Directed Graph Embeddings in Pseudo-Riemannian Manifolds Supplementary Material

Aaron Sim Maciej Wiatrak Angus Brayne Páidí Creed Saee Paliwal

1. Relationship with Disk Embeddings

We prove the following result, relating Euclidean Disk Embeddings (Suzuki et al., 2019) to the Triple Fermi-Dirac (TFP) model in Minkowski spacetime.

Proposition 1. *Let $p = (\mathbf{x}, u)$ and $q = (\mathbf{y}, v)$ denote elements of $\mathbb{R}^n \times \mathbb{R}$, D be the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$, and $T = v - u$ the time difference. Then, for the choice of TFD function parameters $\alpha = 0, r = 0, k = 1$ we have $\mathcal{F}(p, q) \geq \frac{1}{2}$ if and only if*

$$D \leq \left(\tau_1 \log \frac{3 - e^{-T/\tau_2}}{1 + e^{-T/\tau_2}} + T^2 \right)^{1/2} \tag{1}$$

Proof. From the definition of the Triple Fermi-Dirac probability on Minkowski space-time we have

$$\mathcal{F}(p, q) = \left(\frac{1}{e^{(D^2 - T^2)/\tau_1} + 1} \frac{1}{e^{-T/\tau_2} + 1} \frac{1}{2} \right)^{1/3}$$

Then,

$$\begin{aligned} \mathcal{F}(p, q) \geq \frac{1}{2} &\iff \frac{1}{e^{(D^2 - T^2)/\tau_1} + 1} \frac{1}{e^{-T/\tau_2} + 1} \geq \frac{1}{4} \\ &\iff e^{(D^2 - T^2)/\tau_1} \leq \frac{3 - e^{-T/\tau_2}}{1 + e^{-T/\tau_2}} \\ &\iff D \leq \left(\tau_1 \log \frac{3 - e^{-T/\tau_2}}{1 + e^{-T/\tau_2}} + T^2 \right)^{1/2} \end{aligned}$$ □

The set of points p and q from $\mathbb{R}^n \times \mathbb{R}$ that satisfies the condition for inclusion of Euclidean Disks, which determines the embeddings that are connected by directed edges, is given by $D \leq T$. For $T \geq 0$, every pair p and q which satisfy $D \leq T$ also satisfy (1); hence the set of points $p, q \in \mathbb{R}^n \times \mathbb{R}$ which correspond to a directed edge in the Euclidean Disk Embeddings model is strictly contained in the set of pairs of points which have $\mathcal{F}(p, q) \geq \frac{1}{2}$ in the Triple Fermi-Dirac probability function on (flat) Minkowski space-time. Moreover, the difference between these two sets is a small neighbourhood of the Minkowski light cone, with the size of this set dependent on the parameters τ_1 and τ_2. Figure 1 illustrates this in the $\mathbb{R} \times \mathbb{R}$ case for $\tau_1 = \tau_2 = 0.05$ (the parameters used by the model in the experiments on the WordNet dataset.)

Figure 1. Boundaries of the regions containing points $q \in \mathbb{R} \times \mathbb{R}$ with $\mathcal{F}(p, q) \geq \frac{1}{2}$ (red) and with $D \leq T$ (blue) for $p = (0, 0)$

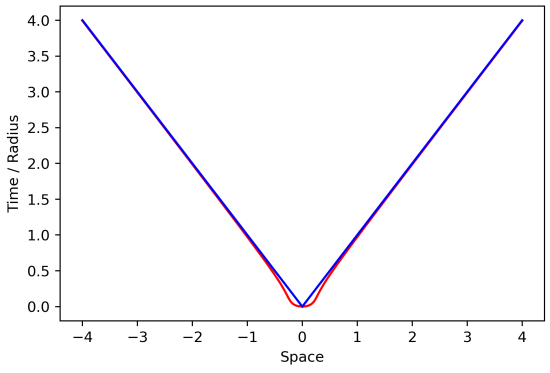

Table 1. Statistics of datasets used in the experiments.

Dataset	Nodes	Edges	Cyclic
Duplication Divergence	100	1026	True
DREAM5: *In silico*	1,565	4,012	True
DREAM5: *E. coli*	1,081	2,066	True
DREAM5: *S. cerevisiae*	1,994	3,940	True
WordNet	82,115	743,086	False

2. Experiment details

2.1. Graph Datasets

This section describes in detail the experiments conducted in our work, including dataset and evaluation specifications.

2.2. Hyperparameters and training details

The model hyperparameters for the experiments on the Duplication Divergence Model and DREAM5 datasets are given in Table 4. We performed a Cartesian product grid search and selected the optimal values based on the median average precision on the test set over three random initializations. We implemented a simple linear learning rate decay schedule to a final rate of a quarter of the initial rate λ; hence although we selected the best test set performance across all the training run epochs, the maximum number of

epochs affects the performance as it determines the rate of learning-rate decay.

The hyperparameter tuning approach for the WordNet experiments was performed differently as it includes an F1 score threshold selection. This is described separately in Section 2.2.3 below.

A summary of the datasets is provided in Table 2.1.

2.2.1. Duplication Divergence Model

The duplication divergence model is a simple two-parameter model of the evolution of protein-protein interaction networks (Ispolatov et al., 2005). Starting from a small, fully-connected, directed graph of n_i nodes, the network is grown to size n_f as follows. First we duplicate a random node from the set of existing nodes. Next the duplicated node is linked to the neighbors of the original node with probability p_1 and to the original itself with probability p_2. This step is repeated $(n_\mathrm{f} - n_\mathrm{i})$ times. Because the model only includes duplicated edges with no extra random attachments, it is easily shown that if the seed network is a directed acyclic graph (DAG), the network remains a DAG throughout its evolution. We generate one cyclic graph from an initial 3-node seed graph, with parameters $(n_\mathrm{i}, n_\mathrm{f}, p_1, p_2) = (3, 100, 0.7, 0.7)$ and perform a single random 85 / 15 train-test split for all experiments.

2.2.2. DREAM5

Traditionally a network inference problem, the DREAM5 challenge (Marbach et al., 2012) is a set of tasks with corresponding datasets. The goal of the challenge and each task is to infer genome-scale transcriptional regulatory networks from gene-expression microarray datasets. We use only the gold standard edges from three DREAM5 gene regulatory networks for our use case: *In silico*, *Escherichia coli* and *Saccharomyces cerevisiae*. The fourth *Staphylococcus aureus* network from the DREAM5 challenge was not considered here due to the issues in evaluation described in Marbach et al. (2012). We extract only the positive-regulatory nodes from the remaining three networks while omitting the gene-expression data itself. We limit ourselves to positive-regulation edges as our work focuses on single edge type graphs. The extracted node pairs form a directed graph, where the edge between nodes (i, j) represent a regulatory relation between two genes. Each network is then randomly split into train and test sets, following 85 / 15 split. Although the network is cyclic, the number of cycles in each of them is relatively low, which may account for the similarity in performances of the (non-cylindrical) Minkowski embeddings with embeddings on cylindrical Minkowski and AdS (e.g. the *E. Coli* dataset experiments in Table 2.2.3). Similar to the Duplication Divergence model above, we evaluate the performance using the average precision on the test set.

2.2.3. WordNet

To further evaluate the embeddings of directed acyclic graphs (DAGs), we use the *WordNet* (Miller, 1998) database of noun hierarchy. The WordNet dataset is an example of a tree-like network with low number of ancestors, and high number of descendants. To ensure fair comparison we use the same dataset and split as in Suzuki et al. (2019), proposed in Ganea et al. (2018). During training and evaluation, we use nodes $(i, j) \in \mathcal{T}$ connected by an edge such that $c_i \preceq c_j$ as positive examples. As the dataset consists only of positive examples, we randomly sampled a set of non-connected negative pairs for each positive pair. For evaluation, we use an F1 score for a binary classification to assess whether a given pair of nodes (i, j) is connected by a directed edge in the graph. Similarly as in (Suzuki, 2019) we compute the results based on the percentage of transitive closure of the graph.

As WordNet is a DAG, we restrict our attention to the model based on flat Minkowski spacetime. Here, the key hyperparameters are the learning rate and values for α, τ_1 and τ_2 (we set $r = 0$ and $k = 1$ in these experiments). We perform a Cartesian product grid search and select the optimal values based on the mean F1 score on the validation set over two random initializations (seeds), using 50 epochs and batch size 50. The threshold for the F1 score was chosen by line search and tuned. The grid of values is as follows, with the optimal values given in **bold**: $\lambda \in \{0.01, \mathbf{0.02}, 0.1\}$, $\alpha \in \{0., 0.0075, \mathbf{0.075}\}$, $\tau_1 \in \{0.005, \mathbf{0.05}, 1.\}$ and $\tau_2 \in \{0.005, \mathbf{0.05}, 1.\}$. We then trained a model with the selected parameters for 150 epochs using 5 different random seeds to generate the results reported in Table 2. Here, we report the test set F1 score using a threshold selected via line search on the validation set.

2.3. Model complexity and runtimes

The model complexity is similar to standard (Euclidean) embedding models. For the pseudo-Riemannian manifold examples in this paper, the computation of the pseudo-Riemannian SGD vector from the JAX autodiff-computed value for the differential df scales linearly with embedding dimension. The approximation to infinite sum in the wrapped TFD function introduces a $\mathcal{O}(m)$ complexity factor, where m is the number of cycles one uses in the approximation (see eq. (21) in the main text).

The cyclic graph link prediction experiments were performed as CPU-only single process runs. The longest runtime for the 100-dimensional AdS manifold embedding model was $\sim$ 19 sec/epoch for the Duplication Divergence model graph and $\sim$ 112 sec/epoch for the *in silico* DREAM5 dataset. The equivalent runtimes for the other

Table 2. F1 percentage score on the test data on WordNet. The best flat-space performance (top-half) for each dataset/embedding dimension combination has its background highlighted in gray, and the best overall is highlighted in **bold**. The benchmark methods results were taken from (Suzuki et al., 2019). For the results of our method, we report the median together with standard deviation across seeds (N=5).

	$d = 5$				$d = 10$			
Transitive Closure Percentage	0%	10%	25%	50%	0%	10%	25%	50%
Minkowski + TFD (Ours)	21.2 ± 0.5	77.8 ± 0.1	86.2 ± 0.7	92.1 ± 0.3	24 ± 0.5	82 ± 1.2	89.3 ± 0.7	**94.4 ± 0.1**
Order Emb. (Vendrov et al., 2016)	34.4	70.6	75.9	82.1	43.0	69.7	79.4	84.1
Euclidean Disk (Suzuki et al., 2019)	35.6	38.9	42.5	45.1	**45.6**	54.0	65.8	72.0
Spherical Disk (Suzuki et al., 2019)	**37.5**	**84.8**	**90.5**	**93.4**	42.0	**86.4**	**91.5**	93.9
Hyperbolic Disk (Suzuki et al., 2019)	32.9	69.1	81.3	83.1	36.5	79.7	90.5	94.2
Hyperbolic EC (Ganea et al., 2018)	29.2	80.0	87.1	92.8	32.4	84.9	90.8	93.8
Poincare Emb. (Nickel et al., 2014)	28.1	69.4	78.3	83.9	29.0	71.5	82.1	85.4

Table 3. Link prediction for directed cyclic graphs with embedding dimension d. Reported above are the median average precision (AP) percentages with standard deviation across seeds (N=20), calculated on a held-out test set for varying embedding dimension. Annotated in **bold** is the top-performing model for the given dimension. For reference, the random baseline AP is 20%.

		Embedding dimension				
		3	5	10	50	100
Duplication Divergence	Euclidean + FD	37.8 ± 2.8	39.4 ± 2.4	39.0 ± 1.9	38.9 ± 1.9	38.9 ± 1.9
	Hyperboloid + FD	36.3 ± 2.2	37.5 ± 2.4	38.2 ± 2.3	38.2 ± 2.4	38.1 ± 2.3
	Minkowski + TFD	43.7 ± 2.2	47.5 ± 2.5	48.5 ± 3.7	48.5 ± 3.7	48.5 ± 3.7
	Anti de-Sitter + TFD	50.1 ± 3.2	52.4 ± 3.3	56.2 ± 3.2	56.3 ± 3.1	56.8 ± 3.0
	Cylindrical Minkowski + TFD	**55.8 ± 3.6**	**61.6 ± 4.8**	**65.3 ± 4.1**	**65.7 ± 3.1**	**65.6 ± 3.2**
DREAM5: *in silico*	Euclidean + FD	29.4 ± 2.1	32.9 ± 2.5	39.7± 1.8	39.8 ± 1.6	34.8± 1.1
	Hyperboloid + FD	28.8 ± 5.5	46.8 ± 4.6	50.8 ± 7.4	50.9 ± 1.5	52.5 ± 1.5
	Minkowski + TFD	36.3 ± 2.3	43.1 ± 3.1	51.2 ± 3.0	57.7 ± 2.8	58.0 ± 2.7
	Anti de-Sitter + TFD	38.1 ± 4.8	45.2 ± 2.3	51.9 ± 5.2	55.6 ± 4.2	56.0 ± 3.4
	Cylindrical Minkowski + TFD	**41.0 ± 3.6**	**48.4 ± 7.3**	**56.3 ± 8.4**	**58.9 ± 2.9**	**61.0 ± 1.9**
DREAM5: *E. Coli*	Euclidean + FD	33.0 ± 3.9	34.2± 3.4	40.2 ± 4.3	44.5 ± 2.6	49.0 ± 3.2
	Hyperboloid + FD	43.4 ± 4.1	47.2 ± 3.3	52.7 ± 1.9	53.6 ± 1.4	50.6 ± 0.7
	Minkowski + TFD	**51.0 ± 4.0**	**58.4 ± 2.3**	**63.4 ± 3.6**	**67.7 ± 2.7**	**68.2 ± 2.4**
	Anti de-Sitter + TFD	42.7 ± 3.7	56.5 ± 2.6	61.8 ± 6.8	63.3 ± 4.8	63.0 ± 7.5
	Cylindrical Minkowski + TFD	50.3 ± 3.3	56.8 ± 3.4	62.3 ± 3.3	65.8 ± 3.4	63.2 ± 2.4
DREAM5: *S. Cerevisiae*	Euclidean + FD	33.0 ± 2.7	34.2 ± 2.8	40.2 ± 3.3	44.5 ± 3.5	49.0 ± 2.0
	Hyperboloid + FD	29.2 ± 2.5	37.9 ± 1.3	46.5 ± 1.6	48.8 ± 1.4	47.9 ± 1.2
	Minkowski + TFD	34.7 ± 2.2	38.6 ± 1.9	46.4 ± 3.1	52.7 ± 3.0	54.0 ± 2.5
	Anti de-Sitter + TFD	37.2 ± 3.2.	41.3 ± 1.5	44.9. ± 2.5	47.5 ± 3.1	49.4 ± 3.3
	Cylindrical Minkowski + TFD	**37.4 ± 3.2**	**42.7 ± 2.3**	**46.8 ± 3.5**	**53.4 ± 2.2**	**54.6 ± 2.1**

manifolds are approximately an order of magnitude shorter.

The WordNet experiments were performed on NVIDIA V100 GPUs. The run-times were largely similar for the cases with embedding dimension 5 or 10, with the proportion of the transitive closure included in the training data (i.e. the training data size) being the main factor which determined run-time in our experiments. These varied from ~ 60 sec/epoch when 0% was included to $\sim$ 280 sec/epoch when 50% was included.

References

Ganea, O., Becigneul, G., and Hofmann, T. Hyperbolic entailment cones for learning hierarchical embeddings. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, pp. 1646–1655, Stockholmsmässan, Stockholm Sweden, 2018.

Ispolatov, I., Krapivsky, P. L., and Yuryev, A. Duplication-divergence model of protein interaction network. *Phys. Rev. E*, 71:061911, Jun 2005.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M.,

Prill, R. J., Camacho, D. M., Allison, K. R., Aderhold, A., Allison, K. R., Bonneau, R., Camacho, D. M., Chen, Y., Collins, J. J., Cordero, F., Costello, J. C., Crane, M., Dondelinger, F., Drton, M., Esposito, R., Foygel, R., Fuente, A. d. l., Gertheiss, J., Geurts, P., Greenfield, A., Grzegorczyk, M., Haury, A.-C., Holmes, B., Hothorn, T., Husmeier, D., Huynh-Thu, V. A., Irrthum, A., Kellis, M., Karlebach, G., Küffner, R., Lèbre, S., Leo, V. D., Madar, A., Mani, S., Marbach, D., Mordelet, F., Ostrer, H., Ouyang, Z., Pandya, R., Petri, T., Pinna, A., Poultney, C. S., Prill, R. J., Rezny, S., Ruskin, H. J., Saeys, Y., Shamir, R., Sîrbu, A., Song, M., Soranzo, N., Statnikov, A., Stolovitzky, G., Vega, N., Vera-Licona, P., Vert, J.-P., Visconti, A., Wang, H., Wehenkel, L., Windhager, L., Zhang, Y., Zimmer, R., Kellis, M., Collins, J. J., and Stolovitzky, G. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012. ISSN 1548-7091.

Nickel, M., Jiang, X., and Tresp, V. Reducing the rank in relational factorization models by including observable patterns. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 1179–1187, 2014.

Suzuki, R., Takahama, R., and Onoda, S. Hyperbolic disk embeddings for directed acyclic graphs. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, pp. 6066–6075, 2019.

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. Order-embeddings of images and language. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations*, 2016.

Table 4. Cyclic graph inference hyperparameters. Optimal values* are in **bold**. * In a situation where the optimal value was dependent on the embedding dimension, we highlight all the values that recorded best performance on at least one embedding dimension.

	circumference	τ_1	τ_2	r	λ	batch size	α	Max epochs
DD Model (Euclidean)	-	$(0.04, 0.075, 0.15, \mathbf{0.4}, 1.0, 2.0)$	-	$(\mathbf{0.0}, 0.1, 1.0)$	$(0.08, \mathbf{0.02}, 0.001)$	**4**	-	**300**
DD Model (Hyperboloid)	-	$(0.04, \mathbf{0.075}, 0.15, 0.4, 1.0, 2.0)$	-	$(\mathbf{0.0}, 0.1, 1.0)$	$(0.08, 0.02, \mathbf{0.001})$	**4**	-	**300**
DD Model (Minkowski)	-	$(\mathbf{0.075}, 0.15, 0.4)$	$(\mathbf{0.03}, 0.05, 0.07)$	**0.0**	$(0.08, \mathbf{0.02})$	**2**	$(0.03, 0.045, \mathbf{0.06}, 0.075, 0.09)$	**200**
DD Model (Cylindrical Minkowski)	$(6, 8, \mathbf{10}, 12, 14)$	$(0.075, 0.15, \mathbf{0.4})$	$(0.03, 0.05, \mathbf{0.07})$	**0.0**	$(0.08, \mathbf{0.02})$	**2**	$(0.03, 0.045, 0.06, 0.075, \mathbf{0.09})$	**200**
DD Model (AdS)	-	**0.4**	**0.15**	$(-0.25, -0.2, -0.15, \mathbf{-0.1}, -0.05, -0.025, 0.0)$	$(0.02, \mathbf{0.016}, 0.012, 0.008)$	**2**	**0.15**	**150**
DREAM5: In Silico (Euclidean)	-	$(\mathbf{0.04}, 0.075, 0.15, 0.4, 1.0, 2.0)$	-	$(\mathbf{0.0}, 0.1, 0.5)$	$(0.2, 0.08, \mathbf{0.02}, \mathbf{0.002})$	**2**	-	**60**
DREAM5: In Silico (Hyperboloid)	-	$(0.04, 0.075, \mathbf{0.15}, 0.4, 1.0, 2.0)$	-	$(\mathbf{0.0}, 0.1, \mathbf{0.5})$	$(0.2, \mathbf{0.08}, 0.02, \mathbf{0.002})$	**2**	-	**60**
DREAM5: In Silico (Minkowski)	-	$(0.075, \mathbf{0.15}, 0.4)$	$(0.03, \mathbf{0.05}, 0.07)$	**0.0**	$(\mathbf{0.2}, 0.08, \mathbf{0.02})$	**2**	$(0.03, 0.045, \mathbf{0.06}, 0.075, 0.09)$	**60**
DREAM5: In Silico (Cylindrical Minkowski)	$(6, \mathbf{8}, 10, 12, 14)$	$(0.075, \mathbf{0.15}, 0.4)$	$(\mathbf{0.03}, 0.05, 0.07)$	**0.0**	$(\mathbf{0.2}, 0.08, \mathbf{0.02})$	**2**	$(0.03, 0.045, 0.06, \mathbf{0.075}, 0.09)$	**60**
DREAM5: In Silico (AdS)	-	$0.075, \mathbf{0.15}$	$(0.05, \mathbf{0.07})$	$\mathbf{-0.1}$	$(\mathbf{0.01}, \mathbf{0.02})$	**2**	$(0.03, 0.04, 0.05, \mathbf{0.06}, 0.07)$	**60**
DREAM5: E. Coli (Euclidean)	-	$(0.04, 0.075, 0.15, 0.4, \mathbf{1.0}, 2.0)$	-	$(\mathbf{0.0}, 0.1, 0.5)$	$(\mathbf{0.08}, \mathbf{0.02}, 0.002)$	**2**	-	**60**
DREAM5: E. Coli (Hyperboloid)	-	$(0.04, 0.075, \mathbf{0.15}, 0.4, 1.0, 2.0)$	-	$(\mathbf{0.0}, \mathbf{0.1}, 0.5)$	$(0.08, 0.02, \mathbf{0.002})$	**2**	-	**60**
DREAM5: E. Coli (Minkowski)	-	$(0.075, \mathbf{0.15}, 0.4)$	$(0.03, 0.05, \mathbf{0.07})$	**0.0**	$(\mathbf{0.2}, 0.08, \mathbf{0.02})$	**2**	$(0.03, 0.045, \mathbf{0.06}, 0.075, 0.09)$	**60**
DREAM5: E. Coli (Cylindrical Minkowski)	$(6, \mathbf{8}, 10, \mathbf{12}, 14)$	$(\mathbf{0.075}, 0.15, 0.4)$	$(0.03, \mathbf{0.05}, 0.07)$	**0.0**	$(0.2, 0.08, \mathbf{0.02})$	**2**	$(0.06, \mathbf{0.075}, 0.09)$	**60**
DREAM5: E. Coli (AdS)	-	**0.15**	**0.07**	$\mathbf{-0.1}$	$(\mathbf{0.01}, \mathbf{0.02})$	**2**	**0.06**	**60**
DREAM5: S. Cerevisiae (Euclidean)	-	$(0.04, 0.075, 0.15, 0.4, 1.0, \mathbf{2.0})$	-	$(\mathbf{0.0}, \mathbf{0.1}, 0.5)$	$(\mathbf{0.2}, \mathbf{0.08}, 0.02, 0.002)$	**2**	-	**60**
DREAM5: S. Cerevisiae (Hyperboloid)	-	$(0.04, 0.075, 0.15, 0.4, 1.0, \mathbf{2.0})$	-	$(\mathbf{0.0}, 0.1, 0.5)$	$(\mathbf{0.2}, \mathbf{0.08}, 0.02, 0.002)$	**2**	-	**60**
DREAM5: S. Cerevisiae (Minkowski)	-	$(0.075, \mathbf{0.15}, 0.4)$	$(\mathbf{0.03}, 0.05, 0.07)$	**0.0**	$(\mathbf{0.2}, 0.08, \mathbf{0.02})$	**2**	$(0.03, \mathbf{0.045}, 0.06, 0.075, 0.09)$	**60**
DREAM5: S. Cerevisiae (Cylindrical Minkowski)	$(6, \mathbf{8}, 10, \mathbf{12}, 14)$	$(\mathbf{0.075}, 0.15, 0.4)$	$(0.03, \mathbf{0.05}, 0.07)$	**0.0**	$(0.2, 0.08, \mathbf{0.02})$	**2**	$(0.03, 0.045, 0.06, \mathbf{0.075}, 0.09)$	**60**
DREAM5: S. Cerevisiae (AdS)	-	**0.15**	**0.07**	$\mathbf{-0.1}$	$(\mathbf{0.01}, \mathbf{0.02})$	**2**	**0.06**	**60**