellinger, Bernhard A Moser, and Susanne Saminger-Platz. On generalization in moment-based domain adaptation. *Annals of Mathematics and Artificial Intelligence*, pp. 333–369, 2021.

A CONVERGENCE RESULTS BETWEEN SPACES

In this section we present some convergence results between samples $\mathbf{x}, \mathbf{y}$ and their binarized activations $\beta(x), \beta(y)$ in Wasserstein-p distance. We consider a ReLU network with L hidden layers and total number of neurons on hidden layers $N = \sum_{k=1}^{L} n_k$, denote the activation vector at layer k by $a_k(x) = (a_1^{(k)}(x), \dots, a_{n_k}^{(k)}(x)) = \mathrm{relu} \circ g_k \circ \mathrm{relu} \circ \dots \circ \mathrm{relu} \circ g_0(x)$, and denote its binarized activation by $\beta(x) := (\beta_1^{(k)}(x), \dots, \beta_{n_k}^{(k)}(x))$, where $\beta_i^{(k)}(x) = 1$ if $a_i^{(k)}(x) > 0$ and $\beta_i^{(k)}(x) = 0$

otherwise. Remark that, for *truncated Hamming distance* $d_{H,\theta}(x,y) := \min\{d_H(x,y),\theta\}$ and ReLU network f, a pair $(\mathcal{X}_f, d_{\mathrm{H},\theta})$, where $\mathcal{X}_f := \{\beta(x) \in \{0,1\}^N \mid x \in \mathbb{R}^d\}$, is a metric space.

For any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_1}, \ldots, \beta_{n_L}) \in \beta(x)$ let $L_N^{\mathbf{x}}$ be the associated empirical measure:

$$L_N^{\boldsymbol{\beta}}(\cdot) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\beta_i}(\cdot),$$

where δ_{β_i} represents the Dirac delta mass at $\beta_i \in \beta(x)$. We also denote with $\mathcal{L}_n$ the following set

$$\mathcal{L}_n = \{\nu : \nu = L_n^{\boldsymbol{\beta}} \quad \text{for some} \quad \boldsymbol{\beta} \in \beta(x)\}$$

Theorem A.1. *For $n \geq m$, let $\mu^1, \mu^2 \in \mathcal{P}(\beta(x))$. Let also $X_1^1, \ldots X_n^1$ and $X_1^2 \ldots X_m^2$, be independent $\beta(x)$ valued random variables with distributions μ^1 and μ^2 respectively. We have $\mathbb{P}(|W_p(\mu^1,\mu^2) - W_p(L_n^{\mathbf{X}^1}, L_m^{\mathbf{X}^2})| > \epsilon) < (n+1)^{|\beta(x)|}\exp(-\epsilon n 2^{-(1+2p)/2p}|\beta(x)|^{-1})$, where $|\beta(x)| :=$ sup$\{d_H(\beta_i,\beta_j),\ \beta_i,\beta_j \in \beta(x)\}$.*

To ease notation, we will write $W(\cdot,\cdot)$ instead of $W_p(\cdot,\cdot)$.

Proof. By triangular inequality we have

$$|W(\mu^1,\mu^2) - W(L_n^{\mathbf{X}^1}, L_m^{\mathbf{X}^2})| \leq W(\mu^1, L_n^{\mathbf{X}^1}) + W(\mu^2, L_m^{\mathbf{X}^1}),$$

Therefore we have

$$\begin{aligned}
&\mathbb{P}\left(|W(\mu^1,\mu^2) - W(L_n^{\mathbf{X}^1}, L_m^{\mathbf{X}^2})| \geq \epsilon\right) \leq \mathbb{P}\left(W(\mu^1, L_n^{\mathbf{X}^1}) + W(\mu^2, L_m^{\mathbf{X}^1}) \geq \epsilon\right) \leq \\
&\mathbb{P}\left(W(\mu^1, L_n^{\mathbf{X}^1}) \geq \epsilon/2 \text{ or } W(\mu^2, L_m^{\mathbf{X}^2}) \geq \epsilon/2\right) \leq \\
&\mathbb{P}\left(W(\mu^1, L_n^{\mathbf{X}^1}) \geq \epsilon/2\right) + \mathbb{P}\left(W(\mu^2, L_m^{\mathbf{X}^2}) \geq \epsilon/2\right) \leq \\
&\max\left(\mathbb{P}\left(W(\mu^1, L_n^{\mathbf{X}^1}) \geq \epsilon/2\right), \mathbb{P}\left(W(\mu^2, L_m^{\mathbf{X}^2}) \geq \epsilon/2\right)\right).
\end{aligned}$$

Let $\Gamma = \{\nu \in \mathcal{P}(\beta(x)) : W(\mu^1,\nu) \geq \epsilon/2\}$. We have

$$\mathbb{P}\left(W(\mu^1, L_n^{\mathbf{X}^1}) \geq \epsilon/2\right) = \mathbb{P}(L_n^{\mathbf{X}^1} \in \Gamma)$$

By equation 2.1.12 in Dembo & Zeitouni (2010) we have that for every closed set, it holds

$$\mathbb{P}(L_n^{\mathbf{X}^1} \in \Gamma) \leq (n+1)^{|\beta(x)|}\exp(-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu^1))$$

where

$$\inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu^1) = \inf\{H(\nu|\mu^1) : \nu \in \mathcal{L}_n \text{ and } W(\nu,\mu^1) \geq \epsilon/2\}.$$

From the relationship between mutual entropy H and Wasserstein-p distance W_p stated in Bolley & Villani (2005) on a bounded and measurable space $\mathcal{X}$ it holds that $W_p(\mu,\nu) \leq 2^{1/2p}|\beta(x)|H(\mu|\nu)^{1/2p}$ for $\mu,\nu \in \mathcal{P}(\beta(x))$. Thus joining above equations we obtain

$$H(\nu,\mu^1) \geq 2^{-1/2p}|\beta(x)|^{-1}W_p(\nu,\mu^1) \geq 2^{-(1+2p)/2p}(|\beta(x)|)^{-1}\epsilon,$$

which results in

$$\mathbb{P}(|W(\mu^1,\mu^2) - W(L_n^{\mathbf{X}^1}, L_m^{\mathbf{X}^2})| > \epsilon) \leq (n+1)^{|\beta(x)|}\exp\big(-\epsilon n 2^{-(1+2p)/2p}|\beta(x)|^{-1}\big).$$

□